Small Study Regression Discontinuity Designs: Density Inclusive Study Size Metric and Performance

Daryl Swartzentruber and Eloise Kaizar
Department of Statistics, The Ohio State University
September 2, 2022

Abstract

Regression discontinuity (RD) designs are popular quasi-experimental studies in which treatment assignment depends on whether the value of a running variable exceeds a cutoff. RD designs are increasingly popular in educational applications due to the prevalence of cutoff-based interventions. In such applications sample sizes can be relatively small or there may be sparsity around the cutoff. We propose a metric, density inclusive study size (DISS), that characterizes the size of an RD study better than overall sample size by incorporating the density of the running variable. We show the usefulness of this metric in a Monte Carlo simulation study that compares the operating characteristics of popular nonparametric RD estimation methods in small studies. We also apply the DISS metric and RD estimation methods to school accountability data from the state of Indiana.

Keywords— regression discontinuity, nonparametric, small sample

1 Introduction

The regression discontinuity (RD) design is an approach to estimating the causal effect of an intervention using observational data when treatment assignment depends on whether or not some continuous variable (known as the running variable) falls above a cut-off. Because assignment to or eligibility for many educational interventions depends on a quantitative measure such as an assessment score, RD has gained popularity in this field since it was introduced by Thistlethwaite and Campbell (1960). For example, Abdulkadiroğlu et al. (2014) used RD to look at the effect that attending magnet schools with entrance exam admissions had on future educational outcomes. Lindo et al. (2010) examined the effect that GPA-based academic probation for university students had on future educational outcomes. RD designs can also be used to evaluate interventions at the school or district level. For example, many states impose sanctions on schools or districts that fail to exceed some accountability score cutoff. These educational applications often have features such as small sample sizes or sparsity around the cutoff that make RD estimation particularly challenging.
Under certain assumptions, individuals with running variable scores near the cutoff have essentially the same probability of having that score falling above the cutoff as below the cutoff (Lee & Lemieux, 2010). This idea of local randomization of treatment assignment motivates common RD modeling techniques. These techniques typically involve fitting separate regressions above and below the cutoff. These regressions yield estimates of the response for those with scores at the cutoff for both treatment and control groups. The difference in these fitted values is an estimate of the local average treatment effect (LATE).

Much of the recent methodological developments in RD designs have come from the field of economics, where sample sizes tend to be relatively large. In a survey of 110 RD studies in economics journals from 1999 to 2017, Pei et al. (2021) reports that the median sample size is 21,561. Only a third of papers have sample sizes of less than 6,000, and only three have sample sizes less than 500. Certainly some educational applications also involve large data sets. Abdulkadiroğlu et al. (2014) analyze subsets of data from around 6000, 11000, and 74000 students, and Lindo et al. (2010) have around 13000 students in their analysis. However, many educational applications involve much smaller sample sizes, particularly if an individual educator or institution wants to estimate the effect of a local intervention. In a meta-analysis of 11 studies with 21 independent samples looking at the effect of developmental classes on future educational outcomes of college students, Valentine et al. (2017) found a median sample size of around 1000, with the smallest sample having only 185 observations. Data collected at the school or district level can also result in relatively small sample sizes. For example, in this paper, we examine the effect of the threat of school sanctions from the state of Indiana on future school accountability scores (Indiana Department of Education, 2022). Less than 2000 schools are evaluated on these scores, and policy makers may be interested in subsets of these data, such as the set of all high schools, that have considerably smaller sample sizes. Popular simulation studies demonstrating the performance of many of the current popular RD estimation methods have typically used 500 as their only or smallest sample size (Armstrong & Kolesár, 2020; Calonico et al., 2014; Imbens & Kalyanaraman, 2012).

Related to sample size, an important consideration that has been neglected in much of the literature on RD designs is the location of the cutoff relative to the distribution of the running variable. In many educational applications, the cutoff is by design in the tail of the distribution of the running variable. The percent of students admitted to magnet schools is small (Abdulkadiroğlu et al., 2014), as is the percent of university students on probation (Lindo et al., 2010) and Indiana schools labeled as failing. The nonparametric regression techniques common for RD estimation give little or no weight to observations with running variable values far from the cutoff. Therefore, overall sample size may not be the best indicator of the RD-relevant size of a study.

This paper aims to provide practical advice for RD analysis of small data sets. First, we propose a simple density inclusive study size (DISS) metric that quantifies RD-relevant size better than overall sample size. Then we illustrate the effect of study and sample size on the operating characteristics of several leading methods of RD estimation via a simulation study. The RD methods we consider use local polynomial regression, a nonparametric approach that relies on a tuning parameter called a bandwidth. These methods differ in their bandwidth selection algorithms as well as their inferential techniques. We consider the bandwidth algorithms in Imbens and Kalyanaraman (2012), Calonico et al. (2014), Calonico, Cattaneo, and Farrell (2020) and Armstrong and Kolesár (2020). We consider the robust, bias corrected inference in Calonico et al. (2014) and the notion of honest inference in Armstrong and Kolesár (2020), as well as the conventional inference referred to in both papers.

The rest of the paper is organized as follows. Section 2 provides an overview of the RD estimation methods we consider. Section 3 introduces our proposed DISS metric. In Section 4 we present
the results of our simulation study. In Section 5 we apply the study size metric and the considered RD estimation methods to the Indiana school accountability data. Section 6 offers a discussion and our recommendations.

2 Popular RD Estimation Methods

We define $X$ to be the running variable, with a cutoff at $X = c$. We focus here on the “sharp” RD case, where the treatment $D$ is assigned solely based on whether or not the value of the running variable is at or above the cutoff, i.e., $D = 1_{\{X \geq c\}}$. Let $\mu(x) = \mu^-(x)1_{\{X < c\}} + \mu^+(x)1_{\{X \geq c\}} = E(Y|X)$ be the true underlying mean function, with $\mu^-(x)$ and $\mu^+(x)$ the mean function below and above the cutoff, respectively. The LATE of interest is $\tau = \mu^+(c) - \mu^-(c)$, which can be estimated by plugging in fitted values from separate regressions on either side of the cutoff to obtain $\hat{\tau} = \hat{\mu}^+(c) - \hat{\mu}^-(c)$.

Thistlethwaite and Campbell (1960) used global linear regression for their original RD estimation. To provide more flexibility, some researchers have used higher order global polynomials instead. However, in these higher order models, points far from the cutoff have large influence relative to points close to the cutoff (Gelman & Imbens, 2019), which is an undesirable property given the desire to minimize the bias of the estimated mean function at the cutoff. Thus local nonparametric techniques, and in particular the local polynomial regression technique used by the methods considered in this paper, may be superior. The estimated mean function in local polynomial regression is a weighted combination of the response values, using a weighting function called a kernel that relies on a bandwidth as a tuning parameter. See Fan and Gijbels (1996) for more details. The choice of kernel function from amongst those commonly employed, such as the triangular and Epanechnikov functions, typically makes little difference in the estimated mean function because they produce similar weights. Different bandwidth values, on the other hand, can lead to larger differences in weights and therefore larger differences in the estimated mean functions. Many data-driven bandwidth selection algorithms have been proposed in the literature, and these algorithms represent one of the key differences among the methods we consider. Since RD estimation typically involves separate regressions above and below the cutoff, some algorithms produce a separate bandwidth for each regression. The algorithms considered for this paper, however, select a single bandwidth for both regressions. Another key difference among the RD estimation methods in this paper is the inferential techniques employed by each. We address these two important features in turn.

2.1 Bandwidth Selection Algorithms

One popular bandwidth selection approach for local polynomial regression is to choose a bandwidth that optimizes an estimated objective function of interest. Imbens and Kalyanaraman (2012), henceforth IK, use as their objective function an asymptotic approximation of the mean squared error (AMSE) of their RD estimator. Minimization of their AMSE expression with respect to the bandwidth, $h$, leads to an infeasible optimal bandwidth $\hat{h}_{IK}$. They estimate the unknown quantities in this expression to yield a fully data-driven bandwidth

$$\hat{h}_{IK} = \left[\frac{1}{n} (C_K) \left( \frac{\hat{\sigma}_+^2(c) + \hat{\sigma}_-^2(c)}{\hat{\sigma}(c)(\hat{\mu}_+''(c) - \hat{\mu}_-''(c))^2 + \hat{\tau}(c)} \right) \right]^{1/5}$$

(1)

where $C_K$ is a function of the kernel, $\hat{\mu}_+''(c)$ and $\hat{\mu}_-''(c)$ are the estimated second-order one-sided derivatives of the underlying mean function above and below the cutoff, respectively, $\hat{\sigma}_+^2(c)$ and $\hat{\sigma}_-^2(c)$
\( \sigma^2(c) \) are the right and left hand limits, respectively, of the estimated error function \( \hat{\sigma}^2(x) = \text{Var}(Y|X) \) at the cutoff, and \( \hat{f}(x) \) is the estimated density of the running variable. IK also include a regularization term, \( \delta(c) \), that is not present in the infeasible expression, to help prevent similar estimated second derivatives from leading to extremely large bandwidth values.

Calonico et al. (2014), henceforth CCT, use the same AMSE and infeasible optimal bandwidth from IK as their starting point, although they generalize the expressions to allow for different model and inferential choices (see Section 2.2) and use a different regularized algorithm to calculate their data-driven bandwidth \( \hat{h}_{CCT} \).

Armstrong and Kolesár (2020), henceforth AK, perform RD inference based on the assumption that the mean functions on either side of the cutoff lie in a class with second derivatives globally bounded by \( M \). The AK approach effectively replaces \( (\hat{\mu}'_+(c) - \hat{\mu}'_-(c))^2 \) in \( h_{IK} \) with \( 4M^2 \) to obtain an infeasible optimal bandwidth

\[
\hat{h}_{AK} = \left[ \frac{1}{n} (C_K) \left( \frac{\sigma^2_+(c) + \sigma^2_-(c)}{4f(c)M^2} \right) \right]^{1/5}.
\]

Their data-driven bandwidth \( \hat{h}_{AK} \) does not incorporate regularization and estimates the bandwidth by directly minimizing the finite sample mean squared error (MSE). AK contend that the choice of \( M \) should be made \textit{a priori} to maintain the honesty of their confidence intervals (see section 2.2). However, they also present a data-driven estimate of \( M \), denoted \( \hat{M} \), which may be valid in some cases.

One notable alternative to these MSE optimal bandwidth algorithms is that of Calonico, Cattaneo, and Farrell (2020), henceforth CCF, who choose instead to optimize the asymptotic coverage error of robust bias corrected intervals. Their rule of thumb bandwidth,

\[
\hat{h}_{CCF} = n^{-p/(2(p+3)(p+3))} \hat{h}_{CCT},
\]

involves rescaling an MSE optimal bandwidth in order to achieve an optimal coverage error decay rate.

### 2.2 Inferential Procedures

Currently popular RD methods involve using the chosen bandwidth to fit local polynomial regressions to the left of the cutoff to estimate \( \mu^-(x) \) and separately to the right of the cutoff to estimate \( \mu^+(x) \). Recall that the difference in these estimated mean functions at the cutoff is the estimated LATE, \( \hat{\tau} \). The conventional (CV) inferential procedure pairs this point estimate with an estimate of the standard error of \( \hat{\tau} \) to calculate a confidence interval:

\[
I_{CV} = \hat{\tau}_{CV} \pm z_{\alpha/2} SE_{CV}(\hat{\tau})
\]

CCT and AK both use a nearest neighbors approach to calculate conventional standard errors, \( SE_{CV}(\hat{\tau}) \).

Two concerns with the conventional approach are non-negligible bias and empirical coverage well below the nominal level. CCT propose a robust bias-corrected (RBC) inferential procedure designed to address these concerns. They estimate the bias using higher order local polynomial regression and subtract that estimate, \( \hat{B} \), from the conventional point estimate. They also use a correction term \( C \) in their variance estimate to account for the extra variation that comes from bias estimation. This yields the interval

\[
I_{RBC} = \left( \hat{\tau}_{CV} - \hat{B} \right) \pm z_{\alpha/2} \sqrt{SE^2_{CV}(\hat{\tau}) + C} = \hat{\tau}_{RBC} \pm z_{\alpha/2} SE_{RBC}(\hat{\tau}).
\]
AK present an alternative approach to improving coverage based on the idea of honest confidence intervals. They simply inflate the critical value in the conventional interval (Equation 4) by using a critical value $z_{1-\alpha}(t)$ from a folded normal reference distribution with scale parameter 1 and shape parameter $t$ based on an estimated worst case scenario. This procedure produces what AK call a fixed length confidence interval (FLCI):

$$I_{FLCI} = \hat{\tau}_{CV} \pm z_{1-\alpha}(t)SE_{CV}(\hat{\tau}).$$

(6)

Despite the increased critical value, they show their intervals actually tend to be shorter than those produced by $I_{RBC}$ while maintaining proper coverage.

### 3 Study Size Metric

As with most statistical techniques, we expect the quality of bandwidth estimation and RD inference to degrade as sample size shrinks. For RD inference based on local polynomial regression, the distribution of the running variable relative to the cutoff is just as critical to inferential quality as the raw sample size. A small data set that has most of its observations around the cutoff may lead to better inference than a much larger data set that has most of its observations far from the cutoff. Thus, an RD-specific size metric would be much more useful than raw sample size for considerations such as initial feasibility assessments or choice of estimation method for RD-based analyses.

We propose a new density inclusive study size (DISS) metric that factors in the location of the cutoff and the density of the running variable near the cutoff. Kernels popular for RD local polynomial regression, such as the triangular and Epanechnikov kernels, give zero weight to observations more than one bandwidth value away from the cutoff. Thus we propose to quantify the size of RD data sets at a population or generating model level with $\bar{m}(n)$, the expected number of observations within one bandwidth of the cutoff for a given overall sample size $n$:

$$\bar{m}(n) = nP(c - h_{ROT}(s^*) < X < c + h_{ROT}(s^*)).$$

(7)

Rather than using one of the bandwidths from Section 2.1 that involve model assumptions and can differ substantially, we use a population-level version of the classic rule of thumb from Silverman (1986),

$$h_{ROT}(s^*) = 0.9(s^*)n^{-1/5}.$$  

(8)

Here $s^*$ is the minimum of the population IQR of the running variable divided by 1.34 and the population standard deviation of the running variable.

This population-level metric is useful if the distribution of the running variable is assumed or known, as in the case of our simulation study. When the distribution of the running variable is unknown, we propose a sample-level version which is simply the number of observations within one rule of thumb bandwidth of the cutoff,

$$m = \sum_{i=1}^{n} 1_{[c-h_{ROT}(s^*) \leq X_i \leq c+h_{ROT}(s^*)]}.$$  

(9)

Here for the Silverman bandwidth we have replaced $\sigma^*$ with $s^*$, which is the minimum of the sample IQR of the running variable divided by 1.34 and the sample standard deviation of the running variable.
We note that $h_{\text{ROT}}$ is based only on the running variable and does not factor in the values of the response, unlike the bandwidth algorithms for local polynomial regression. One benefit of this approach is that a researcher working with multiple response variables would not need to calculate a separate study size metric for each. Of course, if a researcher has a single response of interest and a certain technique in mind, they may choose to use the bandwidth algorithm from that technique rather than that of Silverman when quantifying study size. We show via simulation that the true population $\bar{m}$ explains RD estimator operating characteristics better than overall sample size.

4 Simulation and Results

We use an extensive Monte Carlo simulation study to explore the small-sample performance of common RD estimation methods and demonstrate the usefulness of our proposed DISS metric. Our simulation settings vary across five dimensions: bandwidth selection algorithm, inferential technique, running variable distribution, study size, and underlying mean function.

We consider the four data-driven bandwidth selection algorithms from Section 2.1, based on IK, AK, CCT, and CCF. Some of these algorithms can be generalized to a local polynomial regression degree $p$. When possible, we use degrees 0, 1, and 2 and denote the bandwidth-degree algorithm by concatenating the degree to the approach name. The IK algorithm is only available for local linear regression, while the AK algorithm is only available for degrees of at least 1. Furthermore, from Equation 3 it is clear that for local constant regression the CCF approach will yield the same value as the CCT approach, and thus CCF0 is omitted. This leaves a total of eight bandwidth algorithms: IK1, AK1, AK2, CCT0, CCT1, CCT2, CCF1, and CCF2.

For each bandwidth algorithm, point estimates and confidence intervals are constructed based on the three different inferential techniques from Section 2.2: CV, RBC, and FLCI. We choose the degree of the local polynomial for the inferential procedures to match that of the bandwidth algorithm, so CCT1/RBC performs local linear regression while CCT2/RBC performs local quadratic regression. The FLCI algorithm is not available for $p = 0$, so we do not consider CCT0/FLCI. Taking all other combinations of bandwidth algorithm and inferential technique yields 23 primary estimation methods under consideration. For the AK bandwidth algorithms and the FLCI technique we use the data-driven estimate $\hat{M}$ as we expect this may be more common in practice. In the supplemental appendix we present results from using the true value of $M$ calculated from the underlying mean function as a sensitivity analysis. All methods use the triangular kernel. We perform all calculations in R (R Core Team, 2020). We implement the IK, CCF, and CCF algorithms and the CV and RBC techniques using the R package \textit{rdrobust} (Calonico, Cattaneo, Farrell, & Titiunik, 2020). We implement the AK algorithm and FLCI technique using the R package \textit{RDHonest} (Kolesar, 2022).

To generate values $X_i$ of the running variable, we follow the simulation studies in IK and CCT in first independently drawing values $Z_i$ from a Beta distribution and then applying a transformation $X_i = 2Z_i - 1$. Thus the support of the transformed distribution of the running variable is $[-1,1]$. The cutoff is set to be $c = 0$. IK and CCT use the same Beta distribution for all design settings. However, we are interested in examining the effects of different levels of sparsity around the cutoff, and choose three different Beta distributions to aid in this analysis. The first running variable, denoted RV1, has a transformed Beta(1,1) distribution, and has half of its density on either side of the cutoff. This underlying distribution was used in the simulation study of AK. The second running variable, RV2, has the transformed Beta(2,4) distribution from IK and has approximately 19% of its density above the cutoff. The third running variable, RV3, has a transformed Beta(14,7) distribution and has less than 6% of its density below the cutoff. This distribution is modeled after
FIGURE 1. Running variable densities and underlying mean functions. The vertical gray line represents the cutoff.

TABLE 1.
Sample sizes and rule of thumb bandwidths for the different running variable distributions and values of \(\bar{m}\).

| \(\bar{m}\) | RV1  | n   | \(h_{ROT}(\sigma^*)\) | RV2  | n   | \(h_{ROT}(\sigma^*)\) | RV3  | n   | \(h_{ROT}(\sigma^*)\) |
|----------|------|-----|------------------------|------|-----|------------------------|------|-----|------------------------|
| 10       | 40   | 0.124 |                       | 56   | 0.143 |                       | 140  | 0.034 |                       |
| 20       | 96   | 0.104 |                       | 133  | 0.121 |                       | 337  | 0.028 |                       |
| 35       | 193  | 0.091 |                       | 267  | 0.105 |                       | 681  | 0.025 |                       |
| 50       | 301  | 0.083 |                       | 417  | 0.096 |                       | 1066 | 0.022 |                       |

the Indiana school accountability data discussed in Section 5. The upper left panel of Figure 1 depicts these densities.

To determine the sample sizes for our simulated data sets we start by choosing reasonable values of our population DISS \(\bar{m}\). We then calculate the sample size needed to achieve those values for each of the running variable distributions, which we report in Table 1. This allows us to compare data sets with similar numbers of observations near the cutoff.

We consider three different underlying mean functions \(\mu(x)\). The first, referred to as \(\mu_1\), is a modified version of the mean function from Design 3 in the simulation of AK. That simulation explored inference at a point and not regression discontinuity, so a jump at the cutoff has been added to the original mean function, which consists of quadratic splines with knots away from
the cutoff. Thus the second derivatives immediately on either side of the cutoff are equal. The second derivative bound of this function is \( M = 6 \). The second mean function, \( \mu_2 \), comes from Design 3 in the simulation of IK and is based on a modification of the data from Lee (2008). It consists of quintic polynomials on either side of the cutoff that differ only in their intercept. This function has more curvature than \( \mu_1 \), particularly below the cutoff, with a second derivative bound of \( M = 233.26 \). The third mean function, \( \mu_3 \), is based on the Indiana data described in Section 5. It consists of separate cubic polynomials on either side of the cutoff, with differing second derivatives and a bound of \( M = 16.2 \). All three mean functions have a vertical discontinuity of 0.1 at the cutoff and are continuous elsewhere. The equations of the three mean functions are

\[
\begin{align*}
\mu_1(x) &= (x + 1)^2 - 2s(x + 0.2) + 2s(x - 0.2) - 2s(x - 0.4) + s(x - 0.7) - 0.92 \\
& \quad + (0.1)1_{[x \geq 0]} \\
\mu_2(x) &= 0.42 + 0.84x - 3.0x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 + (0.1)1_{[x \geq 0]} \\
\mu_3(x) &= (0.05 + 1.5x + 3.2x^2 + 2.7x^3)1_{[x < 0]} + (0.15 - 0.15x + 2.5x^2 - 1.5x^3)1_{[x \geq 0]}
\end{align*}
\]

where \( s(x) = (x)^2_+ = \max\{x, 0\}^2 \) is the square of the plus function. Graphs of these functions are included in Figure 1.

We consider all pairs of a running variable density and a mean function, which gives us nine data generating processes (DGP). We denote these by concatenating the designations for the running variable and mean function, e.g., RV1\( \mu_1 \). The response values \( Y_i \) are generated from the values of the running variable \( X_i \) as \( Y_i = \mu(X_i) + \epsilon_i \), where \( \epsilon_i \sim N(0, .1295^2) \). This error distribution was also used in the IK simulation. We generate 50,000 simulated data sets for each DGP at each value of \( \bar{m} \).

We analyze the performance of the methods in four categories. For small sample sizes, bandwidth algorithms and inferential techniques do not always produce finite estimates, so an important consideration in evaluating methods of RD estimation is simply how often those methods work. Thus in Section 4.1, we look at bandwidth success rates, the percentage of repetitions in which an algorithm produces a finite bandwidth value. In this section we also examine the distributions of the calculated bandwidth values, as the size of the bandwidth plays an important role in the subsequent operating characteristics. Second, in Section 4.2 we look at the interval estimate success rate, the percentage of repetitions where a method produces a finite confidence interval. Third, in Section 4.3 we evaluate the performance of finite point estimates using the bias, empirical standard error (EmpSE), and mean squared error (MSE). Fourth, in Section 4.4, we evaluate interval estimates using the median interval width and empirical coverage. Throughout, we use the RV2\( \mu_2 \) DGP as a baseline to describe typical patterns across methods, then highlight deviations across the various DGPs.

### 4.1 Bandwidth Calculation

One concern for RD estimation with small sample sizes is the ability to produce finite estimates. To produce a finite estimate a method first must produce a finite bandwidth. Thus we begin our analysis by looking at the success rate of the included bandwidth algorithms and the distributions of the bandwidths they produce. We omit the CCF algorithm from this analysis as it produces bandwidth values that differ only from the CCT bandwidth values in scale. The square symbols in Figure 2 give the percentage of simulated RV2\( \mu_2 \) data sets for which each bandwidth algorithm produces a finite value across different study sizes. The IK and AK algorithms achieve near universal success for all study sizes considered (and all DGPs; see Supplemental Appendix Figures SA1-SA3).
FIGURE 2. *Bandwidth success rates (red) and distributions (black) for RV2μ2.*

But the CCT algorithm starts to break down for the very small study size of \( \bar{m} = 10 \). For this study size the CCT algorithm also has more difficulty as the degree of the local polynomial (and thus the order of the estimated mean functions) increases. These trends continue for other DGPs (see Supplemental Appendix Figures SA1-SA6.)

The distribution of the running variable also plays a role in the CCT bandwidth algorithm success rates for the smallest sized data. Figure 3 gives the success rates of the CCT1 algorithm across all nine DGPs when \( \bar{m} = 10 \). The success rates decrease from RV1 to RV2 to RV3 despite the fact that the overall sample size increases across these distributions. The CCT algorithm estimates quantities separately on either side of the cutoff and thus may fail if there are too few observations on one side. The relative frequency of observations on the sparse side of the cutoff decreases from RV1 to RV2 to RV3 giving the algorithm less data with which to calculate a bandwidth. This trend holds for the other CCT algorithms with \( \bar{m} = 10 \) and across all DGPs (see Supplemental Appendix Figures SA4-SA6.)

The bandwidth values themselves are important in understanding the the performance of the estimation methods as we will see. Figures 2-3 include the distributions of all finite bandwidth values produced by the given algorithms. (This means that the distributions have different numbers of observations due to differing success rates.) In Figure 2, we see that regardless of study size \( \bar{m} \), the IK algorithm tends to produce the largest median bandwidth values among those optimized for local linear regression \( (p = 1) \), followed by the CCT and then the AK algorithms. Within algorithm type, both median bandwidth size and variability among bandwidth values tends to increase with the regression degree \( p \).

We are also interested in the relationship between study size and bandwidth distribution. As
study size increases, the bandwidth distributions for each algorithm naturally become less variable and have fewer outliers. Given the $n$ in the denominator of Equations 1 and 2 and the direct relationship between $\bar{m}$ and $n$, we also expected an inverse relationship between study size and median bandwidth value. This inverse relationship is present for each of IK1, AK1, AK2, and CCT0, but CCT1 and CCT2 produce surprising results. The approximate CCT1 median bandwidth remains roughly constant across study sizes, while the CCT2 median slightly but consistently increases until it is higher than even the IK1 algorithm for larger study sizes. Across DGP, these trends relative to study size are consistent for each bandwidth, although relative magnitudes across bandwidth algorithm may prevent the CC2/IK1 inversion we see here (see Supplemental Appendix Figures SA1-SA6). So, the larger sample sizes that come from increasing $\bar{m}$ do not always lead to smaller median bandwidth values the way we might have predicted for a fixed running variable distribution.

Alternatively, if we fix $\bar{m}$ and allow the running variable distribution to vary, an increase in sample size does consistently lead to decreased median bandwidth values. We can see this by comparing boxplots within panels in Figure 3 for CCT1 (and in Supplemental Appendix Figures SA1-SA6 for other DGPs and values of $\bar{m}$). Each panel has the same $\bar{m}$ value but the overall sample sizes increase from left to right as the more unbalanced running variable distributions require larger sample sizes to achieve the same number of points near the cutoff. For the CCT algorithm, the differing relationships between sample size and median bandwidth value depending on whether we fix RV or $\bar{m}$ may indicate instability in the plug-in estimates of quantities in the infeasible bandwidth expression for small sample sizes.
### Table 2.
Interval estimate success rates (in percents) for estimation methods at various study sizes for RV2μ2.

|       | \( \bar{m} = 10 \) |       | \( \bar{m} = 20 \) |       | \( \bar{m} = 35 \) |
|-------|----------------------|-------|----------------------|-------|----------------------|
|       | RBC                  | CV    | FLCI                 | RBC    | CV    | FLCI                 | RBC    | CV    | FLCI                 |
| IK1   | 97.31                | 97.31  | 91.06                | 100.00 | 100.00 | 99.98               | 100.00 | 100.00 | 100.00               |
| AK1   | 68.02                | 68.22  | 13.37                | 96.31  | 96.32  | 75.36               | 99.92  | 99.92  | 99.06               |
| AK2   | 86.68                | 86.76  | 74.88                | 99.71  | 99.72  | 99.41               | 100.00 | 100.00 | 100.00               |
| CCT0  | 67.30                | 67.30  | NA                   | 95.07  | 95.07  | NA                  | 99.63  | 99.63  | NA                  |
| CCT1  | 65.50                | 65.47  | 36.83                | 99.34  | 99.34  | 97.09               | 100.00 | 100.00 | 99.99               |
| CCT2  | 55.64                | 55.70  | 43.39                | 99.49  | 99.48  | 98.86               | 100.00 | 100.00 | 100.00               |
| CCF1  | 59.11                | 59.11  | 26.01                | 98.49  | 98.50  | 92.65               | 99.99  | 99.99  | 99.94               |
| CCF2  | 47.86                | 47.91  | 32.65                | 98.67  | 98.67  | 96.69               | 100.00 | 100.00 | 100.00               |

#### 4.2 Interval Estimate Success Rates

Even for those iterations where a finite bandwidth is calculated, there is no guarantee that RD estimation methods will produce finite point or interval estimates. All of the considered methods rely on fitting separate local polynomial regressions on either side of the cutoff, which can be impossible if there is not enough data “close” to the cutoff, as quantified by the bandwidth value. (Calculating the bias estimates for the RBC technique and the standard errors for all methods can be even more challenging.) Thus, methods using algorithms that tend to produce larger bandwidths, such as IK1, tend to have higher success rates. Table 2 gives these rates, defined as the percentage of simulated data sets that produce finite bandwidth values and interval estimates. Methods using the IK1 algorithm all have at least a 91% success rate when \( \bar{m} = 10 \), but the success rates of methods using the other algorithms are quite a bit lower. Success rates are higher across the board for \( \bar{m} = 20 \), but the methods most likely to fail are still those using algorithms with relatively small bandwidths such as AK1, CCT0, and CCF1. For \( \bar{m} = 35 \) and \( \bar{m} = 50 \) (see Supplemental Appendix Table SA2 for the latter) there is near universal success, indicating that with enough data even methods using relatively small bandwidth values can still produce finite interval estimates. These patterns hold for all DGPs as seen in Tables SA1-SA3 in the supplemental appendix.

Methods employing FLCI inference have much lower success rates than those using the other inferential techniques. This may indicate that it is difficult to estimate the worst case scenario needed for the inflated critical value in Equation 6. This seems to be particularly true for relatively small bandwidth values, as the AK1/FLCI method has a success rate of only 13% for \( \bar{m} = 10 \) and 75% for \( \bar{m} = 20 \).

#### 4.3 Point Estimation

Comparing methods in terms of point estimation (and later interval estimation) is made more challenging by the differing success rates. To provide a fair comparison, we summarize only iterations that produce finite interval estimates for all methods for a particular DGP and value of \( \bar{m} \). This restriction is most problematic for \( \bar{m} = 10 \), where the vast majority of iterations are not considered (recall the 13% success rate for AK1/FLCI), but for larger study sizes we are less concerned about any differences in analysis due to this restriction. Note that the CV and FLCI point
estimates are theoretically the same, and while there are minute differences that exist between the results due to algorithmic implementations, we choose to present only the CV results here.

Bias is certainly a concern for small sample RD estimation. Most of the methods have an absolute relative bias of more than 5% for \( \mu_2 \) when \( \bar{m} = 20 \), although there is less bias when \( \bar{m} = 50 \). These bias values are given in Figure 4 along with EmpSE and MSE values. For \( \bar{m} = 20 \), the RBC methods are effective at reducing the bias for most, but not all, of the algorithms. The cost of the attempted bias correction is larger empirical standard errors for all bandwidth algorithms. This allows the CV/FLCI methods to have lower MSE values than the RBC methods for all algorithms in this DGP. In particular, RBC methods that use the AK1 and AK2 algorithms, which tend to produce small bandwidth values, have very high EmpSE and thus very high MSE. This may indicate that for small study sizes, the extra estimation needed to do bias correction may not be worth it, especially when paired with relatively small bandwidths.

We also see here that methods that use quadratic regression tend to be quite variable and thus do not perform well in terms of MSE for small study sizes. Again, the extra parameter estimation seems to result in worse performance when there are insufficient data close to the cutoff. Methods with small MSE values in this setting include those using IK1, CCT0, and CCT1, particularly when paired with CV/FLCI inference.

RBC methods are somewhat more effective at reducing bias for larger study sizes, but still lag behind the CV/FLCI methods in terms of EmpSE and MSE when \( \bar{m} = 50 \). Methods using an AK algorithm continue to perform poorly at this study size. Methods using CCT2 or CCF2 are more competitive in terms of MSE, indicating that for larger study sizes quadratic regression may be a more reasonable option. However methods pairing IK1, CCT0, and CCT1 with CV/FLCI tend to have the lowest MSE values for all study sizes considered (see Supplemental Appendix Figure SA7).

The tendency of methods using CV/FLCI to dominate those using RBC in terms of MSE extends to the other considered DGPs, as seen in Figure 5, and is again due to large EmpSE values for methods using RBC not being mitigated by enough bias reduction (see Supplemental Appendix Figures SA8-SA9). Likewise methods using CCT2 and CCF2 continue to perform poorly in terms of MSE due primarily to relatively high EmpSE, while methods using IK1, CCT0, and CCT1 continue to perform well. A notable difference is that the method using AK1 paired with CV/FLCI is among the leaders in terms of MSE for several of the other DGPs, particularly in situations where the median AK1 bandwidth value is relatively closer to those from the other algorithms, such as RV3.

If we allow both the running variable distribution and the value of \( \bar{m} \) to vary, we can obtain a comparison of different DGPs with similar overall sample sizes but differing number of observations close to the cutoff. For example, RV1 with \( \bar{m} = 50 \) has \( n = 301 \), while RV3 with \( \bar{m} = 20 \) has \( n = 337 \). In Figure SA10 in the Supplemental Appendix we see that the MSE values for RV1/\( \bar{m} = 50 \) tend to be lower than the MSE values for RV3/\( \bar{m} = 20 \), despite the former having more than 30 fewer observations. These gains tend to come from decreased EmpSE that make up for estimates that are often more biased.

### 4.4 Interval Estimation

To compare interval estimation we look at the empirical coverage for the nominally 95% confidence intervals, as well as the median interval width. By construction the CV intervals will have the smallest median widths for a given bandwidth value, discounting minor differences in algorithmic implementation, since both the RBC and FLCI techniques widen the intervals. The FLCI technique uses an inflated critical value but is centered around the same point estimate as the CV interval, resulting in higher empirical coverage. The RBC technique adds a term to the standard
FIGURE 4. Performance measures for methods using CV and RBC inference for RV2μ2 at two study sizes. FLCI values are essentially the same as CV and thus omitted. Values based on iterations in which all estimates were finite, 66.8% of all iterations for \( \bar{m} = 20 \) and 99.9% for \( \bar{m} = 50 \). The graph with \( \bar{m} = 20 \) omits EmpSE and MSE values for RBC/AK1 (1.07 and 1.15, respectively), AK2/RBC (1.43, 2.02), CCF2/RBC (0.70, 0.49), and CCF2/CV (0.69, 0.48). For the bias, all Monte Carlo standard errors (MCSE) were less than 0.004 except for AK2/RBC (0.008) and AK1/RBC (0.006). For the EmpSE, all values shown had MCSE less than 0.002. For the MSE, all values shown had MCSE less than 0.0006 except CCT2/RBC (0.04) and CCT2/CV (0.04).

error to account for the variability in the bias estimation, but since it is centered in a different place it may have higher or lower empirical coverage than the other techniques. Thus the top performing methods should provide reasonable coverage with the smallest possible interval widths.

For RV2μ2 at all study sizes, the intervals typically produced by many of the methods using FLCI are so wide that these methods may be undesirable despite their high coverage, as seen in Figure 6. However, the AK1/FLCI method yields the smallest median interval width from among the FLCI methods and still has coverage above the nominal rate, making it a good overall choice. Methods using CV and RBC are more similar to each other, with RBC methods yielding slightly larger median widths. For the smaller study sizes this coincides with slightly higher coverage for RBC methods compared to those of CV, but for the larger study sizes most CV and RBC methods have similar coverage. Thus the RBC methods fail to significantly improve upon the CV methods.
FIGURE 5. MSE of point estimates for different methods and designs at $\bar{m} = 20$. FLCI values are essentially the same as CV and thus omitted. Values based on iterations in which all estimates were finite. The proportion of finite iterations for each DGP is given in the top right corner of the graph. The RV1 graph omits values for AK1/RBC (65.2) and AK2/RBC (3.81). The RV1 graph omits CCF2 values for $\mu_1$ (0.84 for RBC, 0.84 for CV), $\mu_2$ (0.49, 0.48), and $\mu_3$ (1.02, 1.01) as well as AK2/RBC (2.02) and AK1/RBC (1.15) for $\mu_2$. The RV3 graph omits CCF2 values for $\mu_1$ (1.41 for RBC, 1.40 for CV), $\mu_2$ (1.38, 1.37), and $\mu_3$ (1.42, 1.41). Of the values shown for RV1, all MCSE are below 0.004 except RV1 for RBC and CV. Of the values shown for RV2, all MCSE are below 0.002 except for CCT2 (0.03 for $\mu_1$, 0.04 for $\mu_2$, 0.03 for $\mu_3$ for both RBC and CV) and AK1/RBC (0.01 for $\mu_1$ and $\mu_3$). Of the values shown for RV3, all MCSE are below 0.002 except for CCT2 (0.10 for $\mu_1$, 0.01 for $\mu_2$, 0.01 for $\mu_3$ for both RBC and CV), and AK1/RBC and AK2/RBC (0.01 for $\mu_1$, $\mu_2$, and $\mu_3$).

in our study. Perhaps the best performing of the CV methods is the one that uses IK1, which yields quite narrow intervals and yet has reasonable coverage at larger study sizes.
FIGURE 6. Empirical coverage and median interval width for all methods and study sizes with RV2\(\mu_2\). The maximum Monte Carlo standard errors for the coverage estimates are 0.008 (\(\bar{m} = 10\)), 0.0019 (\(\bar{m} = 20\)), 0.0016 (\(\bar{m} = 35\)), and 0.0015 (\(\bar{m} = 50\)).

TABLE 3. Median \(\hat{M}\) values for different DGPs. True values of \(M\) provided for comparison.

| Truth | RV1  | RV2  | RV3  | RV1  | RV2  | RV3  |
|-------|------|------|------|------|------|------|
| \(\mu_1\) | 6.0  | 23.0 | 13.8 | 10.0 | 8.3  | 90.2 | 34.4 | 21.1 | 16.3 | 749.5 | 210.1 | 109.7 | 76.6 |
| \(\mu_2\) | 233.3 | 210.5 | 219.6 | 223.3 | 224.7 | 203.4 | 206.1 | 211.9 | 214.8 | 748.1 | 212.3 | 114.6 | 85.1 |
| \(\mu_3\) | 16.2 | 25.5 | 17.6 | 14.4 | 13.0 | 90.0 | 34.3 | 21.5 | 17.2 | 749.6 | 210.2 | 109.5 | 76.6 |

The critical values used by the FLCI technique are based on a worst case estimate of the ratio between the bias and the standard deviation of the effect estimate and increase as the estimated second derivative bound \(\hat{M}\) increases. Intuitively, a scenario with a large bandwidth and large estimated curvature will have a large worst case bias estimate. Table 3 shows that the estimated curvature for \(\mu_2\) is quite high, and when combined with larger bandwidth values such as those produced by IK1 yields the very wide intervals we see in Figure 6.

The estimated curvature for the other mean functions is much lower than for \(\mu_2\), at least in the case of RV1 and RV2, and the median widths of FLCI intervals are more in line with (though still typically larger than) those produced by the other methods for \(\bar{m} = 20\), as seen in Figure 7. For RV3 the \(\hat{M}\) values are similar across mean functions, possibly because the mean functions are somewhat similar above the cutoff where the vast majority of the RV3 density is. Thus the median FLCI interval widths are also similar across mean functions for RV3, and are still typically somewhat larger than the other methods.

The relative performance of methods using RBC and CV remains essentially the same for the
Figure 7. Empirical coverage and median interval width at $\bar{m} = 20$ for all DGPs. Values based on iterations in which all estimates were finite (see Figure 5). The graph omits values of CCT0/CV, which have coverages lower than 0.9 and widths between 0.3 and 0.43, as well as values of IK1/FLCI, which have coverages greater than 0.999 and widths between 1.3 and 3.9. The maximum Monte Carlo standard errors for the given coverage estimates are 0.0028 for RV1, 0.0019 for RV2, and 0.0015 for RV3.

DGPs pictured in Figure 7. The RBC intervals are typically wider but do not improve upon coverage enough to justify the extra width. The IK1/CV method continues to yield narrow intervals and competitive coverage, and along with AK1/FLCI is one of the top performing methods in terms of the width-coverage tradeoff. Figure SA11 in the supplemental appendix looks at the relationship between coverage and interval width for the same DGPs for the larger study size of $\bar{m} = 50$, and we see similar patterns to those that exist for the smaller study sizes. Figure SA12 in the supplemental appendix compares performance between the RV1/$\bar{m} = 50$ and RV3/$\bar{m} = 20$ DGPs, and we see that the former, with more observations close to the cutoff, typically yields smaller median interval widths but similar (and in some cases larger) coverage than the latter, which has a larger overall sample size.
TABLE 4.
Overall sample size, sample size below the cutoff, Silverman rule of thumb bandwidth, number of observations within $h_{\text{ROT}}$ of the cutoff, and estimated second derivative bounds for the original and scaled data.

| Subset                      | Sample Size | Sample Size Below | $h_{\text{ROT}}(s^*)$ | $m$   | $\hat{M}$ | $\hat{M}_S$ |
|-----------------------------|-------------|-------------------|--------------------|-------|-----------|-------------|
| All                         | 1933        | 88                | 2.31              | 51    | 2.173     | 130.4       |
| Low                         | 1676        | 86                | 2.42              | 51    | 2.141     | 128.5       |
| Traditional Public          | 1626        | 77                | 2.30              | 42    | 1.980     | 118.8       |
| Traditional Public, Low     | 1391        | 75                | 2.39              | 43    | 1.959     | 117.6       |
| High                        | 408         | 7                 | 1.88              | 4     | 5.080     | 304.8       |
| Private                     | 307         | 11                | 3.60              | 10    | 1.725     | 103.5       |

5 Real Data Example

In 2015 the state of Indiana revamped their K-12 school accountability system (Indiana Department of Education, 2022). The new system, which began with the year 2015-2016, gives each school an accountability score between 0 and 120 based on a various performance and growth measures. Those scores are then translated into A-F letter grade categories. A school that scores below a 60 is automatically given an F. The state imposes various consequences if a school receives an F for a certain number of years. These consequences vary based on the type of school and the number of consecutive years the school has received an F.

This situation can be thought of as a sharp regression discontinuity design. We choose the running variable to be the 2017 school accountability scores (Indiana Department of Education, 2017) and the response variable to be the 2018 school accountability scores (Indiana Department of Education, 2018). These are the first two years after the new system was implemented that full data are available. Note that the state does not implement sanctions based on just one year of earning a failing grade, so the treatment can be thought of as the threat of sanctions or the lack of such a threat that occurs when a school is given a failing grade, rather than the sanctions themselves. Schools with 2017 scores greater than the cutoff do not receive the threat of sanctions, so to be consistent with the definition of $\hat{\tau}$ in Section 2.2 the LATE estimated in this section will be the effect of not receiving the threat of sanctions. Thus a negative value of this LATE would indicate that the threat of sanctions is improving scores, which is presumably what the state of Indiana would like to be the case.

Indiana schools are categorized according to type and grade level for the purpose of the accountability system. School types include traditional public, charter, and choice schools. The latter two categories will be grouped together here as private schools for the sake of convenience, even though there is differing opinion as to how charter schools should be categorized. In terms of grade level, low schools contain at least some of the K-8 grades, while high schools contain at least some of the 9-12 grades. Note that a K-12 school is considered part of both categories, so there is some overlap. Subsets of the overall set of Indiana schools can be formed based on combinations of school type and grade level. These subsets may be of interest as it is possible that the effect of the threat of sanctions may be different across subsets. These subsets also highlight differences in sample sizes and sparsity around the cutoff that is relevant to our discussion of RD LATE estimates for small samples. Table 4 contains the overall sample sizes, sample sizes below the cutoff, and values of $h_{\text{ROT}}$ and $m$ for the overall data set and selected subsets.
In 2017 only 88 schools in the data set were labeled as failing, representing less than five percent of the total of 1933 schools. This percentage is even smaller for some of the subsets of interest. This data set provides a good example of an education application with a cutoff in the tail of the running variable distribution, and thus care must be taken when considering the overall sample size. In the group of all Indiana schools, only 51 are within a Silverman bandwidth of the cutoff, and yet the sample size of 1933 is much larger than those DGPs in Table 1 with $\bar{m} = 50$. The subset of high schools is larger than that of private schools, but the latter has more values below the cutoff and more values within a rule of thumb bandwidth of the cutoff, providing a useful test case to compare our DISS to overall sample size.

We use the methods of Section 2 to obtain interval estimates for the RD treatment effect for each of the above subsets. We choose again to use the data driven choice of $\hat{M}$ for our second derivative bound for the AK bandwidth algorithms and FLCI inference. We are using data from a new accountability system and thus do not have the luxury of historical data that might give us a different value of $M$ to use instead. The values of $\hat{M}$ for each subset of interest are also recorded in Table 4. For most of the subsets the maximum second derivative is estimated to be around 2, with the subset of high schools being the exception with a value above 5. Note however that the scale here is different than in the simulations. To facilitate a comparison, we also divide the running and response variables by 60 and then subtract 1, which scales the support of the variables to resemble those of the RV3_{μ3} DGP. This scaling has the effect of multiplying the estimated derivative bounds by 60, and these are presented in the table as $\hat{M}_S$. For the overall set of schools as well as the subset of low schools and the subset of traditional public schools, the AK2 algorithm as implemented in the RDHonest package is unable to calculate bandwidths using the original scaling. In those cases the data is scaled in the manner above and a bandwidth is calculated on the scaled data, and then that bandwidth value is scaled back to get an approximate bandwidth for the original data. When this process was performed on the algorithms that produced bandwidth values with the original scaling, the two sets of bandwidths were essentially the same.

The IK1 bandwidth values are the highest for each of the considered subsets by a substantial margin, while the CCT0 and AK1 bandwidths are the smallest, as seen in Figure 8. This pattern is consistent with that in the simulation study. Note that the CCT1 and CCT2 algorithms are unable to produce a bandwidth for the set of high schools, but all algorithms produce bandwidths for the set of private schools. There are 101 more high schools than private schools, but the set of private schools has a larger value of $m$. These extra observations near the cutoff likely play a large role in the success of the bandwidth algorithms.

The different subsets allow us to see similar patterns to those in the simulation study involving sample sizes and values of $\bar{m}$. An important caveat is that while we could fix the underlying mean function in the simulation, the different subsets here may have underlying mean functions that differ. However, we would expect those differences to be small due to the overlap between subsets. The subsets of all schools and low schools have the same values of $m$ but differ in overall sample size by more than 250, similar to the within panel comparison from Figure 3. In the simulation study the DGP with the largest overall sample size in each panel had the smallest median bandwidth value. The all and low subsets are consistent with this pattern except in the case of the CCT0 algorithm. The bandwidth value differences are much smaller than we saw in the simulation, but the sample sizes are larger, which may contribute to these smaller differences along with possible differences in the underlying mean function. The subset of traditional public schools has a smaller overall sample size and fewer points around the cutoff than the overall set of Indiana schools, similar to the across panel comparison in Figure 3. In the simulation study the DGP with the larger sample size in this situation did not necessarily have the smaller bandwidth. We see that pattern here with the CCT2
bandwidth value being larger for the overall set compared to the traditional public school subset, although the opposite is true for the other algorithms.

Figure 9 gives estimated treatment effects and confidence intervals for three of the subsets of Indiana schools. The relationships between method and interval width mostly mimic those of the simulation study. The FLCI intervals tend to be wider than the RBC intervals, except when using the AK algorithms. The FLCI/IK1 method produces extremely wide intervals. The width of the interval tends to increase slightly as the local polynomial degree increases. These patterns differ little across subsets, which is also consistent with the simulation results. Figure SA13 in the supplemental appendix shows similar results for the other Indiana subsets of interest.

The results show mostly statistically insignificant treatment effects. For the set of all Indiana schools, the values of $\hat{\tau}$ are positive for each method, but only significantly so for the CV/CCT0 method. The latter would show some evidence that being labelled a failing school actually decreases future performance measures. For traditional public schools the values of $\hat{\tau}$ tend to be slightly larger, with CV/CCT0, CV/CCF2, RB/CCF2, and CV/AK1 all producing intervals with positive endpoints. If we switched to 90% intervals, which may be more consistent with educational applications with small sample sizes, nearly all of the methods would produce statistically significant treatment effect estimates, but in the opposite direction than the state would likely intend. For private schools, the values of $\hat{\tau}$ are smaller and even negative in a few cases, with no statistically significant results. These results may indicate differences in the effect of the treatment among different subgroups, but as no comparative methods were used we must be careful not to overstate this case. We failed to see a statistically significant negative effect across the six subsets. Thus our preliminary analysis indicates that the threat of sanctions does not seem to be an effective way to improve school accountability scores. However a more thorough analysis, including checks for model fit, may come to a different conclusion.

6 Discussion

Regression discontinuity designs will continue to be a popular way to estimate causal effects for educational interventions thanks to the prevalence of cutoff-based programs. Many of the samples used in real world applications may be somewhat small or have sparsity around the cutoff, and these characteristics may help determine the type of methodology to use. Rather than using the overall
FIGURE 9. Treatment effect estimates and 95% confidence intervals for select Indiana subsets.
sample size to guide this process, we recommend using our proposed DISS metric, which quantifies the number of observations near the cutoff. In our simulation study and real data example, we showed that \( \bar{m} \) was a more useful metric than sample size \( n \) in determining whether RD estimates could be obtained from a data set. Furthermore, \( \bar{m} \) is useful in understanding trends in the size of calculated bandwidth values, which in turn affects the interval estimate success rates as well as the performance of point and interval estimates.

We also used our simulation study to compare the performance of popular RD methods when applied to small data sets. Certain bandwidth algorithms performed poorly almost no matter which inferential technique they were paired with. The CCT0 algorithm tended to produce the smallest bandwidths values, which can be a problem when there are few points near the cutoff. These small bandwidth values likely contributed to relatively low interval success rates, and the CV/CCT0 method in particular had very poor coverage. The methods using the CCT2 and CCF2 bandwidth algorithms also had relatively low success rates, large MSE values, and wide intervals that didn’t result in superior coverage. The AK2 bandwidth algorithm had some of the same issues with width/coverage. Using local quadratic regression rather than local linear regression might produce better results at large study sizes (see Pei et al. (2021)), but at the small study sizes considered here the extra estimation it entails does not appear to be worth it. Thus for the methods under consideration, using local linear regression for the bandwidth and inferential techniques seemed to have the overall best operating characteristics.

In terms of inferential techniques, the methods using RBC seemed to suffer from some of the same problems as the methods using local quadratic regression. Bias correction takes extra estimation, and if there is not enough data around the cutoff this extra estimation can lead to much higher variability. Thus the RBC methods had larger MSE values than those using CV/FLCI in nearly all DGPs. The RBC methods also tended to produce larger intervals, but not always better coverage. As the study sizes increased, RBC methods tended to improve in terms of point estimation, but not as much in terms of interval estimation.

The AK bandwidth algorithms and the FLCI inferential technique are perhaps the hardest to evaluate, as their performance varied the most across DGP. The AK1/FLCI method had a very low success rate in some settings, but in others was much more competitive with the other methods. The FLCI technique did lead to improved coverage but often at the cost of much wider intervals. The interval widths stem in part from using the worst-case bias-sd ratio, which can be rather large when samples are small and curvature is large. In many situations the AK1/FLCI method would be considered one of the top methods when considering both point and interval estimation, but its low success rate for very small study sizes is somewhat concerning.

In addition to AK1/FLCI, another method that consistently performed well was the IK1/CV method. The IK1 bandwidth algorithm tended to produce relatively large bandwidth values, which helped the IK1/CV and IK1/RBC to have the highest success rates for each DGP. Perhaps when studies are very small, it may be better to have an algorithm that produces relatively large bandwidths and hope that the increased bias can be compensated for with decreased variability, as was often the case in our simulation results.

In some ways then, our simulation results highlight the difficulty of small study RD estimation. The conventional approach is far from perfect, and in particular leads to lower than desired empirical coverage. But attempts to improve this coverage using the bias correction of CCT or the fixed length confidence intervals of AK have only inconsistent success and often come at a cost of other performance measures. Clearly there is more work to be done in developing RD methodology that works well for small studies.
References

Abdulkadiro˘ glu, A., Angrist, J., & Pathak, P. (2014). The elite illusion: Achievement effects at Boston and New York exam schools. *Econometrica, 82*(1), 137–196. https://doi.org/10.3982/ECTA10266

Armstrong, T. B., & Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics, 11*(1), 1–39. https://doi.org/10.3982/QE1199

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal, 23*(2), 192–210. https://doi.org/10.1093/ectj/utz022

Calonico, S., Cattaneo, M. D., Farrell, M. H., & Titiunik, R. (2020). *Rdrobust: Robust data-driven statistical inference in regression-discontinuity designs* [R package version 0.99.6]. https://CRAN.R-project.org/package=rdrobust

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica, 82*(6), 2295–2326. https://doi.org/10.3982/ECTA11757

Fan, J., & Gijbels, I. (1996). Local polynomial modelling and its applications.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics, 37*(3), 447–456. https://doi.org/10.1080/07350015.2017.1366909

Imbens, G., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies, 79*(3), 933–959. https://doi.org/10.1093/restud/rdr043

Indiana Department of Education. (2017). 2017 A-F school grade reports [Data file] [Retrieved May 27, 2022 from https://www.in.gov/doe/files/af-2017-school-grade-placement-03192018.xlsx].

Indiana Department of Education. (2018). 2018 state school grade reports [Data file] [Retrieved May 27, 2022 from https://www.in.gov/doe/files/2018-state-school-grade-results.xlsx].

Indiana Department of Education. (2022). Indiana Student-Centered Accountability [Retrieved August 18, 2022 from https://www.in.gov/doe/it/accountability-dashboard/indiana-student-centered-accountability/].

Kolesar, M. (2022). *Rdhonest: Honest inference in regression discontinuity designs* [R package version 0.3.2]. https://github.com/kolesarm/RDHonest

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. house elections. *Journal of Econometrics, 142*(2), 675–697. https://doi.org/10.1016/j.jeconom.2007.05.004

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature, 48*(2), 281–355. https://doi.org/10.1257/jel.48.2.281

Lindo, J. M., Sanders, N. J., & Oreopoulos, P. (2010). Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics, 2*(2), 95–117. https://doi.org/10.1257/app.2.2.95
Pei, Z., Lee, D. S., Card, D., & Weber, A. (2021). Local polynomial order in regression discontinuity designs. *Journal of Business & Economic Statistics*. https://doi.org/10.1080/07350015.2021.1920961

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Taylor & Francis. https://doi.org/10.1201/9781315140919

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology, 51*(6), 309–317. https://doi.org/10.1037/h0044319

Valentine, J. C., Konstantopoulos, S., & Goldrick-Rab, S. (2017). What happens to students placed into developmental education? A meta-analysis of regression discontinuity studies. *Review of Educational Research, 87*(4), 806–833. https://doi.org/10.3102/0034654317709237
Supplemental Appendix to “Small Study Regression Discontinuity Designs: Density Inclusive Study Size Metric and Performance”

Daryl Swartzentruber and Eloise Kaizar
Department of Statistics, The Ohio State University

September 2, 2022
### SA1 Additional Tables and Figures

**TABLE SA1.**
*Interval estimate success rates (in percents) for RV1.*

|        | $\hat{m} = 10$          | $\hat{m} = 20$          | $\hat{m} = 35$          | $\hat{m} = 50$          |
|--------|--------------------------|--------------------------|--------------------------|--------------------------|
|        | RBC  | CV  | FLCI | RBC  | CV  | FLCI | RBC  | CV  | FLCI | RBC  | CV  | FLCI |
| $\mu_1$ |      |     |      |      |     |      |      |     |      |      |     |      |
| IK1    | 99.85 | 99.85 | 99.53 | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| AK1    | 90.20 | 90.24 | 52.51 | 99.94  | 99.94  | 99.37  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| AK2    | 98.79 | 98.79 | 97.57 | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| CCT0   | 74.39 | 74.39 | NA    | 95.34  | 95.34  | NA    | 99.56  | 99.56  | NA    | 99.94  | 99.94  | NA    |
| CCT1   | 82.48 | 82.49 | 54.97 | 99.69  | 99.68  | 98.60  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| CCT2   | 80.14 | 80.18 | 67.19 | 99.91  | 99.90  | 99.77  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| CCF1   | 76.54 | 76.55 | 41.47 | 99.21  | 99.21  | 95.86  | 100.00  | 100.00  | 99.96  | 100.00  | 100.00  | 100.00  |
| CCF2   | 71.99 | 71.96 | 53.12 | 99.64  | 99.61  | 98.90  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| $\mu_2$ |     |     |      |     |     |      |     |     |      |     |     |      |
| IK1    | 97.14 | 97.14 | 88.36 | 99.99  | 99.99  | 99.93  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| AK1    | 53.72 | 54.14 | 3.41  | 82.93  | 82.99  | 31.51  | 98.14  | 98.14  | 84.71  | 99.88  | 99.88  | 98.33  |
| AK2    | 77.26 | 77.37 | 56.88 | 97.27  | 97.27  | 94.46  | 99.92  | 99.92  | 99.86  | 100.00  | 100.00  | 100.00  |
| CCT0   | 67.64 | 67.64 | NA    | 84.72  | 84.72  | NA    | 96.59  | 96.59  | NA    | 99.30  | 99.30  | NA    |
| CCT1   | 82.52 | 82.52 | 45.52 | 98.87  | 98.87  | 93.37  | 99.99  | 99.99  | 99.99  | 100.00  | 100.00  | 100.00  |
| CCT2   | 86.30 | 86.30 | 73.39 | 99.93  | 99.93  | 99.79  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| CCF1   | 74.55 | 74.57 | 29.89 | 96.94  | 96.93  | 83.34  | 99.94  | 99.94  | 99.26  | 100.00  | 100.00  | 99.98  |
| CCF2   | 77.70 | 77.69 | 57.60 | 99.64  | 99.64  | 98.83  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| $\mu_3$ |     |     |      |     |     |      |     |     |      |     |     |      |
| IK1    | 99.64 | 99.64 | 98.70 | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| AK1    | 88.49 | 88.55 | 47.16 | 99.88  | 99.88  | 98.85  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| AK2    | 98.44 | 98.44 | 96.87 | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| CCT0   | 77.95 | 77.95 | NA    | 97.46  | 97.46  | NA    | 99.86  | 99.86  | NA    | 99.99  | 99.99  | NA    |
| CCT1   | 80.35 | 80.35 | 51.35 | 99.69  | 99.69  | 98.52  | 100.00  | 100.00  | 100.00  | 99.99  | 100.00  | 100.00  |
| CCT2   | 77.81 | 77.71 | 63.57 | 99.89  | 99.88  | 99.73  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
| CCF1   | 73.97 | 73.97 | 38.21 | 99.18  | 99.18  | 95.82  | 100.00  | 100.00  | 99.97  | 100.00  | 100.00  | 100.00  |
| CCF2   | 68.96 | 68.95 | 49.47 | 99.58  | 99.58  | 98.77  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  | 100.00  |
TABLE SA2.

Interval estimate success rates (in percents) for RV2.

|       | \( \hat{m} = 10 \) | \( \hat{m} = 20 \) | \( \hat{m} = 35 \) | \( \hat{m} = 50 \) |
|-------|-------------------|-------------------|-------------------|-------------------|
|       | RBC   | CV    | FLCI  | RBC   | CV    | FLCI  | RBC   | CV    | FLCI  | RBC   | CV    | FLCI  |
| \( \mu_1 \) |       |       |       |       |       |       |       |       |       |       |       |       |
| IK1   | 98.83 | 98.83 | 96.98 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| AK1   | 81.87 | 82.06 | 40.82 | 99.80 | 99.81 | 98.21 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| AK2   | 92.94 | 93.00 | 86.14 | 99.99 | 99.99 | 99.96 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| CCT0  | 73.12 | 73.12 | NA    | 97.66 | 97.66 | NA    | 99.89 | 99.89 | NA    | 99.89 | 99.89 | NA    |
| CCT1  | 66.35 | 66.37 | 39.31 | 99.35 | 99.35 | 97.36 | 100.00| 100.00| 99.99 | 100.00| 100.00| 100.00|
| CCT2  | 55.71 | 55.67 | 43.56 | 99.48 | 99.47 | 98.85 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| CCF1  | 60.34 | 60.33 | 29.26 | 98.61 | 98.61 | 93.48 | 99.99 | 99.99 | 99.96 | 100.00| 100.00| 100.00|
| CCF2  | 47.97 | 47.97 | 32.98 | 98.67 | 98.69 | 96.73 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| \( \mu_2 \) |       |       |       |       |       |       |       |       |       |       |       |       |
| IK1   | 97.31 | 97.31 | 91.06 | 99.98 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| AK1   | 68.02 | 68.22 | 13.37 | 96.31 | 96.32 | 75.36 | 99.92 | 99.92 | 99.06 | 100.00| 100.00| 99.99|
| AK2   | 86.68 | 86.76 | 74.88 | 99.71 | 99.72 | 99.41 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| CCT0  | 67.30 | 67.30 | NA    | 95.07 | 95.07 | NA    | 99.63 | 99.63 | NA    | 99.97 | 99.97 | NA    |
| CCT1  | 65.50 | 65.47 | 36.83 | 99.34 | 99.34 | 97.09 | 100.00| 100.00| 99.99 | 100.00| 100.00| 100.00|
| CCT2  | 55.64 | 55.70 | 43.39 | 99.49 | 99.48 | 98.86 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| CCF1  | 59.11 | 59.11 | 26.01 | 98.49 | 98.50 | 92.65 | 99.99 | 99.99 | 99.94 | 100.00| 100.00| 100.00|
| CCF2  | 47.86 | 47.91 | 32.65 | 98.67 | 98.67 | 96.69 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| \( \mu_3 \) |       |       |       |       |       |       |       |       |       |       |       |       |
| IK1   | 98.72 | 98.72 | 96.44 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| AK1   | 81.70 | 81.86 | 40.77 | 99.80 | 99.80 | 98.28 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| AK2   | 92.80 | 92.86 | 86.01 | 99.98 | 99.98 | 99.96 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| CCT0  | 73.66 | 73.66 | NA    | 98.24 | 98.24 | NA    | 99.95 | 99.95 | NA    | 99.95 | 99.95 | NA    |
| CCT1  | 65.36 | 65.35 | 38.03 | 99.36 | 99.36 | 97.30 | 100.00| 100.00| 99.99 | 100.00| 100.00| 100.00|
| CCT2  | 54.96 | 55.04 | 42.63 | 99.47 | 99.45 | 98.85 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
| CCF1  | 59.04 | 59.08 | 28.22 | 98.57 | 98.57 | 93.38 | 99.99 | 99.99 | 99.95 | 100.00| 100.00| 100.00|
| CCF2  | 47.14 | 47.16 | 32.10 | 98.67 | 98.68 | 96.70 | 100.00| 100.00| 100.00| 100.00| 100.00| 100.00|
TABLE SA3.
Interval estimate success rates (in percents) for RV3.

| μ | \( \bar{m} = 10 \) | \( \bar{m} = 20 \) | \( \bar{m} = 35 \) | \( \bar{m} = 50 \) |
|---|---|---|---|---|
|   | RBC | CV | FLCI | RBC | CV | FLCI | RBC | CV | FLCI | RBC | CV | FLCI |
| μ₁ |   |   |   |   |   |   |   |   |   |   |   |   |
| IK1 | 95.78 | 95.78 | 90.20 | 99.99 | 99.99 | 99.99 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| AK1 | 73.15 | 73.40 | 29.16 | 99.02 | 99.03 | 93.89 | 100.00 | 100.00 | 99.99 | 100.00 | 100.00 | 100.00 |
| AK2 | 84.33 | 84.45 | 70.73 | 99.83 | 99.83 | 99.65 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| CCT0 | 61.71 | 61.71 | NA | 97.15 | 97.15 | NA | 99.95 | 99.95 | NA | 100.00 | 100.00 | NA |
| CCT1 | 48.73 | 48.73 | 25.62 | 97.81 | 97.81 | 92.63 | 99.99 | 99.99 | 99.97 | 100.00 | 100.00 | 100.00 |
| CCT2 | 37.20 | 37.18 | 27.93 | 97.46 | 97.48 | 95.38 | 100.00 | 100.00 | 99.99 | 100.00 | 100.00 | 100.00 |
| CCF1 | 42.49 | 42.49 | 17.27 | 95.78 | 95.80 | 85.11 | 99.96 | 99.96 | 99.79 | 100.00 | 100.00 | 100.00 |
| CCF2 | 30.22 | 30.27 | 19.16 | 94.65 | 94.68 | 89.37 | 99.97 | 99.97 | 99.95 | 100.00 | 100.00 | 100.00 |

| μ₂ |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IK1 | 95.61 | 95.61 | 89.62 | 99.99 | 99.99 | 99.99 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| AK1 | 72.95 | 73.20 | 29.09 | 98.99 | 99.00 | 93.86 | 100.00 | 100.00 | 99.98 | 100.00 | 100.00 | 100.00 |
| AK2 | 84.15 | 84.25 | 70.59 | 99.84 | 99.84 | 99.64 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| CCT0 | 61.84 | 61.84 | NA | 97.20 | 97.20 | NA | 99.95 | 99.95 | NA | 100.00 | 100.00 | NA |
| CCT1 | 48.82 | 48.82 | 25.52 | 97.81 | 97.81 | 92.61 | 99.99 | 99.99 | 99.97 | 100.00 | 100.00 | 100.00 |
| CCT2 | 37.21 | 37.26 | 27.88 | 97.50 | 97.44 | 95.39 | 100.00 | 100.00 | 99.99 | 100.00 | 100.00 | 100.00 |
| CCF1 | 42.50 | 42.50 | 17.12 | 95.79 | 95.80 | 85.06 | 99.96 | 99.96 | 99.79 | 100.00 | 100.00 | 100.00 |
| CCF2 | 30.27 | 30.21 | 19.13 | 94.61 | 94.56 | 89.35 | 99.97 | 99.97 | 99.95 | 100.00 | 100.00 | 100.00 |

| μ₃ |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| IK1 | 95.69 | 95.69 | 90.00 | 99.99 | 99.99 | 99.99 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| AK1 | 73.23 | 73.46 | 29.31 | 99.02 | 99.03 | 93.97 | 100.00 | 100.00 | 99.98 | 100.00 | 100.00 | 100.00 |
| AK2 | 84.25 | 84.34 | 70.85 | 99.82 | 99.83 | 99.65 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| CCT0 | 61.74 | 61.74 | NA | 97.27 | 97.27 | NA | 99.95 | 99.95 | NA | 100.00 | 100.00 | NA |
| CCT1 | 48.46 | 48.48 | 25.44 | 97.80 | 97.80 | 92.61 | 99.99 | 99.99 | 99.97 | 100.00 | 100.00 | 100.00 |
| CCT2 | 37.06 | 37.09 | 27.75 | 97.44 | 97.46 | 95.39 | 100.00 | 100.00 | 99.99 | 100.00 | 100.00 | 100.00 |
| CCF1 | 42.23 | 42.22 | 17.19 | 95.79 | 95.80 | 85.13 | 99.96 | 99.96 | 99.80 | 100.00 | 100.00 | 100.00 |
| CCF2 | 29.97 | 30.00 | 19.07 | 94.60 | 94.61 | 89.36 | 99.97 | 99.97 | 99.95 | 100.00 | 100.00 | 100.00 |
FIGURE SA1. Bandwidth success rates (red) and distributions (black) for the IK1 algorithm.
FIGURE SA2. Bandwidth success rates (red) and distributions (black) for the AK1 algorithm.
FIGURE SA3. Bandwidth success rates (red) and distributions (black) for the AK2 algorithm.
FIGURE SA4. Bandwidth success rates (red) and distributions (black) for the CCT0 algorithm.
FIGURE SA5. Bandwidth success rates (red) and distributions (black) for the CCT1 algorithm.
FIGURE SA6. Bandwidth success rates (red) and distributions (black) for the CCT2 algorithm.
FIGURE SA7. Performance measures for methods using CV and RBC inference for RV2μ2 at two study sizes. FLCI values are essentially the same as CV and thus omitted. Values based on iterations in which all estimates were finite, 4.4% of all iterations for \( \bar{m} = 10 \) and 98.3% for \( \bar{m} = 50 \). The graph with \( \bar{m} = 10 \) omits EmpSE and MSE values for RBC/AK2 (4.04 and 16.35, respectively). For the bias when \( \bar{m} = 10 \), all Monte Carlo standard errors (MCSE) were at most 0.03 except for AK2/RBC (0.09) and AK1/RBC (0.04). For the EmpSE when \( \bar{m} = 10 \), all values shown had MCSE at most 0.02 except AK1/RBC (0.03). For the MSE when \( \bar{m} = 10 \), all values shown had MCSE at most 0.1 except AK1/RBC (0.3), CCT2 (0.2 for RBC and CV) and CCF2 (0.4 for RBC and CV). For \( \bar{m} = 35 \), maximum MCSE values were 0.002 (Bias), 0.001 (EmpSE), and 0.006 (MSE).
FIGURE SA8. Bias of point estimates for different methods and designs at $m = 20$. FLCI values are essentially the same as CV and thus omitted. Values based on iterations in which all estimates were finite. The proportion of finite iterations for each DGP is given in the top right corner of the graph. All MCSE are below 0.008 except RV1$\mu_2$ RBC for AK1 (.07) and AK2 (.02).
FIGURE SA9. Empirical standard errors of point estimates for different methods and designs at $\bar{m} = 20$. FLCI values are essentially the same as CV and thus omitted. Values based on iterations in which all estimates were finite. The proportion of finite iterations for each DGP is given in the top right corner of the graph. The graph of $RV_{1}\mu_{2}$ omits the value for RBC/AK1 (8.07). All estimates shown have MCSE below 0.006 except $RV_{1}\mu_{2}$ RBC/AK2 (0.01).
FIGURE SA10. MSE of point estimates and mean bandwidth values for two DGPs with similar overall sample sizes. FLCI values are essentially the same as CV and thus omitted. The panel on the left with $\mu_2$ omits the value of AK1/RBC (0.24). The panel on the right omits CCF2 values for $\mu_1$ (1.41 for RBC, 1.40 for CV), $\mu_2$ (1.38, 1.37), and $\mu_3$ (1.42, 1.41). All estimates shown in the left panel have MCSE values below 0.0003 except AK2/RBC/$\mu_2$ (0.002). All estimates shown in the right panel have MCSE values below 0.002 except CCT2 (0.04 for $\mu_1$ and $\mu_3$, 0.01 for $\mu_2$, for both RBC and CV), and AK1/RBC and AK2/RBC (0.01 for all $\mu$).
FIGURE SA11. Empirical coverage and median interval width at $\bar{m} = 50$ for all DGPs. Values based on iterations in which all estimates were finite (see Figure SA8). The graph omits values of CCT0/CV, which have coverages lower than 0.89 and widths between 0.16 and 0.28, as well as values of IK1/FLCI, which have coverages at least 0.999 and widths between 0.6 and 2.6. The maximum Monte Carlo standard error for the coverage estimates is 0.0014.
FIGURE SA12. Empirical coverage and median interval width for two DGPs with similar overall sample sizes. The graph omits methods using CCT0, all of which have coverage at most 92%. All estimates shown have MCSE less than 0.002.
FIGURE SA13. Treatment effect estimates and 95% confidence intervals for select Indiana subsets.
SA2  Sensitivity Analysis

In our main simulation results we used the data driven $\hat{M}$ for the second derivative bound in the AK and FLCI methods. However, Armstrong and Kolesár (2020) argue that the choice of $M$ should be made a priori in order to maintain the honesty of their intervals. Clearly knowledge of the true value of $M$ is an advantage as it allows estimation of fewer unknown quantities. To determine how much of an improvement this knowledge provides we include methods in our simulation that make use of this true value of $M$. We let AKM1, AKM2, and FLCIM refer to the algorithms and techniques using the true value of $M$.

There is naturally less variability in bandwidth values when using the consistent value of $M$ rather than estimating $\hat{M}$ each time, which we see in Figure SA14 for $\bar{m} = 20$. The relative size of the bandwidths depends on the relative values of $M$ and $\hat{M}$, which was given in Table 3 in the main text. Larger values of $\hat{M}$ relative to $M$, as in $\mu_1$ and $\mu_3$, lead to smaller bandwidth values for AK1 than AKM1. This is not surprising, because if we suspect larger curvature in the underlying mean function we would not want to include values as far away from the cutoff as we would if we suspected smaller curvature. We see the opposite trend in the $\mu_2$ setting, although the relative differences are smaller. We see very similar results when considering AK2 and AKM2, and these trends continue for other values of $\bar{m}$.

The AKM1 and AKM2 bandwidth algorithms have near universal success rates similar to those of AK1 and AK2. The interval success rates are once again tied to the bandwidth size. Table
TABLE SA4.
Interval estimate success rates (in percents) for RV2 for bandwidth algorithms using the true value of $M$.

|      | $\bar{m} = 10$ | $\bar{m} = 20$ | $\bar{m} = 35$ | $\bar{m} = 50$ |
|------|----------------|----------------|----------------|----------------|
|      | RBC            | CV             | FLCI           | RBC            | CV             | FLCI           | RBC            | CV             | FLCI           | RBC            | CV             | FLCI           |
| $\mu_1$ |                |                |                |                |                |                |                |                |                |                |                |
| AKM1  | 99.49          | 99.50          | 96.82          | 100.00         | 100.00         | 100.00          | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         |
| AKM2  | 99.64          | 99.65          | 99.31          | 100.00         | 100.00         | 100.00          | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         |
| $\mu_2$ |                |                |                |                |                |                |                |                |                |                |                |                |
| AKM1  | 67.06          | 67.27          | 10.98          | 95.46          | 95.47          | 70.85           | 99.91          | 99.91          | 98.72          | 100.00         | 100.00         | 99.98          |
| AKM2  | 86.76          | 86.82          | 74.92          | 99.59          | 99.59          | 99.20           | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         |
| $\mu_3$ |                |                |                |                |                |                |                |                |                |                |                |                |
| AKM1  | 98.21          | 98.22          | 88.66          | 100.00         | 100.00         | 99.97           | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         |
| AKM2  | 99.29          | 99.29          | 98.56          | 100.00         | 100.00         | 100.00          | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         | 100.00         |

SA4 shows the interval success rates for AKM1 and AKM2 for RV2. For $\mu_1$ and $\mu_3$, the larger bandwidths of AKM1 and AKM2 led to higher success rates than those of AK1 and AK2 (see Table SA2). For $\mu_2$, the methods using the true $M$ had lower success rates. This trend holds for the other running variable distributions as well. Once a bandwidth has been calculated, using FLCIM rather than FLCI does not have a meaningful effect on the interval success rates.

For methods using CV inference, there is a relatively small difference in MSE values for methods using AK1 and those using AKM1, as seen in Figure SA15. However the change is larger for methods using RBC inference, especially for smaller sample sizes. In those cases the methods using the true value of $M$ fair better, as would be expected. Even with this improvement the CV methods tend to outperform the RBC methods in terms of point estimation. A comparison of methods using AK2 and AKM2 yields similar results.

The effect of using the true value of $M$ on interval estimation varies considerably depending on DGP. Figure SA16 shows the median interval widths and coverage for all DGPs and methods using AK1 and AKM1 when $\bar{m} = 20$. It includes all four types of inference. However, it excludes AK1/FLCIM and AKM1/FLCI because in practice the same derivative bound would be used for both bandwidth and interval calculations. For $\mu_2$, there is little difference between AK1/FLCI and AKM1/FLCIM. However for $\mu_1$ and $\mu_3$, using the true value of $M$ leads to narrower intervals for about the same coverage. For methods using CV inference, using the true value of $M$ leads to better coverage and narrower intervals for $\mu_1$, but less of a difference for the rest. For methods using RBC inference, using the true value of $M$ leads to less coverage but narrower widths for $\mu_1$ and $\mu_3$. For larger study sizes the trends are not meaningfully different.

Based on our sensitivity analysis using the true value of $M$ can provide a substantial benefit in certain situations, including some DGPs where the $\hat{M}$ estimates are not very close to the truth. If a researcher has strong evidence for a derivative bound, they can certainly use that value. However, it would seem to be a good idea to also use the data driven $\hat{M}$ value in practice. If the two derivative bounds do not match, further investigation into the data should probably take place.
FIGURE SA15. Mean squared error values for methods using AK1 and AKM1 algorithms. The RV1$\mu_2$ graph omits values when $\bar{m} = 20$ for AK1/RBC (1.2) and AKM1/RBC (1.7), and when $\bar{m} = 50$ for AK1/RBC (0.24) and AKM1/RBC (0.27). The RV2$\mu_2$ graph omits values when $\bar{m} = 20$ for AK1/RBC (65.2) and AKM1/RBC (66.2). The maximum Monte Carlo standard errors (MCSE) for the estimates shown is 0.003 (RV1), 0.009 (RV2), and 0.01 (RV3).
FIGURE SA16. Empirical coverage and median interval widths for methods using AK1 and AKM1 algorithms when \( \bar{m} = 20 \). The maximum MCSE value for these estimates is 0.002.
References

Armstrong, T. B., & Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics, 11*(1), 1–39. https://doi.org/10.3982/QE1199