# The Power Asymmetry in Fuzzy Regression Discontinuity Designs ⋆

Daniel Kaliski, Michael Keane and Timothy Neal[a]

[a] *Birkbeck, University of London, Johns Hopkins University, and University of New South Wales,*

May 30, 2025

**Abstract**

Regression discontinuity (RD) is a popular method to estimate the effect of endogenous treatment. In a Fuzzy RD design, the probability of treatment jumps discontinuously when a running variable ($R$) passes a threshold ($R_0$). Fuzzy RD estimates are obtained via a procedure analogous to two-stage least squares (2SLS), where an indicator $I(R > R_0)$ plays the role of the instrument. Recently, Keane and Neal (2023, 2024) showed that 2SLS $t$-tests suffer from a "power asymmetry" problem: 2SLS standard errors are spuriously small (large) when the 2SLS estimate is close to (far from) the OLS estimate. Here we show that a similar problem arises in fuzzy RD designs. As a result, if the OLS bias is positive, the Fuzzy RD $t$-test has little power to detect true negative effects; And, if the true effect is zero, it has inflated power to find false positives. The problem persists even if the instrument (first stage) is very strong. A simple way to avoid this problem is to rely exclusively on the intent to treat (ITT) regression to assess significance of the treatment effect, where the ITT regression is simply a sharp RD of the outcome on $I(R > R_0)$.

*Keywords:* Regression discontinuity design; Treatment effect; 2SLS; Endogeneity; Intent to Treat Effect; Size distortion; Anderson–Rubin test.

*JEL: C12, C14, C26, C36*

## 1. Introduction

Recently, Keane and Neal (2023, 2024) showed that two stage least squares (2SLS) suffers from a "power asymmetry" problem: The 2SLS standard error estimate is artificially small (large) when the 2SLS parameter estimate is near (far from) OLS. In contrast to the well-known weak instrument problem (Bound et al., 1995; Stock et al., 2002), the power asymmetry exists even if instruments are very strong by conventional standards. In this paper we show that an analogous power asymmetry problem also arises in fuzzy regression discontinuity designs (Fuzzy RDs). We explore implications for estimation and inference in Fuzzy RDs, and show how inference based on the intent-to-treat (ITT) regression can avoid the power asymmetry problem.

⋆Kaliski: d.kaliski@bbk.ac.uk, Keane: mkeane14@jhu.edu, Neal: timothy.neal@unsw.edu.au. We thank Josh Catalano, Vadim Marmer and Claudia Noack for helpful comments. Additional thanks to Josh Catalano and Vadim Marmer for sharing R and MATLAB code to produce AR confidence sets as in Feir et al. (2016).

The 2SLS power asymmetry creates two serious inferential problems: First, if the true $\beta = 0$, a one-tailed 2SLS $t$-test of $H_0$:$\beta \leq (\geq)0$ will reject at an inflated (deflated) rate if the OLS endogeneity bias is positive.[1] The converse is true if the OLS bias is negative. Hence, an estimate $\hat{\beta}_{2SLS}$ is more likely to be judged significant by a one- or two-tailed $t$-test if that estimate departs from zero in the same direction as the OLS endogeneity bias. Second, 2SLS has very poor power to detect true negative (positive) effects when the OLS bias is positive (negative).

For example, in a typical application of IV, one cares if a treatment like a training program has a positive effect on an outcome like wages. But one is concerned that selection into the program is positive – e.g., more motivated people are more likely to participate, and they have have high wages anyway. Hence, one instruments using an exogenous variable that encourages participation. The 2SLS $t$-test is biased in favor of finding positive results: It has poor power to detect a true negative treatment effect; And, if the true treatment effect is zero, it has inflated power to find false positives.

Here, we show Fuzzy RD suffers from a similar power asymmetry problem. This follows from the close analogy between Fuzzy RD and 2SLS noted by Hahn et al. (2001). As a result, Fuzzy RD estimates shifted in the direction of the endogeneity bias (that motivates use of RD in the first place) have artificially small standard errors, rendering the $t$-test unreliable. Furthermore, this problem exists even if the "first stage" regression of treatment on the instrument $I(R > R_0)$ is strong.

The solution to the power asymmetry in 2SLS is inference via the reduced form; see Anderson and Rubin (1949). Due to the non-parametric nature of Fuzzy RD, there is no exact reduced form. However, the ITT regression, a sharp RD of the outcome on $I(R > R_0)$ is a "quasi-reduced form." Feir, Lemieux, and Marmer (2016) and Noack and Rothe (2024) have proposed an "AR approach" to inference via the ITT regression if identification is weak. We show this AR approach has the added benefit of avoiding the power asymmetry, and should be used even if identification is strong.

Next, in Section 2, we explain the relation between Fuzzy RD and 2SLS. Section 3 describes two modern approaches to RD: The robust bias correction (RBC) approach of Calonico, Cattaneo, and Titiunik (2014b) and the bounded second derivative (BSD) approach of Kolesár and Rothe (2018). Then, in Sections 4-5, we assess the power asymmetry in Fuzzy RD quantitatively, and show it can be avoided by using the AR approach. Finally, Section 6 illustrates the importance of these issues by revisiting the Ambrus et al. (2020) study of the impact of cholera deaths on rents. Their Fuzzy RD estimate, which is far from OLS, is insignificant according to the $t$-test, but significant according

[1] For example, in a context where a two-tailed $t$-test has correct size of 5%, and the OLS bias is positive, the power asymmetry may cause a one-tailed $t$-test of $H_0$:$\beta \leq 0$ to have inflated size of 5%, while a test of $H_0$:$\beta \geq 0$ has size 0.

to the ITT regression, at least if using the RBC approach. Our analysis shows that, in their context, an AR-type test based on the ITT regression test has at least 5 times the power of the $t$-test.

## 2. Problems with 2SLS $t$-tests and their Relevance to Fuzzy RDDs

### *2.1.* ***Fuzzy RDs as a Special Case of 2SLS***

Let $Y$ be the outcome of interest, $X$ the endogenous explanatory variable (or "treatment") of interest. Let $R$ be an ordered running variable. In a "sharp" RD design all units with $R \geq R_0$ are assigned to treatment, while no units with $R < R_0$ are treated. The object of interest is the difference between the left- and right- limits of the expected outcome at the point $R = R_0$,

$$\xi_1 = \lim_{r \to R_0^+} E[Y|R = r \geq R_0] - \lim_{r \to R_0^-} E[Y|R = r < R_0], \tag{1}$$

which can be written compactly as $\xi_1 = Y^+ - Y^-$. This identifies the effect of treatment $\beta_1$.[2]

The fuzzy RD design relaxes the assumption of deterministic assignment, allowing the probability of treatment conditional on $R$ to jump discontinuously at $R_0$. We may express the jump as:

$$\pi_1 = \lim_{r \to R_0^+} E[X|R = r \geq R_0] - \lim_{r \to R_0^-} E[X|R = r < R_0], \tag{2}$$

which can be written compactly as $\pi_1 = X^+ - X^-$. In the case where X is a continuous treatment, equation (2) is the jump in the expected level of treatment at $R_0$.

The fuzzy RD has as its object of interest the ratio of the intent to treat (ITT) effect $Y^+ - Y^-$ to the jump in treatment probability $X^+ - X^-$. This ratio $\beta_1 = \frac{\xi_1}{\pi_1} = \frac{Y^+ - Y^-}{X^+ - X^-}$ identifies the effect of treatment itself. In principle, a consistent estimator of the effect of treatment may be obtained using kernel estimators of $Y^+, Y^-, X^+, X^-$ based on observations near $R_0$, and constructing $\hat{\beta}_1 = \frac{\hat{Y}^+ - \hat{Y}^-}{\hat{X}^+ - \hat{X}^-}$, provided the bandwidth shrinks with sample size. However, Hahn et al. (2001) pointed out that such an approach has poor finite-sample properties when estimating the discontinuities of interest. Hence, local linear regression estimators are preferred – see Fan (1992).

As Hahn et al. (2001) and Imbens and Lemieux (2008) point out, once a bandwidth is chosen, a local linear regression based on a uniform kernel is equivalent to 2SLS. Specifically, using only observations within a window $[R_0 - h, R_0 + h]$, a regression of $Y$ on $X$, instrumenting $X$ by the treatment indicator $D = 1[R \geq R_0]$, while including the local linear terms $R$ and $D \times R$ as controls,

[2] With the usual caveat that if treatment effects are heterogeneous this estimates the local average treatment effect (LATE) for subjects in the vicinity of $R_0$. Identification also requires an assumption that selection into treatment is not based on gains from treatment, referred to as "non-manipulation" in the the RD literature.

is numerically equivalent to obtaining the local linear estimates of $Y^+, Y^-, X^+, X^-$ individually, using that same window and a uniform kernel, and then constructing $\hat{\beta}_1 = \frac{\hat{Y}^+ - \hat{Y}^-}{\hat{X}^+ - \hat{X}^-}$. Thus, fuzzy RDs can be estimated via the system of equations:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 R_i + \beta_3 R_i D_i + u_i, \quad \forall i \quad s.t. \quad R_0 - h \leq R_i \leq R_0 + h, \tag{3}$$

$$X_i = \pi_0 + \pi_1 D_i + \pi_2 R + \pi_3 R_i D_i + e_i, \quad \forall i \quad s.t. \quad R_0 - h \leq R_i \leq R_0 + h, \tag{4}$$

Analogous to 2SLS, equation (3) is the outcome equation that contains the endogenous variable $X$, and (4) is the "first stage" that contains $D = 1[R \geq R_0]$ as the excluded instrument. As with 2SLS, one can estimate $\beta_1$ via a two-step procedure: Take fitted values $\hat{X}$ from the first stage, which can be estimated as a sharp RD, and plug them in for $X$ in the outcome equation, which can then be estimated by OLS using data on the interval $[R_0 - h, R_0 + h]$. The difference with 2SLS is that the "instrument" $D = 1[R \geq R_0]$ is not exogenous, as $Cov(u, D)$ is not in general zero. The exogeneity assumption is replaced with an assumption that $E[u|R]$ is continuous at $R_0$.

Fuzzy RD estimates are biased in finite samples because $E[u|R]$ and $E[e|R]$ may vary as one moves away from $R_0$, generating bias in the estimates $Y^+, Y^-, X^+, X^-$. One may think of the $R$ and $RD$ terms as serving to "sop up" some of that bias. In fact, they do so perfectly if $E[u|R]$ and $E[e|R]$ are linear in $R$ in the region $[R_0 - h, R_0 + h]$. However, a researcher never knows the true relationship between $Y$ or $X$ and $R$, so the local linear terms $R$ and $DR$ may be insufficient to eliminate bias. This motivates more recent methods that use bias correction or bias-aware inference (Calonico et al., 2014b; Kolesár and Rothe, 2018; Noack and Rothe, 2024).[3] We discuss these in Section 3. Our results show that the power asymmetry problem affects fuzzy RD regardless of whether one uses simple 2SLS or these more sophisticated approaches.

A large literature (Ludwig and Miller, 2007; Imbens and Kalyanaraman, 2012; DesJardins and McCall, 2014; Calonico et al., 2014b) discusses how to choose the bandwidth $h$. We find that the power asymmetry arises no matter which bandwidth selection method is used. As for the choice of kernel, we first consider the uniform kernel case, as this gives the close analogy to 2SLS noted above. But the default in most RD software is the triangular kernel, which produces weighted two-stage least squares estimates, with observations closer to the cutoff being given more weight. Cattaneo and Titiunik (2022) point out that each kernel has advantages, with no consensus in the literature as to which is to be preferred. We will show that the power asymmetry arises in either case.

[3] An alternative is to control for higher order terms in $R$. But Gelman and Imbens (2019) argue this places too much weight on observations far from the cutoff. We show it has also makes the power asymmetry problem worse.

### *2.2.* ***The "Reduced Form" or Intent-to-Treat (ITT) Regression***

We can substitute the first stage equation (4) into the outcome or second stage equation (3) to obtain the "reduced form" or Intent-to-Treat (ITT) regression:

$$Y_i = \xi_0 + (\beta_1\pi_1)D_i + \xi_2 R_i + \xi_3 R_i D_i + v_i, \quad \forall i \quad s.t. \quad R_0 - h \leq R_i \leq R_0 + h \tag{5}$$

where $v = (u + \beta e)$, $\xi_1 = (\beta_1\pi_1)$, $\xi_k = (\alpha_k + \beta_1\theta_k)$ for $k = 0, 2, 3$. This takes the form of a sharp RD where the outcome $Y$ is regressed on treatment assignment $D = 1[R \geq R_0]$, rather than treatment itself. This estimates the Intent-to-Treat (ITT) effect $\xi_1 = \beta_1\pi_1$. Lee and Lemieux (2010) refer to (5) as the "reduced form," but we prefer to call it a "quasi-reduced form," as (i) in a non-parametric setting the true model is not specified, and (ii) the "instrument" here is not exogenous.

The effect of treatment $\beta_1 = \frac{Y^+ - Y^-}{X^+ - X^-}$ may be estimated by the ratio $\hat{\beta}_1 = \hat{\xi}_1/\hat{\pi}_1$. This is equivalent to the two-step procedure described in the previous section. The delta-method is used to form the standard error $se(\hat{\beta}_1)$ of this ratio, which may be used to form the $t$-test $t = \hat{\beta}_1/se(\hat{\beta}_1)$ to test $H_0$:$\beta_1 = 0$. Feir, Lemieux, and Marmer (2016) and Noack and Rothe (2024) both note that the Fuzzy RD $t$-test suffers from size distortion when the first stage is weak.

Hence, they propose an alternative that is analogous to the Anderson and Rubin (1949) approach in 2SLS. This AR approach is based on the ITT regression: The idea is to form a $t$-test for significance of $D$ in the reduced form ITT regression (5), and use it to assess whether there is a significant effect of treatment. That is, unless we see a significant effect of treatment assignment on the outcome, we conclude there is no effect of treatment itself. Below we call this a $t_{AR}$ test.

Our main contribution is to show that the Fuzzy RD $t$-test suffers from the power asymmetry problem even when the first stage is strong. But the AR approach, based on the $t_{AR}$ test from the ITT regression, avoids the power asymmetry problem. Hence, we advocate that the $t_{AR}$ test be used in lieu of the Fuzzy RD $t$-test even when the first stage is strong.

### *2.3.* ***The "Infeasible Fuzzy RD" Estimator and the "OLS" Estimator***

An important point is that if $\pi$ were known *a priori*, we could estimate $\beta_1$ directly via a sharp RD regression of $Y$ on $\pi D$, controlling for $R$ and $R \cdot D$. We call this the "Infeasible Fuzzy RD" regression. It provides an important benchmark, as it defines the upper bound of the power curve for feasible fuzzy RD regression where $\pi$ must be estimated; i.e., one cannot gain efficiency by discarding information on true $\pi_1$. It is also important for us to define what we call the "OLS" estimate of $\beta_1$. In the RD context this as what one obtains by estimating the sharp RD regression (3) treating $X$ as exogenous (conditional on the controls for $R$ and $R \cdot D$).

As we show below, the delta-method standard errors $se(\hat{\beta}_1)$ of the ratio $\hat{\beta}_1 = \hat{\xi}_1/\hat{\pi}_1$ tend to be far below the Infeasible Fuzzy RD standard errors when $\hat{\beta}_1$ is near the OLS estimate of $\beta_1$. This is the sense in which the delta method standard errors are too small when $\hat{\beta}_1$ is near the OLS estimate. This, in turn, generates the power asymmetry problem that afflicts the Fuzzy RD $t$-test.

In contrast, a remarkable fact is that the $t_{AR}$ test from the ITT regression is numerically equivalent to the $t$-test on $\beta_1$ from the Infeasible Fuzzy RD regression. This is a good way to understand why $t_{AR}$ has desirable properties, as it is the test one could form if the first stage were known!

### *2.4.* ***The Power Asymmetry Problem in 2SLS***

Here we explain the point in Keane and Neal (2024, 2023) that 2SLS estimates have relatively small (large) standard errors when $\hat{\beta}_{2SLS}$ is close to (far from) $\hat{\beta}_{OLS}$. Consider the triangular system:

$$Y = \beta X + u \tag{6}$$

$$X = \pi Z + e \tag{7}$$

where $Y$ is the outcome, $X$ is an endogenous variable whose effect on $Y$ is of interest, and $Z$ is a valid instrument for $X$. The correlation $\rho$ between teh errors $u$ and $e$ is non-zero, creating the endogeneity problem. But $Z$ satisfies $Cov(Z, u) = 0$ and $\pi \neq 0$. We may obtain the 2SLS estimator of $\beta$ by taking the fitted values $\hat{X} = \hat{\pi}Z$ from the first stage (7), and then running an OLS regression of $Y$ on $\hat{X}$. We also define $\hat{\beta}_{OLS}$ as the estimate obtained via OLS regression of $Y$ on $X$, ignoring the endogeneity problem. The 2SLS standard errors are obtained from the formula:

$$\sqrt{N^{-1}\sum(Y_i - X_i\hat{\beta}_{2SLS})^2/TSS_{X,Z}} \tag{8}$$

where $TSS_{X,Z}$ is the total sum of squares of $X$ explained by the instrument $Z$. From this formula, we can see there are two reasons the 2SLS estimates have relatively small (large) standard errors when $\hat{\beta}_{2SLS}$ is close to (far from) $\hat{\beta}_{OLS}$. First, the numerator, which is the standard error of regression, is minimized when $\hat{\beta}_{2SLS} = \hat{\beta}_{OLS}$. This follows immediately from the definition of $\hat{\beta}_{OLS}$ as minimizing the sum of squared residuals. Hence, the 2SLS estimated standard error of regression is mechanically smaller for realizations of $\hat{\beta}_{2SLS}$ that happen to be close to $\hat{\beta}_{OLS}$. Second, a positive finite-sample realization of $\widehat{cov}(Z, u)$ when $\rho > 0$ both increases $TSS_{X,Z}$, and shifts $\hat{\beta}_{2SLS}$ towards OLS, which also contributes to smaller standard errors when 2SLS is shifted in the direction of the OLS bias.

### *2.5.* ***The Power Asymmetry Problem in Fuzzy RD***

In the Fuzzy RD context, an analogous problem arises: the Fuzzy RD standard error on $\hat{\beta}_1$ tends to be too small (large) when the Fuzzy RD estimate is close to (far from) the OLS estimate, defined as in Section 2.3. In Section 4 we present simulations showing this power asymmetry renders the

Fuzzy RD $t$-test unreliable, even when the first stage is strong. Then in Section 5 we show how the $t_{RA}$ test avoids the problem. But first, we discuss two modern approached to RD inference.

## 3. Modern Approaches to RD Inference: Bias Aware and Bias-Corrected

As we mentioned in Section 2.1, methods have recently been developed that try to deal with bias in estimating the discontinuities $Y^+ - Y^-$ and $X^+ - X^-$ arising if the controls for $R$ and $R \cdot D$ are inadequate to fully capture variation in $E[u|R]$ and $E[e|R]$ as one moves away from $R_0$. Here, we describe the two main methods that we focus on in our Monte Carlo experiments:

### *3.1.* ***Robust Bias-Corrected (RBC) approach of Calonico, Cattaneo, and Titiunik (2014b)***

RBC inference is implemented in the popular "rdrobust" package for R and Stata – see Calonico et al. (2015) and Calonico et al. (2014a). In the sharp RD case, if we estimate $\hat{\xi}_1 = Y^+ - Y^-$ using an MSE optimal bandwidth, and $f(R)$ is the true $E(u|R)$ function, then the leading term in a Taylor expansion of the nonparametric asymptotic bias of a local linear estimator is:

$$b = h^2(f''(R|R \geq R_0)\mathcal{B}^+_{\mathcal{K}}(h) - f''(R|R < R_0)\mathcal{B}^-_{\mathcal{K}}(h))(1 + o_p(1)), \tag{9}$$

where $\mathcal{B}^+, \mathcal{B}^-$ are weights that depend on the bandwidth and kernel used to weight observations.[4] Obviously the bias depends on the second derivative of $f(R)$ near $R_0$. This is unknown to the researcher, but can be estimated. Calonico et al. (2014b) propose a multi-step procedure for choosing an initial "wide" bandwidth for estimating the bias correction $\hat{b}$, and then an MSE-optimal bandwidth to form the bias-corrected estimator $\hat{\xi}_1 - \hat{b}$.[5]

In the fuzzy RD case, they show that the bias of the ratio estimator ratio $\hat{\beta}_1 = \hat{\xi}_1/\hat{\pi}_1$ can be written as a linear function of the biases in estimating the sharp RD parameters $\hat{\xi}_1$ and $\hat{\pi}_1$, plus higher order terms. Estimates of these two bias terms may be constructed as described above.

Second, Calonico et al. (2014b) also adjust the variance of the estimator to account for the extra variability introduced by the bias correction term. The resulting confidence intervals can be written $\hat{\beta}_1 - \hat{b} \pm 1.96 \times \sqrt{\hat{V} + \hat{W}}$ where $\hat{V}$ denotes the variance of the uncorrected estimator, and $\hat{W}$ is the extra variance induced by bias correction (the "robust" part of "robust bias-corrected" inference).

[4] The MSE is the squared bias of the estimator, which is proportional to $h^4$, plus the variance, which proportional to $1/Nh$. Hence, the MSE minimizing bandwidth is $h \propto N^{-1/5}$. This is shrinking too slowly with $N$ to eliminate asymptotic bias – hence the need for bias correction. An alternative is to promise to have $h$ shrink more quickly with $N$, which is known as "under-smoothing."

[5] The MSE optimal bandwidth has the form $h = (\mathcal{V}/(4\mathcal{B}^2 + \mathcal{R}))^{1/5} N^{-1/5}$, where $\mathcal{B}$ is a bias term that depends on $f''(R)$ near the boundary, $\mathcal{R} > 0$ is a regularization term, and $\mathcal{V}$ depends on the variance of the estimated intercepts of the local linear regression. A key difference between Calonico et al. (2014b) and Imbens and Kalyanaraman (2012) is that CCT estimate $\mathcal{V}$ in this formula using the estimated variance of the intercepts in the local linear regression, while IK use estimates of $V(Y)$ and the density of $R$ near $R_0$ (on which the variance of the intercepts depend). The CCT bandwidth shrinks at a slightly faster rate, and we find it tends to be smaller in practice, see Appendix B.

### *3.2.* ***The Bounded Second Derivative (BSD) approach of Kolesár and Rothe (2018)***

Kolesár and Rothe (2018) introduced bounded second derivative (BSD) confidence intervals (CIs), implemented in the RDHonest package for R and Stata. This "bias aware" approach does not correct for bias in local linear estimators. Rather, it seeks to bound the bias, by bounding $|f''(R)|$. The basic idea of an "honest" confidence interval is to expand the conventional $t$-test CI in an attempt to insure that the adjusted CI covers the true value at the correct nominal rate $\alpha$, despite the existence of bias. The necessary CI expansion is increasing in the magnitude of $|f''(R)|$. Given an assumed bound $\max|f''(R)| < M$, the honest BSD confidence interval can be written:

$$\hat{\beta}_1 \pm cv_{1-\alpha}(\psi(M,h)) \times se(\hat{\beta}_1) \tag{10}$$

where $\psi(M,h)$ is maximum bias (normalized by the standard error), and the critical value $cv_{1-\alpha}(\cdot)$ is the $1-\alpha$ quantile of the absolute value of a standard normal $|N(\psi(M,h),1)|$, which is increasing in $M$ and bandwidth $h$.[6] Kolesár and Rothe (2018) recommend use of judgement to choose $M$, but the RDHonest default is the Armstrong and Kolesár (2020) approach, where a global fourth-order polynomial fit either side of the cutoff is used to estimate $M = max|f''(R)|$. The bandwidth is chosen to minimize "worst-case MSE," defined as $MSE_{wc} = Bias^2_{max} + se(\hat{\beta}_1)^2$, where $Bias_{max} = \psi(M,h) \cdot se(\hat{\beta}_1)$ depends on $M$ and $h$. See Appendix A for more details.

As BSD accommodates bias, rather than relying on a shrinking bandwidth to eliminate bias asymptotically, its CIs are valid regardless of whether the running variable is continuous or discrete, unlike RBC confidence intervals, which assume the running variable is continuous.[7] Of course, it is rare for a running variable to be truly continuous, but if $R$ takes on many discrete values near the cutoff it can reasonably be treated as continuous – see Cattaneo et al. 2024.

There is currently controversy over whether the RBC or BSD method is preferred. Hence, we consider both approaches in our analysis.[8] We also consider what Calonico et al. (2014a) refer to as the "conventional" approaches – with no bias correction or CI adjustment – in Appendix B.

[6] Note that $\psi(0,h)$=0, so when $M = 0$ we get the usual critical value of 1.96. See Appendix A for details.

[7] Note that the RBC approach relies on a shrinking bandwidth as $N$ approaches infinity, so that asymptotically one limits the analysis to observations within epsilon of the cutoff. The RBC bias correction relies on an asymptotic formula. But with a discrete running variable this apparatus doesn't work, as there will never be observations within epsilon of the cutoff, regardless of how large is $N$, due to the discrete nature of $R$ – see Cattaneo et al. 2024.

[8] Simulation studies in Calonico et al. (2018) and Ganong and Jäger (2018) find RBC exhibits good finite-sample properties. But Armstrong and Kolesár (2020) show it can undercover the parameter of interest in some circumstances, that are hard for an applied researcher to assess. Noack and Rothe (2024) find BSD confidence intervals cover the true parameter more reliably than RBC, even if true $M$ is double the assumed value. This is because BSD CIs are conservative; see Beckert and Kaliski (2024). But Cattaneo and Titiunik (2022) criticize BSD inference on the grounds that (i) manually choosing $M$ amounts to manually choosing the bandwidth, as the former largely determines the latter, while (ii) estimating $M$ from data sacrifices the uniform validity property that motivates use of BSD CIs.

## 4. Examples of the Power Asymmetry in Fuzzy Regression-Discontinuity Designs

In this section we present four Monte Carlo examples to illustrate the power asymmetry problem in fuzzy RDs. The data generating processes (DGPs) we consider are all simple cases that satisfy the assumptions for valid fuzzy RD inference. We focus on these simple cases to emphasize the problems we uncover are fundamental – i.e., they do not only emerge in complex or pathological cases. They are: (i) A randomized controlled trial with perfect compliance, (ii) a randomized controlled trial with imperfect compliance, (iii) a case with a near-linear relationship between the running variable and $Y$, and (iv) a case with quadratic relationship between the running variable and $Y$.

### *4.1.* ***Case 1: RD Inference in an RCT With Perfect Compliance***

Suppose a researcher is interested in the effect of a continuous variable $X$ on the outcome $Y$. Treatment in this experiment exogenously shifts up the level of $X$. Units are assigned to the treatment group if and only if $1[R \geq 0]$, where $R$ is uniformly distributed on $[-A, A]$. Thus, treatment assignment is akin to drawing a lottery number, or a coin flip, with an equal probability of being assigned to either the treatment or control group. The DGP for this case, of an RCT with perfect compliance, is as follows:

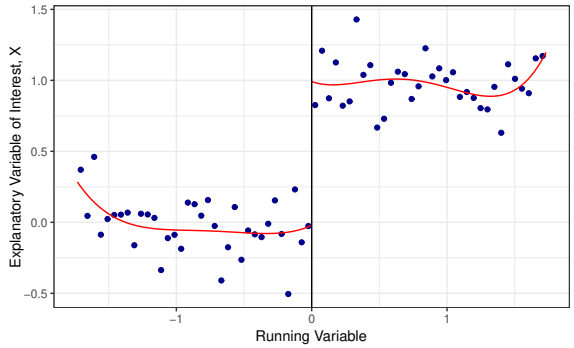
$$\begin{aligned} Y_i &= \beta X_i + u_i, \\ D_i &= 1[R_i \geq 0], \\ X_i &= \pi D_i + e_i, \\ e_i &= \rho u_i + \sqrt{1-\rho^2}\eta_i, \\ R \backsim Unif(-\sqrt{3}, \sqrt{3}), &\quad (u, \eta) \sim iidN(0,1) \end{aligned}$$

We assume the stochastic terms $\eta$ and $u$ are iid standard normal random variables, and we construct $e$ so it is also standard normal, with $\rho$ the correlation between $u$ and $e$. These variance normalizations are without loss of generality, as $Y$ and $X$ can always be normalized so this is true. The support of $R$ is set at $(-\sqrt{3}, \sqrt{3})$ so that $\sigma_R^2 = 1$, but this is also innocuous.

We set $\rho = 0.8$ so the endogeneity problem the experiment is designed to solve is severe. Because $\rho > 0$ the OLS bias from naive regression of $Y$ on $X$ is positive. We set the true effect of interest to be $\beta = 0$, so $E(\hat{\beta}_{OLS}) = Cov(Y, X)/Var(X) = 0.80/1.25 = .64$, increasing to .75 if we control for $R$ and $RD$. We also set the sample size to $N = 2,000$. This allows us to maintain a reasonable effective sample size after bandwidth selection.

Finally, we set $\pi = 1$, so that assignment to treatment causes a one standard deviation increase in $X$. Given the sample size in our runs ($N$=2000) this insures the first stage regression of $X$ on $D$ is "strong" in the sense that first-stage $F$s are well above the conventional threshold of $F = 10$, even if the majority of observations are discarded due to bandwidth selection. Figure 1 displays a representative first-stage RD plot for the above process.

Figure 1: **Representative First Stage, DGP 1: RCT with Perfect Compliance**



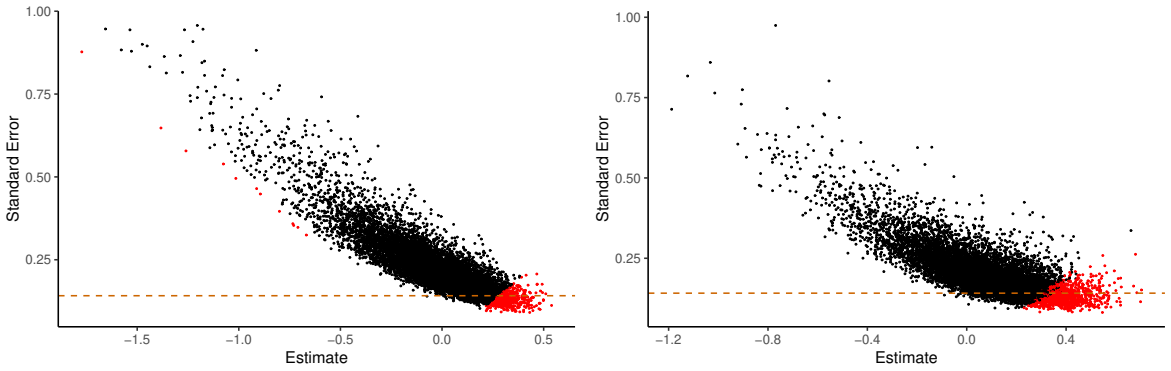

*Notes:* Binned averages of the explanatory variable of interest $X$ in intervals of the running variable $R$ for the 10,000$^{th}$ of 10,000 replications with 2,000 observations each.

We generate 10,000 artificial data sets from this DGP, and then apply rdrobust and RDHonest to each one. In this simple case treatment $D$ is exogenous (i.e., $Cov(D, u) = 0$) so RD is not necessary. One could run 2SLS on the full sample using $D$ as an instrument for $X$. However, the rdrobust and RDHonest algorithms do not know this *a priori*, so they choose bandwidths, bias corrections and CI corrections based on sample estimates of the CEF function $E(u|R)$ and its derivatives.

Figure 2 presents the results from the 10,000 runs, with RBC results (from rdrobust) on the left, and BSD results (from RDHonest) on the right. In both cases we use a uniform kernel. The figure plots the standard error from each run on the $y$-axis, and the $\hat{\beta}$ on the $x$-axis. Estimates that are significant according to the $t$-test are shaded in red. The RBC and BSD estimates and standard errors in left/right panels are different for two reasons: (i) RBC implements bias correction, and (ii) RBC inflates the standard error (In contrast BSD inflates the $t$-test critical values). For BSD we plot the conventional standard errors, but we shade in red only cases that are significant according to the inflated $t$-test critical values in equation 10.

The key thing to notice in Figure 2 is the strong negative association between the estimates and their standard errors. Estimates that are positive (in the direction of the OLS bias) have spuriously small standard errors. As a result, positive estimates $\hat{\beta}$ are much more likely to be judged significant by the $t$-test than negative estimates. This pattern emerges for both RBC and BSD.

Figure 2: **DGP 1: An RCT with Perfect Compliance**




*Note:* Based on 10,000 replications with 2000 observations each. *Left:* Robust bias-corrected (RBC) inference via the rdrobust R package. *Right:* Bounded second derivative (BSD) inference via the RDHonest R package. Both use a uniform kernel. We plot $\hat{\beta}_1$ against $se(\hat{\beta}_1)$. Runs with a standard error $> 1$ excluded. Red dots indicate $H_0$:$\beta = 0$ rejected at 5% level. The dashed line indicates the lower bound standard error of .141 obtained assuming first-stage $\pi$ is known, as we explain in Section 4.2.

As we see in Table 1, row 1, RBC inference rejects the true null that $\beta = 0$ at almost exactly the correct 5% nominal rate (5.1%). But 98% of rejections occur when $\hat{\beta}$ is positive. BSD rejects $H_0$:$\beta = 0$ at an inflated rate of 9.8%, and 100% of these rejections occur when $\hat{\beta}$ is positive. So for both procedures, the power asymmetry problem is severe.

Table 1: **RBC and BSD Inference with Uniform Kernel on DGPs 1 to 4**

| | RBC | | | | | | | BSD | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ |
| 1 | 5.1% | 97.7% | | -.003 | .207 | 32.04 | 508 | 9.8% | 100% | | .094 | .185 | 24.40 | 334 |
| 2 | 4.5% | 76.1% | | -.005 | .417 | 63.17 | 498 | 3.3% | 99.7% | | .139 | .413 | 44.93 | 341 |
| 3 | 4.5% | 79.7% | | .004 | .442 | 49.59 | 451 | 3.2% | 99.7% | | .157 | .438 | 36.92 | 303 |
| 4 | 4.7% | 84.0% | | .012 | .456 | 40.00 | 357 | 5.3% | 100% | | .284 | .451 | 33.41 | 289 |

*Notes:* Summary results for 10,000 artificial datasets of size $N = 2000$ each. The 4 rows report results for the 4 DGPs discussed in Sections 4.1 to 4.5. RBC and BSD indicate results from the rdrobust and RDHonest packages, respectively. Both use a uniform kernel. We report the rate of rejecting $H_0$:$\beta = 0$, the fraction of these rejections that occur when $\hat{\beta} > 0$, and the medians of the estimate, estimated standard error, first stage F, and effective observations.

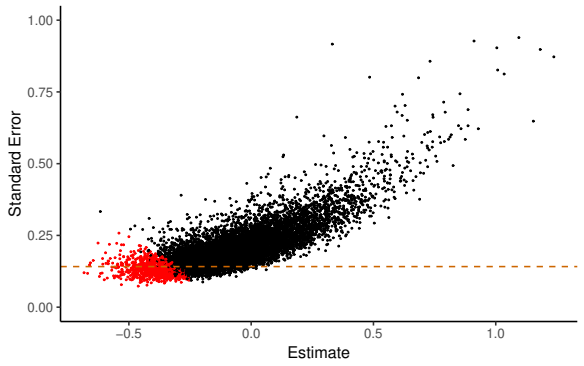
As we see in Table 1, row 1, RBC generates essentially no median bias, while the median bias of BSD is 0.094. It is important to understand that Fuzzy RD suffers from no inherent bias due to mis-specification of the local linear regression in this case, as $f(R) = 0$, $M = 0$. This is reflected in the fact that the RBC bias correction terms are very small, and ignoring them has almost no impact on the RBC results. Thus, it is interesting to ask why BSD exhibits median bias.

The median bias in BSD is due to the power asymmetry. As we noted in Section 3.2, the BSD

bandwidth is chosen to minimize $MSE_{wc} = Bias^2_{max} + se(\hat{\beta}_1)^2$, which depends on the estimated standard error of the fuzzy RD estimate $se(\hat{\beta}_1)$. As we have explained, fuzzy RD estimates close to $\hat{\beta}_{OLS}$ tend to have small standard errors. Hence, BSD has a tendency to choose bandwidths that generate estimates close to $\hat{\beta}_{OLS}$.[9] In DGP 1, this positive bias compounds the power asymmetry problem to generate a highly inflated rate (9.8% vs. 2.5%) of rejecting $H_0$:$\beta \leq 0$, combined with a zero rate of rejecting $H_0$:$\beta \geq 0$. (In Section 4.7 we show the bias problem with BSD is much less if one uses a triangular kernel instead of the uniform.)

Figure 3 illustrates what occurs if we take DGP 1 and flip the sign of $\rho$, from 0.80 to -0.80, so that the OLS bias is now negative. Then we run BSD on 10,000 artificial data sets from this process. Note how the association between the BSD estimates and their standard errors flips from negative to positive. As the OLS bias is negative, standard errors tend to be smaller for negative estimates. Furthermore, the power asymmetry also induces negative median bias in the BSD estimator. Here the median estimate is $-.099$, compared to .094 when $\rho$ is positive.

Figure 3: **BSD Inference: DGP 1 with** $\rho = -0.80$




*Notes:* Based on 10,000 replications with 2000 observations each. We plot $\hat{\beta}_1$ against $se(\hat{\beta}_1)$. BSD inference is implemented via the RDHonest R package. Red dots indicate $H_0$:$\beta = 0$ rejected at 5% level.

Finally, it is important to emphasize that the power asymmetry problem is not due to a weak first stage. As we see in Table 1, in DGP 1, the median first stage $F$ is 32 for the RBC runs and 24 for the BSD runs. These differ as the two approaches choose different bandwidths. But in each case the first-stage $F$ statistics are typically well above conventional levels used to reject the null of weak identification in just-identified IV.

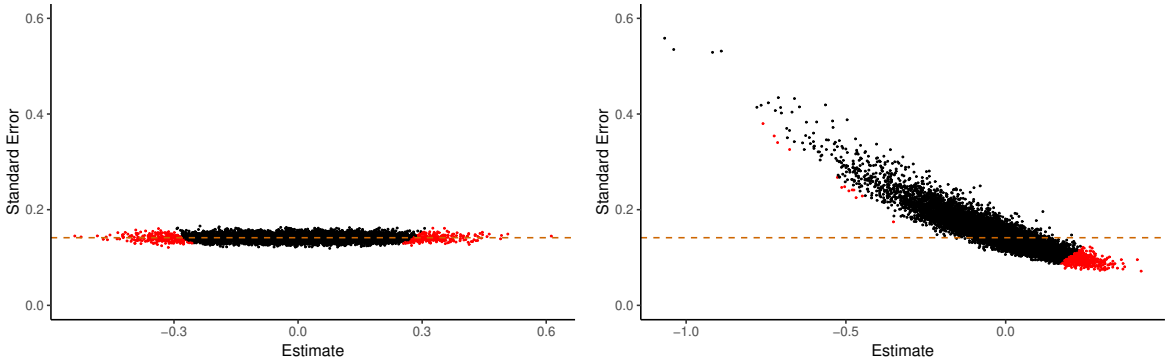
[9] RBC does not suffer from this problem, as it chooses the bandwidth to minimize the MSE of the fuzzy RD estimator based on an analytical formula: The bias depends on $f''(R)$, while the variance is proportional to $1/Nh$. The power asymmetry does not affect this variance calculation, but, in contrast, it does affect CI length.

### *4.2.* ***Comparison with Infeasible Fuzzy RD***

We claim RD estimated standard errors are spuriously small when $\hat{\beta}$ is near OLS. To illustrate this point, we compare them to the best-case standard errors of the "Infeasible Fuzzy RD" estimator that we could construct if the first-stage coefficient $\pi$ was known: If $\pi$ were known *a priori*, we could estimate $\beta$ directly via a sharp RD regression of $Y$ on $\pi D$, controlling for $R$ and $R \cdot D$. Here we apply this estimator to the same 10,000 artificial data sets we just analyzed in Section 4.1. Furthermore, to drive down the standard errors even further, we chose a fixed bandwidth of $0.4 \times \sqrt{3}$, so the median number of observations used in the runs is 800 (i.e., 40% of $N = 2000$). This exceeds the maximum (median) sample size of 799 (508) used by rdrobust, and is greater than 99.9% of the samples selected by RDHonest, which uses a median sample size of only 334.

The left panel of Figure 4 displays the infeasible fuzzy RD results. Notice the estimated standard errors are tightly clustered around 0.141, and they are unrelated to the $\hat{\beta}$ estimates. The true $Var(\hat{\beta})$ is $\sigma_u / \sqrt{800 \cdot Var(\pi D | D, RD)} = 1/\sqrt{800 \cdot 0.0625} = 1/\sqrt{50} = 0.141$.[10] Thus, the estimated standard errors are an accurate reflection of the true parameter uncertainty. This is not surprising, as the infeasible fuzzy RD estimator is based on an OLS regression. We reject $H_0$:$\beta = 0$ at almost exactly the correct 5% rate, and rejections on the positive and negative side are almost perfectly balanced.[11]

Figure 4: **Infeasible Fuzzy RD, and Feasible Fuzzy RD with a Fixed Bandwidth**




*Note:* Based on 10,000 replications with 2000 observations each. We plot $\hat{\beta}_1$ against $se(\hat{\beta}_1)$. *Left:* Infeasible Fuzzy RD regression of $Y$ on $\pi D$. *Right:* Feasible Fuzzy RD regression of $Y$ on $X$ using $D$ as an instrument. $R$ and $D \times R$ are controls in both cases. Both cases use the DGP and parameter values from Case 1 in Section 4, and a fixed bandwidth of $0.4 \times \sqrt{3}$, for a median number of observations of 800. Red dots indicate $H_0 : \beta = 0$ rejected at 5% level. Runs with standard error $> 1$ excluded. The dashed line indicates the true infeasible fuzzy RD standard error of .141, which assumes first-stage $\pi$ is known.

[10] Of course $Var(\pi D) = .25$ because $\pi = 1$ and $D$ is a dummy variable equal to 1 with probability 0.50. But what is relevant for the standard error of $\hat{\beta}$ is the conditional variance $Var(\pi D | D, RD) = 0.25^2 = 0.0625$.

[11] The median infeasible fuzzy RD estimate is -0.002 (i.e., extremely close to zero). The true null that $\beta = 0$ is rejected in 5.12% of runs. Of those, 2.51% occur when $\hat{\beta} > 0$ and 2.61% when $\hat{\beta} \leq 0$, so we have near perfect balance.

Returning to Figure 2, we see that the RBC and BSD standard error estimates are often far below the theoretical lower bound of 0.141, despite using fewer than 800 observations. Even more importantly, this typically occurs when the estimates are shifted towards OLS. That is why almost all significant RBC and BSD estimates occur when $\hat{\beta} > 0$, the direction of the OLS bias.

Are the RBC and BSD standard errors too small when $\hat{\beta}$ is shifted in the direction of the OLS bias because of some problem with the RBC and BSD approaches? (i.e., how they choose the bandwidth or implement bias correction?) Or is this a fundamental problem with fuzzy RD itself?

To address this question, the right panel of Figure 4 shows what happens if we apply "vanilla" fuzzy RD to the same 10,000 datasets, using a fixed bandwidth of $0.4 \times \sqrt{3}$, a uniform kernel, and with no bias correction or standard error adjustment.[12] As we see, the results look very similar to the RBC and BSD results in Figure 2. There is a strong negative association between the Fuzzy RD estimates and their standard errors. As a result, only estimates that are heavily shifted in the direction of the OLS bias are likely to be judged significant by the $t$-test.

To be precise, the median vanilla fuzzy RD estimate is -0.002, with the null $\beta = 0$ rejected in 4.83% of runs, which is close to the correct 5% rate. However, 4.70% of those rejections when $\hat{\beta} > 0$ and only 0.13% occur when $\hat{\beta} \leq 0$. Thus, a one-tailed 2.5% level $t$-test of $H_0 : \beta \leq 0$ rejects at almost twice the nominal rate, while a one-tailed $t$-test of $H_0 : \beta \geq 0$ has almost no power whatsoever.

*These results illustrate that the problem with Fuzzy RD standard errors is fundamental, and not caused by – or solved by – any of the current most popular methods for estimating Fuzzy RDs.*

### *4.3.* ***Case 2: RD Inference in an RCT With Imperfect Compliance***

Next we consider a DGP where treatment $X$ is discrete and there is imperfect compliance:

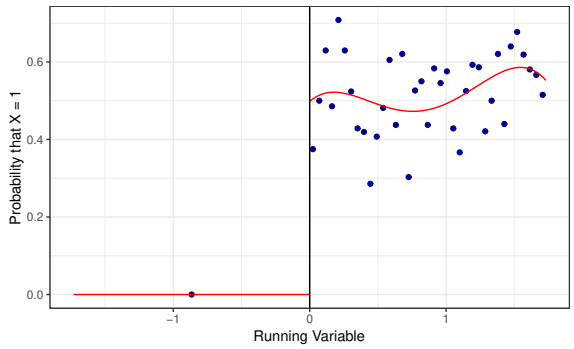
$$
\begin{aligned}
Y_i &= \beta X_i + u_i, \\
D_i &= 1[R_i \geq 0], \\
X_i &= \begin{cases} 0 & if \;\; D_i = 0, \\ 1[\pi - 1 + e_i \geq 0] & if \;\; D_i = 1, \end{cases} \\
e_i &= \rho u_i + \sqrt{1-\rho^2}\eta_i, \\
R \backsim Unif(-\sqrt{3}, \sqrt{3}), &\quad (u, \eta) \sim iidN(0,1)
\end{aligned}
$$

[12] This is simply a 2SLS regression of $Y$ on $X$ with $D$ and $R \cdot D$ as controls, using data on the $(-0.4 \times \sqrt{3} < R < 0.4 \times \sqrt{3})$ interval. So we implement Fuzzy RD via 2SLS just as in equations (3)-(4).

Here individuals with $R$ above a threshold $R_0 = 0$ are accepted into a treatment program, but only a fraction of them chose to participate. We set $\pi = 1$, so the condition for participation ($X = 1$) is $e_i \geq 0$, which is satisfied for approximately half of observations, given that $e$ is $N(0, 1)$. In other words, setting $D = 1$ increases the probability of participation from 0 to 1/2. In this example $R$ is like a lottery number, so again $D = I(R > 0)$ is exogenous.

As in DGP 1, we set the true $\beta = 0$, and $\rho = 0.8$, so the OLS bias from naive regression of $Y$ on $X$ is positive. Here $E(\hat{\beta}_{OLS}) = .85$, increasing to 1.13 if one controls for $R$ and $R \cdot D$. We also set the sample size to $N = 2,000$. Figure 5 displays a representative first-stage RD plot for this process:

Figure 5: **Representative First Stage when X is Discrete with** $F = 54$

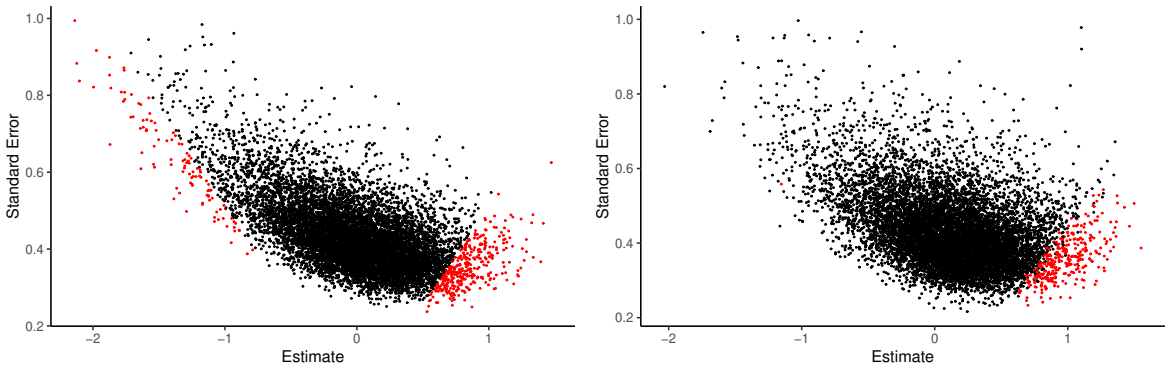



*Notes:* Binned averages of the explanatory variable $X$ in intervals of the running variable $R$ for the 10,000$^{th}$ of 10,000 replications with $N$=2,000 each. Created via the "rdplot" function from the R package rdrobust. Only one bin is plotted left of the cutoff since by design $X_i = 0$ when $R_i < 0$, so all observations fit in a single bin with mean zero.

Figure 6 shows the results of applying RBC and BSD, both with a uniform kernel, to 10,000 artificial data sets from this process. A strong negative association between the estimates and their standard errors is again evident: Positive estimates $\hat{\beta}$ are much more likely to be judged significant by the $t$-test than negative estimate. As we see in Table 1, row 2, RBC inference rejects the true null that $\beta = 0$ at close to the correct 5% nominal rate (4.5%). But 76% of rejections occur when $\hat{\beta}$ is positive, again illustrating the power asymmetry problem. Note that the power asymmetry is less severe here than in DGP 1, presumably because here the median first stage $F$ is higher at 63.

The power asymmetry for BSD is more severe: In the right panel we see it rejects at a 3.2% rate, and 99.7% of these occur when $\hat{\beta} > 0$. RBC and BSD reject on the positive side at about the same rate, but only RBC generates non-negligible rejections on the negative side. As we see in Table 1 the median bias of RBC is essentially zero, while that of RBC is .139. (As we explained in Section 4.1, the power asymmetry generates this bias in BSD.) The positive bias of BSD interacts with the power asymmetry to generate excessive positive rejections, and almost no negative rejections.

Figure 6: **DGP 2: RCT with Imperfect Compliance**




*Notes:* Based on 10,000 replications with 2000 observations each. *Left:* Robust bias-corrected (RBC) inference via the rdrobust R package. *Right:* Bounded second derivative (BSD) inference via the RDHonest R package. Both use a uniform kernel. We plot $\hat{\beta}_1$ against $se(\hat{\beta}_1)$. Runs with a standard error $> 1$ excluded. Red dots indicate $H_0$:$\beta = 0$ rejected at 5% level.

### *4.4. Case 3: RD Inference for a Case with $E[Y|R]$ Linear in $R$*

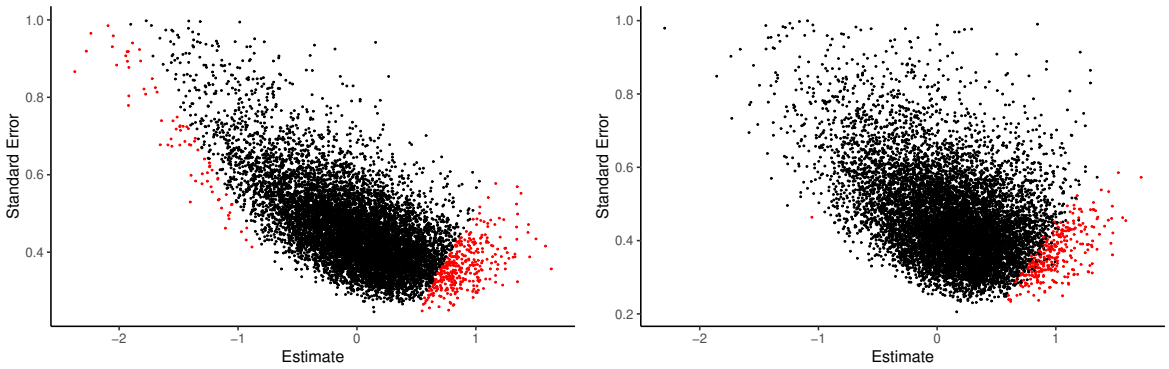
In many regression-discontinuity designs, the object of interest is the impact of admission, $X$, to a school or an educational program, on subsequent test scores or log earnings $Y$, where admission is based on a test score, $R$. The DGP for Case 3 mimics this situation:

$$
\begin{aligned}
Y_i &= \beta X_i + R_i + u_i, \\
D_i &= 1[R_i \geq 0], \\
X_i &= \begin{cases} 0 & if \quad D_i = 0, \\ 1[\pi - 1 + .645R_i + e_i \geq 0] & if \quad D_i = 1, \end{cases} \\
e_i &= \rho u_i + \sqrt{1-\rho^2}\eta_i, \\
R &\backsim N(-1, 1),
\end{aligned}
$$

In this context the test score measures ability, so it affects the outcome $Y$ directly, as well as indirectly through its impact on admission. This violates the IV exclusion restriction, so in contrast to our first two cases, conventional 2SLS would not work in this case. Also, in contrast to the first two examples, here we assume $R$ is normally distributed, as that is typical for distributions of test scores. The test score cutoff is set to $R_0 = 0$, and we set the mean of $R$ to $-1$, so that only students who are at least 1 standard deviation above the mean are admitted. We set $\pi = 1$, so students who are admitted participate in the program if $e + .645R \geq 0$. Thus, students with higher ability $R$ are also more likely to accept if admitted.

In this example, the error terms $u$ and $e$ capture "motivation." We set $\rho = 0.8$, so more motivated students are both more likely to participate in the program and tend to have better outcomes $Y$ regardless. As in the first two DGPs, we set $\beta = 0$. In this setup, $E(\hat{\beta}_{OLS}) = 2.29$, falling to 0.94 if one controls for $R$ and $RD$. We again set the sample size to $N = 2,000$.

Figure 7: **DGP 3: Effect of Admission to Education Program on Wages, $E[Y|R]$ Linear in $R$.**




*Notes:* Based on 10,000 replications with 2000 observations each. *Left:* Robust bias-corrected (RBC) inference via the rdrobust R package. *Right:* Bounded second derivative (BSD) inference via the RDHonest R package. Both use a uniform kernel. We plot $\hat{\beta}_1$ against $se(\hat{\beta}_1)$. Runs with a standard error $> 1$ excluded. Red dots indicate $H_0$:$\beta = 0$ rejected at 5% level.

Figure 7 shows results of applying RBC and BSD, both with a uniform kernel, to 10,000 artificial data sets from this process. The power asymmetry is again evident. Median first stage $F$ is 50 for RBC and 37 for BSD, again illustrating this is not a weak instrument phenomenon. As we see in Table 1, row 3, RBC rejects $H_0$:$\beta = 0$ at at 4.5% rate, and 80% of these rejections occur when $\hat{\beta} > 0$. The power asymmetry for BSD is more severe: It rejects the null at a 3.2% rate, and 99.7% of these rejections occur when $\hat{\beta} > 0$.

Again, the median bias of RBC is essentially zero, but that of BSD is .157. The BSD bandwidth algorithm interacts the power asymmetry to generate this bias, even if the local linear model is correct (so there is no inherent bias in RD). Furthermore, the power asymmetry also causes smaller standard errors on positive $\hat{\beta}$, generating excessive positive rejections and almost no negative rejections.

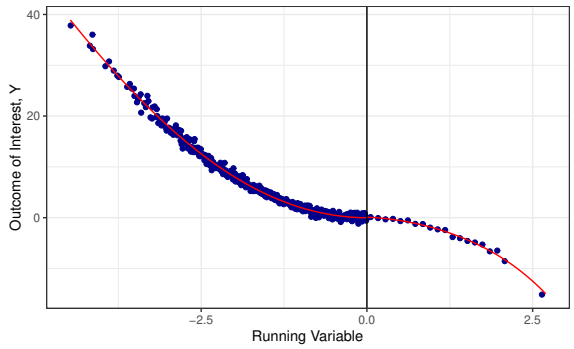
### *4.5. **Case 4: RD Inference for a Case with** $E[Y|R]$ **Quadratic in** $R$*

Our fourth and final DGP mimics admission to a remedial education program. Here $R$ is the *negative* of an ability score, and students with sufficiently high $R$ (hence, low ability) are admitted to the program. We modify DGP 3 so that test score $R$ has a non-linear effect on the outcome $Y$. We assume the effect is quadratic, and that the slope shifts at $R_0$:

$$
\begin{aligned}
Y_i &= \beta X_i + 2R_i^2 - 4D_i R_i^2 + u_i, \\
D_i &= 1[R_i \geq 0], \\
X_i &= \begin{cases} 0 & if \quad D_i = 0, \\ 1[\pi - 1 + 0.645R_i + e_i \geq 0] & if \quad D_i = 1, \end{cases} \\
e_i &= \rho u_i + \sqrt{1-\rho^2}\eta_i, \\
R &\backsim N(-1, 1),
\end{aligned}
$$

Figure 8 illustrates the shape of the conditional expectation function $E[Y|R]$ in this case. Because it is nonlinear, controls for $R$ and $R \cdot D$ will not completely control for variation in $E[Y|R]$ as we move away from $R_0$. However, this case is ideal for RBC inference: If the true conditional expectation function is quadratic, all higher-order bias terms are zero, and so bias correction should on average correct for the bias exactly. Similarly, BSD inference should perform well due as the quadratic form of the true CEF guarantees the boundedness of the second derivative.

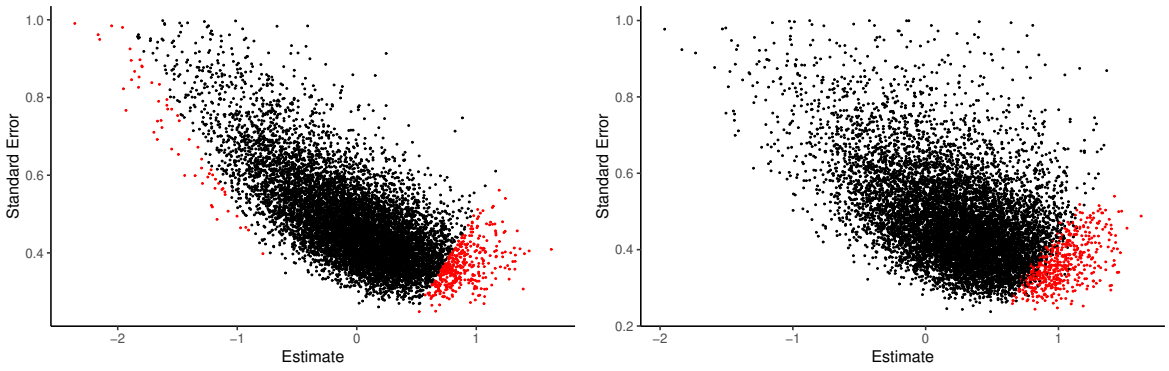
Figure 8: **Representative Quadratic Outcome CEF**




*Note:* Binned averages of the outcome $Y$ in intervals of the running variable $R$, for the 10,000$^{th}$ of 10,000 replications with 2,000 observations each.

We set $\pi = 1$, so students who are admitted participate in the program if $e + .645R \geq 0$. Thus, students with higher $R$ (lower ability) are more likely to accept remediation. The error terms $u$ and $e$ capture "motivation." We set $\rho = 0.8$, so more motivated students are both more likely to participate and tend to have better outcomes $Y$ regardless. We again set the true effect $\beta = 0$. In this setup, a naive OLS regression of $Y$ on $X$ gives $E(\hat{\beta}_{OLS}) = -4.88$, as lower ability students are more likely to have $X = 1$ and also have worse outcomes $Y$. However, if one controls for $R$ and $R \cdot D$, we get $E(\hat{\beta}_{OLS}) = 3.13$. This positive OLS bias is driven by $\rho > 0$. As we will see, it is this positive OLS bias that drives the power asymmetry. We again set the sample size to $N = 2,000$.

Figure 9: **DGP 4: Effect of Admission to Education Program on Wages, $E[Y|R]$ Quadratic in $R$.**




*Notes:* Based on 10,000 replications with 2000 observations each. *Left:* Robust bias-corrected inference (RBC) via the rdrobust R package. *Right:* Bounded second derivative (BSD) inference via the RDHonest R package. Both use a uniform kernel. We plot $\hat{\beta}_1$ against $se(\hat{\beta}_1)$. Runs with a standard error $> 1$ excluded. Red dots indicate $H_0$:$\beta = 0$ rejected at 5% level.

Figure 9 shows results of applying RBC and BSD, both with uniform kernel, to 10,000 artificial data sets from this process. A strong negative association between the estimates and their standard errors is again evident. As we see in Table 1, row 4, RBC rejects $H_0$:$\beta = 0$ at at 4.7% rate, and 84% of these rejections occur when $\hat{\beta} > 0$. The power asymmetry for BSD is more severe: It rejects at a 5.3% rate, and 100% of these rejections occur when $\hat{\beta} > 0$. The median bias of RBC is essentially zero, but that of BSD is a substantial .284. This positive bias again interacts with the power asymmetry to generate excessive positive rejections.

### *4.6.* ***Changing the Sign of*** $\rho$

In this section we briefly report what happens if we change $\rho$ from 0.80 to -0.80, so the OLS bias is now negative. At the end of Section 4.1 we already showed that this flips the median bias of the BSD estimator from positive to negative in DGP 1. Table 2 gives complete results for all four cases.

We see the results are the mirror image of those in Table 1. When the OLS bias is negative, the large majority of rejections of $H_0$:$\beta = 0$ occur when $\hat{\beta} < 0$, due to the power asymmetry problem. Note that the bias of BSD now turns negative in all four cases. As we explained in Section 4.1, choosing the bandwidth by numerically searching over the space of standard errors, when combined with the power asymmetry, causes BSD to choose bandwidths that generate estimates shifted toward OLS, because these estimates have relatively small standard errors.

The one notable difference between the results in Tables 1 and 2 is that in DGP 4 the bias in the BSD estimator shifts from .284 to -.067. So it not only flips sign but also gets smaller in magnitude. The reason is that in DGP 4 the BSD estimator has two sources of bias: (i) the bias towards OLS

induced by the power asymmetry, and (ii) the bias that arises from the mis-specification of the local-linear model, arising because $E[Y|R]$ is quadratic in $R$ in DGP 4. In DGP 4 this second source of bias is positive.[13] Thus, (i) and (ii) reinforce each other when the OLS bias is positive, while partially canceling each other when the OLS bias is negative. Of course there will be fortuitous cases where the two sources of bias happen to cancel, but of greater concern is that the bias may be severe when the two sources of bias reinforce each other – as we see in DGP 4 when $\rho > 0$.

Table 2: **RBC and BSD Inference with Uniform Kernel, DGPs 1 to 4 with $\rho = -0.8$**

| | RBC | | | | | | | BSD | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP | Reject | $\% < 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ | Reject | $\% < 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ |
| 1 | 5.4% | 98.5% | | -.005 | .205 | 32.25 | 508 | 10.1% | 100% | | -.099 | .185 | 24.53 | 334 |
| 2 | 4.8% | 76.7% | | -.006 | .416 | 62.88 | 498 | 3.6% | 99.4% | | -.149 | .414 | 45.08 | 340 |
| 3 | 4.8% | 78.9% | | .002 | .443 | 49.66 | 451 | 3.1% | 99.7% | | -.155 | .437 | 37.27 | 302 |
| 4 | 4.3% | 83.6% | | .013 | .487 | 40.12 | 357 | 2.3% | 98.3% | | -.067 | .469 | 33.72 | 281 |

*Notes:* Summary results from 10,000 artificial datasets of size $N = 2000$ each. The 4 rows report results for the 4 DGPs discussed in Sections 4.1 to 4.5. RBC and BSD indicate results from the rdrobust and RDHonest packages, respectively. Both use a uniform kernel. We report the rate of rejecting $H_0$:$\beta = 0$, the fraction of these rejections that occur when $\hat{\beta} < 0$, and the medians of the estimate, estimated standard error, first stage F, and effective observations.

### *4.7.* ***Using a Triangular Kernel***

As Hahn et al. (2001) pointed out, once a bandwidth is chosen, a local linear regression based on a uniform kernel is equivalent to 2SLS. However, many RD applications instead use a triangular kernel, so this equivalence breaks down. In this section, we show that the power asymmetry problem that affects the fuzzy RD $t$-test is equally important if one uses a triangular kernel.

Table 3 shows results of using a triangular kernel for DGPs 1-4. The RBC results are almost identical to the uniform kernel results in Tables 1 and 2. The number of observations that are used in estimation (shown in the column labeled $N$) increases by about 20 to 25%, as we would expect with a triangular kernel. But the median estimate, standard error and first stage $F$ are little changed. When $\rho = .80$ we see that 75% to 97% of the rejections of the true null $\beta = 0$ occur when $\hat{\beta} > 0$, very close to the figures in Table 1. So the power asymmetry is very similar. The direction of the power asymmetry is reversed when $\rho$ is negative, just as we saw in Table 2.

In contrast, the BSD results are very different: The median bias of the BSD estimates – which was evident when using a uniform kernel – is almost eliminated in cases 1 to 3, and greatly reduced

[13] If we consider the RBC estimates *without* bias correction (i.e., the uncorrected local-linear estimates produced using the RBC MSE-optimal bandwidth), the median estimate is near zero in DGPs 1 to 3, but it is .185 in case 4. So the bias of the local linear estimator in DGP 4 is roughly .185.

in case 4. In Section 4.1 we explained how the power asymmetry interacts with the BSD bandwidth selection algorithm to generate bias. BSD tends to choose bandwidths that generate estimates close to $\hat{\beta}_{OLS}$, as these have smaller standard errors. However, as we explain in Appendix C, this tendency is much weaker when using a triangular kernel, simply because the standard error of the RD estimate varies much less with bandwidth.

On the other hand, the power asymmetry is still very evident for BSD. For all four DGPs over 90% of rejections occur when $\hat{\beta}$ deviates from $\beta = 0$ in the direction of the OLS bias.

Table 3: **RBC and BSD Inference with Triangular Kernel, DGPs 1 to 4**

| $\rho = 0.8$ | RBC | | | | | | | BSD | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ |
| 1 | 5.2% | 96.8% | | -.002 | .208 | 32.61 | 626 | 4.8% | 100% | | .018 | .202 | 28.64 | 414 |
| 2 | 5.2% | 74.7% | | -.005 | .419 | 63.56 | 607 | 2.2% | 96.4% | | .009 | .414 | 55.79 | 411 |
| 3 | 5.2% | 79.0% | | -.001 | .451 | 48.50 | 538 | 2.1% | 97.7% | | .018 | .439 | 44.96 | 373 |
| 4 | 5.1% | 81.1% | | -.002 | .461 | 40.49 | 446 | 3.6% | 99.7% | | .139 | .452 | 41.00 | 373 |
| $\rho = -0.8$ | RBC | | | | | | | BSD | | | | | | |
| Case | Reject | $\% < 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ | Reject | $\% < 0$ | Median | $\hat{\beta}$ | $SE$ | $F$ | $N$ |
| 1 | 5.4% | 97.1% | | -.004 | .206 | 32.92 | 624 | 5.0% | 100% | | -.021 | .200 | 28.86 | 370 |
| 2 | 5.2% | 75.7% | | -.006 | .419 | 63.69 | 607 | 2.1% | 97.6% | | -.019 | .414 | 55.75 | 367 |
| 3 | 5.0% | 81.7% | | -.002 | .451 | 48.50 | 538 | 2.1% | 98.5% | | -.015 | .439 | 44.90 | 390 |
| 4 | 4.8% | 82.4% | | -.002 | .486 | 40.92 | 446 | 1.7% | 90.1% | | .091 | .475 | 41.00 | 365 |

*Notes:* See notes to Table 1 and Table 2. The only change is teh use of a triangular kernel.

### *4.8.* ***Summary of Results***

Each of the four DGPs we considered is favorable to RBC and BSD inference, as the assumptions of these methods are satisfied. Yet for both methods, the power asymmetry causes one-tailed $t$-test size to be distorted, so estimates shifted in the direction of the OLS bias are much more likely to be judged significant by the $t$-test than estimates shifted in the opposite direction. Hence, the estimates for which a two-tailed $t$-test rejects the true null $\beta = 0$ are heavily skewed in the direction of the OLS bias that motivates using RD in the first place.

BSD suffers from an additional problem: The power asymmetry interacts with its bandwidth selection algorithm to induce median bias towards OLS. The problem is severe for the uniform kernel, but minor for the triangular kernel. Hence, it is preferable to use a triangular kernel with BSD.[14]

[14]The RDHonest package also has a minimum MSE bandwidth selection option, that one may choose in lieu of the

## 5. Reduced Form Anderson-Rubin Inference in Fuzzy Regression-Discontinuity Designs

As we explained in Section 2.2 the analogue to the reduced form in RD analysis is a sharp RD regression of the outcome $Y$ on treatment assignment $D = 1[R \geq R_0]$, along with controls for $R$ and $R \cdot D$. This estimates the Intent-to-Treat (ITT) effect $\xi_1 = \beta_1 \pi_1$ rather than the effect of treatment $\beta_1$. The Anderson-Rubin approach to inference in the RD context is to use the $t$-test on $\hat{\xi}_1$ in the ITT regression to assess whether the $\hat{\beta}_1$ is significant in the fuzzy RD regression. We denote this the $t_{AR}$ test. A remarkable fact is that this generates numerically the exact same test that we could form if we knew the first stage coefficient $\pi_1$ *a priori* and could run the "infeasible fuzzy RD regression" of $Y$ on $\xi_1 D$ — see Section 2.3. That is why the $t_{AR}$ has desirable properties.[15] Here we show how this AR-type test performs in the same four cases we examined in Section 4.

Feir et al. (2016) first proposed adopting this AR approach to inference when the first stage of Fuzzy RD is weak. They applied under-smoothing to the ITT regression, and so did not consider bias. Noack and Rothe (2024) proposed to apply the bias aware BSD approach to the ITT regression, using inflated critical values as in equation (10). As in Section 4, we will consider both the BSD approach and the robust bias corrected (RBC) approach. We are not aware of previous Fuzzy RD inference that adopts an AR approach where RBC inference is applied to the ITT regression.

### *5.1.* ***Cases 1 and 2: RCTs with Perfect and Imperfect Compliance***

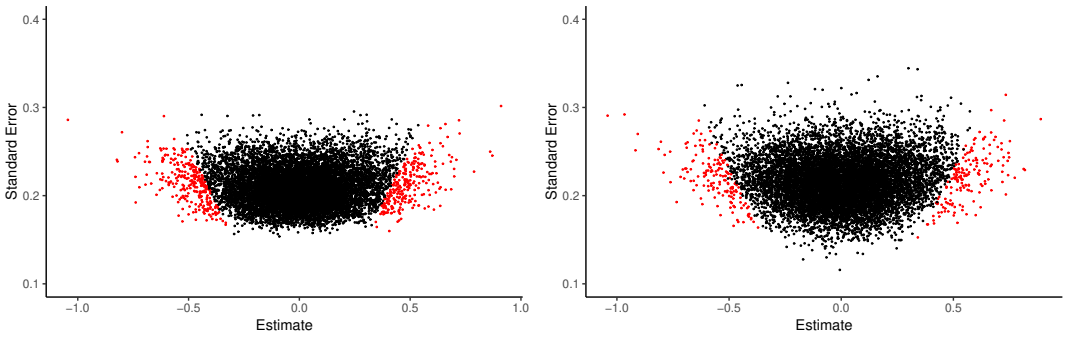
The true reduced from for case 1 is $Y = \beta\pi D + (\beta e + u)$, which, when we set $\beta = 0$ and $\pi = 1$ is simply $y = u$. The true reduced from for case 2 is $Y = \beta D + u$, which, when we set $\beta = 0$ is simply $y = u$. Hence the reduced forms are identical. We use the same draws for $(u, \eta, R)$ in both simulations, so the data for $Y$ and $D$ are identical. Hence each case generates identical reduced form results when we run sharp RD regressions of $Y$ on $D$ along with controls for $R$ and $R \cdot D$.

Reduced form results for DGPs 1 and 2 are shown in Figure 10, which reports results of applying RBC (via rdrobust) and BSD (via RDHonest), both with a uniform kernel, to 10,000 datasets from

minimum CI option. But it defines the variance term in the MSE as the square of the *estimated* standard error of $\hat{\beta}$. Due to the power asymmetry, this variance, and hence the MSE as a whole, tends to be minimized for bandwidths that generate estimates near OLS. In contrast, RDrobust chooses bandwidth to minimize MSE where the variance is defined as $V/nh$, where V is a function of (i) the variance of $Y|R$ near the cutoff and (ii) the density of $R$ near the cutoff. Crucially, these quantities do not vary with the estimated standard error, so the power asymmetry does not induce median bias when bandwidth is chosen to minimize this definition of MSE.

[15] Note that if we run a sharp RD regression of $Y$ on $D$ along with $R$ and $R \cdot D$ as controls we get numerically the same $t$-statistic as if we could run the infeasible regression of $Y$ on $\pi D$ along with $R$ and $R \cdot D$ as controls. The multiplication of $D$ by the scalar $\pi$ does not change the result. Interestingly, one would also get exactly the same result by running a sharp RD regression of $Y$ on $\hat{\pi} D$ along with $D$ and $R \cdot D$ as controls, where $\hat{\pi}$ is the first stage estimate of $\pi$. That is, by running the 2nd stage of fuzzy RD (which is just 2SLS) "by hand" and simply reporting the $t$-test on the estimated coefficient on $\hat{X} = \hat{\pi} D$. All three regressions generate the same $t$-test.

Figure 10: **Reduced Form Inference in RCT DGPs (Cases 1 & 2)**




*Notes:* Based on 10,000 replications with $N$=2000 each. *Left:* RBC inference via the rdrobust R package. *Right:* BSD inference via the RDHonest R package. Both use a uniform kernel. We plot $\hat{\xi}$ against $se(\hat{\xi})$. Red dots indicate $H_0$:$\beta = 0$ rejected at 5% level. For AR inference this corresponds to cases where $\hat{\xi}$ is significant at the 5% level.

this process. We plot the estimated ITT effect $\hat{\xi} = \hat{\beta}\hat{\pi}$ against its estimated standard error. Clearly, there is almost no association between estimates and standard errors in these ITT regressions, in sharp contrast to the Fuzzy RD regressions in Figures 2 and 6. RBC rejects $H_0$:$\beta = 0$ at a 5.47% rate, and 48.3% of these occur when $\hat{\beta} > 0$. BSD rejects at a 3.26% rate, and 50.3% of these occur when $\hat{\beta} > 0$.[16] (Note that, as $\pi = 1$, positive $\hat{\xi}$ correspond exactly to positive $\hat{\beta}$). Thus, rejections are well balanced on the positive and negative side, so the power asymmetry problem is resolved.

### *5.2.* ***Case 3: Reduced Form Inference For a Case with*** $E[Y|R]$ ***Linear in*** $R$

The true reduced form in this case is $Y = \beta D + R + u$, so a sharp RD of $Y$ on $D = 1[R \geq R_0]$ with controls for $R$ and $R \cdot D$ is properly specified. Figure 11 reports the reduced form RD results for this case. Here, RBC inference rejects the false null $\beta = 0$ at a 5.25% rate, and 50.1% of rejections occur when $\hat{\beta} > 0$. Bias-aware inference via RDHonest rejects at a 3.32% rate, and 48.5% of these occur when $\hat{\beta} > 0$. So again, the power asymmetry problem is resolved.[17]

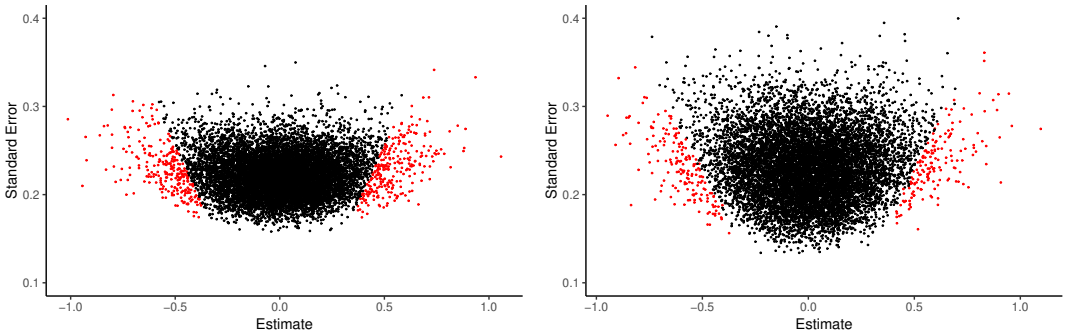
### *5.3.* ***Case 4: Reduced Form Inference for a Case with*** $E[Y|R]$ ***Quadratic in*** $R$

For DGP 4 the true reduced form is $Y = \beta D + 2R^2 - 4DR^2 + u$, so a sharp RD that controls for $R$ and $R \cdot D$ will exhibit bias. However, RBC inference should correct for the bias quite accurately, as higher order bias terms are zero. And BSD should perform well as $M = 4$. Figure 12 shows the results for this case. For RBC the true null $\beta = 0$ is rejected at a 5.28% rate, and 49% of these rejections occur when the estimate is positive. So again the power asymmetry problem is resolved.

[16] Note that the BSD confidence intervals are conservative, so it rejects at less than the nominal 5% rate. This occurs despite the fact that true $M = 0$, as the *estimates* of $|M|$ are positive due to sampling variation.

[17] Again, the BSD confidence intervals are conservative.

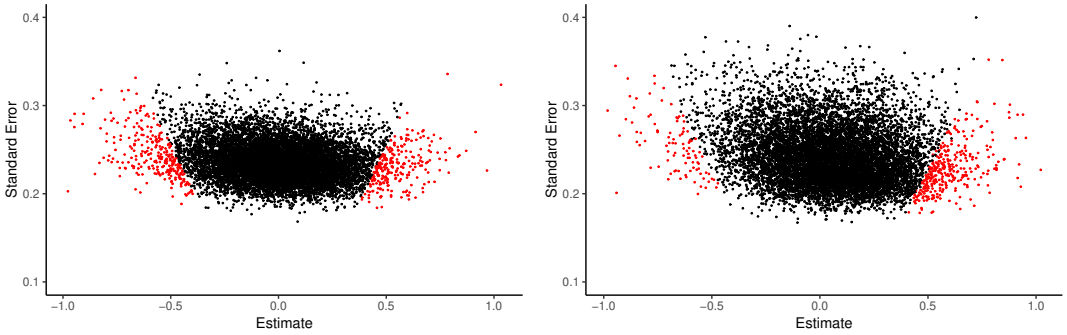
Figure 11: **Reduced Form Inference: DGP 3,** $E[Y|R]$ **Linear in** $R$




*Notes:* Based on 10,000 replications with $N$=2000 each. *Left:* RBC inference via the rdrobust R package. *Right:* BSD inference via the RDHonest R package. Both use a uniform kernel. We plot $\hat{\xi}$ against $se(\hat{\xi})$. Red dots indicate $H_0$:$\beta = 0$ rejected at 5% level. For AR inference this corresponds to cases where $\hat{\xi}_1$ is significant at the 5% level.

In contrast, BSD rejects $\beta = 0$ at a 4.14% rate, which is less than 5% because the BSD confidence intervals are conservative. However, 76.6% of these rejections occur when $\hat{\beta}_1 > 0$. This imbalance is not due to a power asymmetry: As we can see in the right panel of Figure 12, there is no negative association between estimates and standard errors.

Figure 12: **Reduced Form Inference: DGP 4,** $E[Y|R]$ **Quadratic in** $R$




*Notes:* Based on 10,000 replications with $N$=2000 each. *Left:* RBC inference via the rdrobust R package. *Right:* BSD inference via the RDHonest R package. Both use a uniform kernel. We plot $\hat{\xi}$ against $se(\hat{\xi})$. Red dots indicate $H_0$:$\beta = 0$ rejected at 5% level. For AR inference this corresponds to cases where $\hat{\xi}_1$ is significant at the 5% level.

Instead, the source of the imbalance is the bias in the bias-aware estimator. The median bias of the BSD estimates is .062, compared to only .006 for the RBC procedure. Because of this bias, the rate of rejecting $\beta = 0$ on the positive side is 3.2%, which exceeds the nominal 2.5% test of a one-sided test using the same critical values. This illustrates that symmetrically widening confidence intervals, as BSD inference does, does not in general result in a test with symmetric power.

### *5.4.* ***Why Does Inference Via the Reduced Form Avoid the Power Asymmetry?***

Fuzzy RD is a ratio estimator, and the delta-method standard errors $se(\hat{\beta}_1)$ of the Fuzzy RD estimate $\hat{\beta}_1 = \hat{\xi}_1/\hat{\pi}_1$ suffer from a power asymmetry (i.e., they tend to be too small when $\hat{\beta}_1$ is near to OLS), because the standard error of the Fuzzy RD regression is minimized at the OLS estimate – see Section 2.4. In contrast, the ITT estimate $\hat{\xi}_1$ and its standard error can be obtained directly from a sharp RD regression, so no ratio is involved. The standard error of the sharp RD regression is minimized at the sharp RD estimate, thus avoiding the key source of the power asymmetry. And fortunately, the $p$-value for the ITT estimate is also the correct $p$-value for Fuzzy RD estimate $\hat{\beta}_1$, so long as the key assumption that $\pi_1 \neq 0$ is satisfied (since then $\beta_1\pi_1 = 0$ if and only if $\beta_1 = 0$).[18]

### *5.5.* ***AR Confidence Intervals in Fuzzy RDs***

Valid confidence intervals for the Fuzzy RD estimator $\hat{\beta}_1$ cannot be constructed using the usual formula $\hat{\beta}_1 \pm cv_{1-\alpha} \times se(\hat{\beta}_1)$, where $se(\hat{\beta}_1)$ is the Fuzzy RD standard error, as the $t$-stat does not have a symmetric $N(0,1)$ distribution under the null (due to the power asymmetry). Instead, one should obtain a non-symmetric confidence interval by "inverting" the AR test to find the set of values that cannot be rejected. It is straightforward to do this manually. For any hypothesized $\beta_h$, we run the sharp RD regression of $Y - \beta_h X$ on $D$, controlling for $R$ and $R \cdot D$. If $D$ is significant (insignificant) we can (cannot) reject that $\beta_h$ is in the confidence set. Running such regressions on a grid of $\beta_h$ values, one seeks the values $\beta_U$ and $\beta_L$ that form the upper and lower bounds of the set.

Feir et al. (2016) and Noack and Rothe (2024) show how AR confidence intervals can also be formed analytically by finding the set of hypothesized values $\beta_h$ that satisfy the quadratic inequality:

$$(\hat{\xi_1} - \beta_h\hat{\pi}_1)^2 - cv_{1-\alpha}^2 \times (\hat{\sigma}_{RF}^2 + \beta_h^2\hat{\sigma}_{FS}^2 - 2\beta_h\hat{\sigma}_{RF,FS}) \leq 0 \tag{11}$$

where $\hat{\sigma}_{RF}$ and $\hat{\sigma}_{FS}$ denote the standard errors of the sharp RD estimators of $\xi_1 = \beta_1\pi_1$ and $\pi_1$, respectively, and $\hat{\sigma}_{RF,FS}$ denotes their covariance. If $\beta_h = 0$ this reduces to checking if the ITT estimate $\hat{\xi}_1$ is significantly different from 0 in the reduced form, using the $t_{AR}$ test.

The difference between the AR confidence intervals in Noack and Rothe (2024) and Feir et al. (2016) is that Noack and Rothe (2024) use the BSD approach, so they input BSD estimates and the inflated critical values from (10) into (11). One may also use the RBC estimates of the ITT regression to form the CI interval in (11), by plugging in RBC estimates and standard errors.

[18] Apart from avoiding the power asymmetry, Anderson-Rubin $p$-values are preferable to 2SLS $t$-test $p$-values for two further reasons. First, they are robust to weak instruments, since they don't depend on the the first-stage $F$-statistic (Feir et al., 2016; Keane and Neal, 2024, 2023). Second, among unbiased tests, the AR test is optimal (i.e., it is the uniformly most powerful test; see Moreira 2009 and our discussion at start of Section 5).

## 6. Example: Long Run Impact of the 1854 Broad Street Cholera Outbreak on Rents

Here we revisit an interesting RD application to illustrate the empirical relevance of the issues we have discussed. Ambrus, Field, and Gonzalez (2020) study the long-term impact of the 1854 Broad Street cholera outbreak in Soho, London, on local real estate values, as measured by rents. They find evidence that rents were still depressed 10 years later. The treatment effect is identified by the discontinuous change in rents at the boundary of the catchment area of the Broad Street water pump ('BSP'), which drew water from a contaminated well, and was responsible for transmitting cholera throughout the catchment. They estimate the ITT effect of being inside the BSP catchment area on rents in 1864 by estimating the sharp RD regression:

$$ln\, y_i = \xi_0 + \xi_1 BSP_i + \xi_2 R_i + \xi_3 BSP_i \cdot R_i + \boldsymbol{W}_{it}^{'}\boldsymbol{\xi_4} + \epsilon_i \quad \text{for } -h < R_i < h \tag{12}$$

where $ln\, y_i$ is the log rent of property $i$ in 1864, $BSP_i$ is 1 if the property falls within the catchment area, the running variable $R_i$ is the distance in meters to the boundary, $\boldsymbol{W}_{it}^{'}$ is a vector of control variables, and $h$ is the bandwidth. They estimate this local linear regression using the Calonico et al. (2014b) algorithm to determine $h$, along with a triangular kernel.

Ambrus et al. (2020) also report Fuzzy RD estimates of the effect of having at least one cholera death in property $i$ during the outbreak, which we denote by $CD_i$, on rent in 1864. In these regressions $BSP_i$ is the instrument for $CD_i$. [Hence, in our earlier notation, $D = BSP$ while $X = CD$.] Out goal here is to assess (i) how these Fuzzy RD results are affected by the power asymmetry issue, and (ii) if the ITT results are more reliable.

Ambrus et al. (2020) report the Fuzzy RD result in Table B2 column 1 of their article, and the ITT result in Table 3, panel B column 2. We can replicate these results exactly. Notably, Ambrus et al. (2020) use a larger set of controls $W$ in the ITT regression than in the Fuzzy RD regression. We call these the 'Basic' and 'Full' control sets.[19] Furthermore, they do not implement bias correction. We present ITT and Fuzzy RD estimates in Table 4, with and without bias correction, using only the Full set of controls. We also report what we call the 'OLS' local linear regression of $y$ on $CD$ for $|R| < h$, along with controls for $R$, $R \cdot BSP$ and $W$. This OLS regression ignores potential endogeneity of $CD$ that Fuzzy RD using $BSP$ as an instrument is meant to correct.

The 'OLS' regression of $ln\, y$ on $CD$ suggests that a death in a property in 1854 is associated with a small 5% drop in rent in 1864, but the point estimate is not significant. In contrast, the

[19] The Basic set includes: distance (m/100) to the nearest water pump, urinal, Soho centroid, and access to the old/existing sewer. The Full set also adds: distance (m/100) to the public square, fire station, theater, police station, pub, church, bank, presumed plague pit, sewer vent, and whether the property has no access to the sewer.

Fuzzy RD results imply much larger negative effects. For example, the bias corrected estimate is -.829, which implies that a death in the property leads to a 56% drop in rent. But the standard error on the estimate is .612, leading to a $t$-stat of 1.35 (p=.176), so it is not significant at conventional levels.[20] However, ***our analysis in Section 4 indicates this is exactly the type of situation where the fuzzy RD standard error is inflated***. The OLS bias is positive and substantial, so the $t$-test has little power to detect a true negative effect due to the power asymmetry problem.[21]

Table 4: **Impact of Cholera Death on Rent in 1864**

| | BC | Coefficient | Std. Err. | p-value | $h$ (m) | F-stat |
|---|---|---|---|---|---|---|
| 'OLS' – $y$ on $CD$ | | -.046 | .046 | 0.32 | 29.1 | |
| 1st Stage: $CD$ on $BSP$ | | .206 | .068 | .002 | 28.2 | 9.19 |
| 1st Stage: $CD$ on $BSP$ | Yes | .232 | .082 | .005 | 28.2 | 8.04 |
| Fuzzy RD | | -.964 | .531 | .070 | 29.1 | |
| Fuzzy RD | Yes | -.829 | .612 | .176 | 29.1 | |
| ITT | | -.199 | .089 | .025 | 31.5 | |
| ITT | Yes | -.221 | .110 | .044 | 31.5 | |

*Note: Sample size is N=1325 in all regressions. This is the sample Ambrus et al. (2020) use to obtain their Fuzzy RD estimate of -.964. 'BC' indicates bias corrected estimates. All regressions use the 'Full' set of controls, a triangular kernel, and clustered (by street) nearest neighbor standard errors.*

As we argued in Section 5, an AR approach to inference based on the reduced-form ITT regression is preferable, as it avoids the power asymmetry. The AR approach uses the $t$-test from the ITT regression, rather than the Fuzzy RD $t$-test, to assess significance of the Fuzzy RD estimate. We call this the $t_{AR}$ test. Our ITT estimate without bias correction is -0.199, with a standard error of .089, giving a $t_{AR}$-test value of -2.24 (p=.025). If we implement bias correction, the ITT point estimate increases slightly to -.221, and the standard error, which factors in extra variability introduced by bias correction, increases to .110. This gives a $t_{AR}$-test value of -2.01 (p=.044).[22] If we invert the AR test using the grid search method and RBC results, we obtain a 95% CI of [-3.123, -0.025] for $\beta_1$, which excludes zero. If we instead use the analytic formula in (11) we obtain [-2.750, -0.016].

[20] In their paper, Ambrus et al. (2020) only report Fuzzy RD using the much smaller 'Basic' set of controls and no bias correction. The estimate is -.799 with a standard error of .483 (p=.10).

[21] Ambrus et al. (2020) do not discuss why deaths are endogenous. But there are two likely reasons: First, larger properties would have higher rent, and be more likely to have a death, as they house more residents. The control $W$ does not include property size. Using BSP as an instrument for $CD$ removes any influence of individual property size on the Fuzzy RD estimate, eliminating this source of endogeneity. Second, Figures 2A and 3A of their article reveal that both rents and the probability of a death rise as one gets closer to the BSP, so houses in the hardest hit areas were more valuable. Focusing on properties near the boundary avoids the problem created by the price gradient.

[22] Ambrus et al. (2020) use a slightly larger sample of $N$=1,357 to estimate the ITT regression. In Table 3, panel B column 2 they report an ITT estimate without bias correction of -.186 with a standard error of .089, giving a $t_{AR}$-test value of -2.09 (p=.036). We run all regressions (Fuzzy RD, ITT, first stage and OLS) on the exact same sample size of $N$=1,325 that Ambrus et al. (2020) use to estimate their Fuzzy RD regression.

Thus, the $t_{AR}$ test implies the Fuzzy RD estimate is significant at conventionally accepted levels, while the Fuzzy RD $t$-test implies it is not. What are we to conclude about the impact of cholera deaths on rents? It depends on whether the AR or Fuzzy RD $t$-test is more reliable in this data environment.[23] Our analysis in Sections 4 and 5 already suggests the AR test is more reliable. But to investigate this question further in the specific context of the Ambrus et al. (2020) data, we conduct the following Monte Carlo experiment:

Starting from their full sample of $N$=1,325 observations, we "bootstrap" a new artificial dataset by sampling 1,325 observations with replacement. We repeat this process 10,000 times to obtain 10,000 artificial datasets. As our artificial datasets are sampled with replacement from the original Ambrus et al. (2020) data, that data is the "population" from which the artificial datasets are drawn. Then for each artificial dataset, we construct the bias corrected Fuzzy RD and ITT estimates (using a triangular kernel as in the paper). Table 5 summarizes the results.

Table 5: **RBC Results from Monte Carlo Bootstrap Samples**

| | Fuzzy RD (BC) | | | ITT (BC) | | | First Stage |
|---|---|---|---|---|---|---|---|
| | Coefficient | S.E. | $h$ (m) | Coefficient | S.E. | $h$ (m) | $F$ Statistic |
| 'Population' | -.829 | .612 | 29.1 | -.221 | .110 | 31.4 | 8.04 |
| Median | -.699 | .574 | 27.0 | -.199 | .103 | 27.5 | 6.525 |
| Mean | 95.873 | 709.28 | 27.5 | -.197 | .104 | 27.9 | 9.200 |
| Std. Dev. | 6,947 | 50,593 | 5.6 | .110 | .019 | 5.1 | 10.4 |

*Note: $N = 1,325$ for each of the $10,000$ samples used to form the results. All regressions use the 'Full' set of controls and robust bias correction. The RD estimator calculates a new optimal bandwidth in each artificial dataset. The column labeled 'h' reports features of the distribution of bandwidths (in meters).*

The mean ITT estimate is -.197, which is close to the ITT estimate on the population data reported in Table 4 (-.221). Some deviation is expected, as the sharp RD estimator used to obtain the ITT estimate is consistent, but not unbiased in finite samples. This follows from the fact that the local-linear estimators of $Y^+$ and $Y^-$ are consistent but not unbiased in finite samples. The empirical standard deviation of the ITT estimates across the 10k datasets is .110. This is identical to the estimated standard error on the 'population' data, and very similar to the mean of the estimated standard errors across the 10k runs (.104). Thus, the estimated standard errors of the ITT estimates provide a good guide to the actual variability of the ITT estimates.

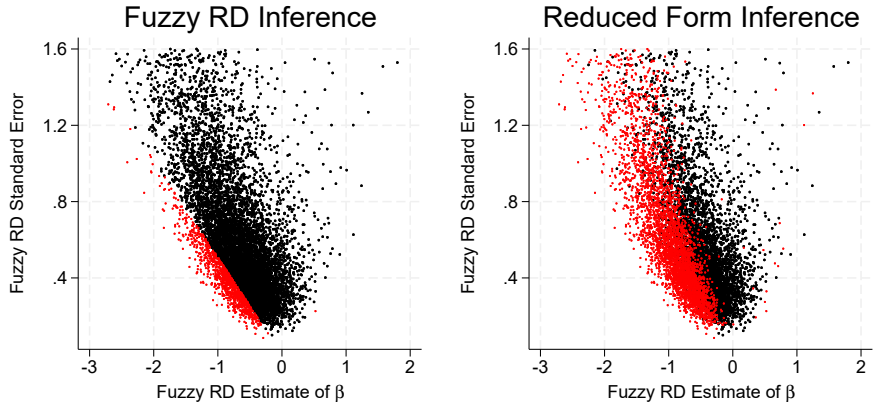
[23] In the IV literature it is common to use first-stage $F$ statistic to assess the reliability of 2SLS $t$-test results: In fuzzy RD the first stage is a sharp RD regression of $CD$ on $BSP$, which we also report in Table 4. The first-stage $F$ statistic is 8.04. Conventional wisdom suggests this should be sufficient for size distortion in the fuzzy RD $t$-test to be modest. But size distortion is not the only issue one should be concerned about. As as we saw in Section 4, the power asymmetry can be a serious problem even when $F$ is over 60.

Turning to the Fuzzy RD estimates, we begin by noting that the mean and variance of the estimator do not exist, just as they do not exist for 2SLS with a single instrument. The problem is that the estimate of $\pi_1 = X^+ - X^-$ can be arbitrarily close to zero in a finite sample, causing the ratio estimator $\hat{\beta}_1 = \frac{\hat{Y}^+ - \hat{Y}^-}{\hat{X}^+ - \hat{X}^-}$ to explode. We see this pathology illustrated by the very large mean (95.9) and empirical standard deviation (6,947) of the Fuzzy RD estimates shown in Table 5.

The median Fuzzy RD estimate is -0.699, which is fairly close to the 'population' value of -0.829. Again, some deviation is expected, as the Fuzzy RD estimator is consistent, but not median unbiased in finite samples. We also note that the median estimated standard error (0.574) greatly understates the observed variability of the Fuzzy RD estimates across the runs (as expected).

Next, we use the simulation results to assess whether the Fuzzy RD $t$-test or the $t_{AR}$ test based on the reduced form ITT regression is a better guide to inference in this context. Figure 13 plots the bias-corrected Fuzzy RD estimate of the effect of a cholera death on rent against the estimated standard error, across the 10,000 bootstrapped samples, using the triangular kernel as in the paper. The plot clearly illustrates the power asymmetry phenomenon – i.e., the strong *negative* association between the estimates and their standard errors (that arises because the OLS bias is *positive*).

Figure 13: **Standard Error of $\hat{\beta}_{2SLS}$ plotted against $\hat{\beta}_{2SLS}$ itself (RBC Inference)**



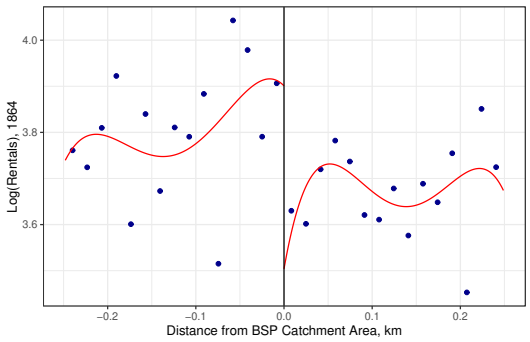

*Note: Runs with standard error > 1.6 are not shown. In the left panel, red dots indicate $H_0$:$\beta = 0$ is rejected at the 5% level in the Fuzzy RD regression using RBC inference. In the right panel red dots indicate $H_0$:$\beta = 0$ is rejected at the 5% level by RBC inference on the ITT. In both cases, the bias corrected estimator is used.*

In the left panel of Figure 13 we shade in red cases where the Fuzzy RD $t$-test rejects the false null that deaths have no effect. Due to the power asymmetry, negative estimates have inflated standard errors, and so the $t$-test has little power to reject the false null hypothesis. The rejection rate is only 10.1%. In contrast, in the right panel of Figure 13 we shade in red cases where the AR test rejects

the null. Here the rejection rate is 47.9%. Thus, the AR test has almost 5 times the power of the $t$-test in this data environment. Thus, we conclude the AR test is a better guide to inference in this context than the $t$-test. This strengthens the conclusion of the original Ambrus et al. (2020) article.

Finally, we repeat the same analysis using BSD inference via the RDHonest package for R, again with a triangular kernel. Recall BSD uses conventional point estimates (no bias correction), but it inflates the $t$-test critical values as shown in (10). BSD generates a Fuzzy RD estimate of -.674 with standard error of .507 ($t$=1.33). Furthermore it inflates the critical value to 2.191, so the p-value is .239. For the ITT regression, BSD gives -.183, with a standard error of .099, so the $t_{AR}$-stat is 1.844. The critical value is inflated to 2.254, giving a $p$-value of .112. The reason BSD is so conservative is that it estimates $M$ to be very large at 8.55, as the CEF is quite "wavy," as we see in Figure 14:

Figure 14: **Curvature of the Conditional Expectation Function Near the Cutoff**




*Notes:* Binned averages of log(rent) in intervals of the running variable (distance from BSP catchment area, in km). Created via rdplot function in rdrobust for R. A $4^{th}$ order polynomial fit is shown either side of the cutoff.

Returning to our Monte Carlo experiment, we find the Fuzzy RD $t$-test using BSD's inflated critical values rejects the false null of a zero effect in only 1.85% of the 10,000 artificial datasets. So power does not approach the nominal 5% level of the test. In contrast, applying BSD inference to the ITT sharp RD regression rejects the false null in 37.1% of datasets. Thus, the power of the $t_{AR}$ test is just over 20 times the power of the Fuzzy RD $t$-test when using BSD critical values.

## 7. Conclusion

Fuzzy RD estimates are obtained via a procedure closely analogous to two-stage least squares (2SLS) regression, where an indicator $I(R > R_0)$ plays the role of the instrument. Recently, Keane and Neal (2023, 2024) showed that 2SLS $t$-tests suffer from a "power asymmetry" problem: 2SLS standard errors are spuriously small (large) when the 2SLS estimate is close to (far from) the OLS estimate. Hence, 2SLS $t$-tests are more likely to judge a result significant if it aligns with the

direction of OLS bias. And they have low power to detect effects that go against the direction of OLS bias. Here we have shown that the same problem arises in Fuzzy RD designs:

In the Fuzzy RD context, the analog of OLS is a local linear regression of the outcome on the endogenous treatment, ignoring the endogeneity problem that the instrument $I(R > R_0)$ is meant to correct. We have shown that, similar to 2SLS, Fuzzy RD $t$-tests are more likely to judge a result significant if it aligns with the direction of OLS bias. If the OLS bias is positive, then the Fuzzy RD $t$-test has little power to detect true negative effects. And, if the OLS bias is positive and the true effect is zero, then the Fuzzy RD $t$-test has inflated power to find false positive effects. This power asymmetry problem persists even if the instrument (first stage) is very strong.

Fortunately, a simple way to avoid this problem is to instead rely on the intent to treat (ITT) regression to assess significance of the treatment effect, where the ITT regression is simply a sharp RD of the outcome on $I(R > R_0)$. This is analogous to the AR approach in 2SLS. The construction of AR confidence sets for Fuzzy RD is discussed in Feir et al. (2016) and Noack and Rothe (2024).

We illustrate the importance of these issues by revisiting the Ambrus et al. (2020) study of the impact of the Broad Street cholera outbreak on rents. They find a large Fuzzy RD estimate of the impact of a death in a property on its rent 10 years later. However, the OLS estimate is close to zero, so Fuzzy RD and OLS are very far apart. The Fuzzy RD $t$-stat indicates the effect of death on rent is statistically insignificant, but this is precisely the context where the Fuzzy RD standard error is inflated. We find the AR test has about 5 times the power of the $t$-test, and it indicates the Fuzzy RD estimate is significant. Thus, our analysis strengthens the result in the original paper.

In future work, we recommend that all Fuzzy RD papers should report not just the Fuzzy RD estimate and $t$-test, but also the first stage, OLS and ITT estimates, and the AR test results. Without seeing the OLS results, it is not possible to assess the likely direction of the power asymmetry in the $t$-test. We also suggest that researchers not use under-smoothing or local quadratic regressions, as these weaken the first stage and make the power asymmetry and size distortions worse. Robust bias correction (RBC) and/or bias aware inference (BSD) should be used instead.

We also recommend that BSD inference should always be done using the triangular kernel rather than the uniform. As we have seen, the bandwidth selection algorithm of BSD interacts with the power asymmetry to generate median bias toward OLS in BSD estimates, even when the local linear model is properly specified. This problem is severe with a uniform kernel but is largely avoided with a triangular kernel. The triangular kernel also smooths the objective function used to search for the optimal bandwidth, making the optimization far more reliable.

## References


Ambrus, A., E. Field, and R. Gonzalez (2020). Loss in the time of cholera: Long-run impact of a disease epidemic on the urban landscape. *American Economic Review 110*(2), 475–525.

Anderson, T. W. and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of mathematical statistics 20*(1), 46–63.

Armstrong, T. B. and M. Kolesár (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics 11*(1), 1–39.

Beckert, W. and D. Kaliski (2024). Honest inference for discrete outcomes in regression discontinuity designs. Unpublished manuscript.

Bound, J., D. A. Jaeger, and R. M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American statistical association 90*(430), 443–450.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018). On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association 113*(522), 767–779.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014a). Robust data-driven inference in the regression-discontinuity design. *The Stata Journal 14*(4), 909–946.

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

Calonico, S., M. D. Cattaneo, and R. Titiunik (2015). Rdrobust: an r package for robust nonparametric inference in regression-discontinuity designs. *R J. 7*(1), 38.

Cattaneo, M. D. and R. Titiunik (2022). Regression discontinuity designs. *Annual Review of Economics 14*(1), 821–851.

DesJardins, S. L. and B. P. McCall (2014). The impact of the gates millennium scholars program on college and post-college related choices of high ability, low-income minority students. *Economics of Education Review 38*, 124–138.

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

Feir, D., T. Lemieux, and V. Marmer (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics 34*(2), 185–196.

Ganong, P. and S. Jäger (2018). A permutation test for the regression kink design. *Journal of the American Statistical Association 113*(522), 494–504.

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), 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), 201–209.

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

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

Keane, M. and T. Neal (2023). Instrument strength in iv estimation and inference: A guide to theory and practice. *Journal of Econometrics 235*(2), 1625–1653.

Keane, M. P. and T. Neal (2024). A practical guide to weak instruments. *Annual Review of Economics 16*, 185–212.

Kolesár, M. (2024). RDHonest R Code. https://github.com/kolesarm/RDHonest/tree/master/R. Accessed on 2025-03-28.

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

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

Ludwig, J. and D. L. Miller (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly journal of economics 122*(1), 159–208.

Moreira, M. J. (2009). Tests with correct size when instruments can be arbitrarily weak. *Journal of Econometrics 152*(2), 131–140.

Noack, C. and C. Rothe (2024). Bias-aware inference in fuzzy regression discontinuity designs. *Econometrica 92*(3), 687–711.

Stock, J. H., J. H. Wright, and M. Yogo (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics 20*(4), 518–529.

## Appendix A. Details on BSD Confidence Intervals

The standardized bias function $\psi(M,h)$ in equation (10) in the main text takes the form:

$$\psi(M,h) = (M/2)g(h)/se(\hat{\beta}_1), \tag{A.1}$$

where $g(h) = \sum_{i=1}^{N_h}[-w_i sgn(R_i)] \cdot R_i^2$, and where $[-w_i sgn(R_i)]$ is a positive weight – see Kolesár and Rothe (2018) for details. The function $g(h)$ is positive and increasing in bandwidth $h$, as the sum of squares of the running variable is increasing in $N_h$. Then the confidence interval is:

$$C_{1-\alpha} = (\hat{\beta}_1 \pm cv_{1-\alpha}(\psi(M,h)) \times se(\hat{\beta}_1)), \tag{A.2}$$

where $\hat{\beta}_1$ and $se(\hat{\beta}_1)$ are the "conventional" local-linear estimate and nearest-neighbor standard error. Note that $se(\hat{\beta}_1)$ tends to decrease with $\sqrt{h}$, while $cv_{1-\alpha}(\psi(M,h))$ increases with $h^2$, causing their product to have a U-shape. Thus one can search for the $h$ that minimizes CI length.

## Appendix B. The Power Asymmetry in Conventional Fuzzy RD Estimates

Here we consider "Conventional" Fuzzy RD inference, meaning we do not implement any bias correction or CI adjustment. We consider two local linear estimators, using the MSE-optimal bandwidth algorithms proposed in Imbens and Kalyanaraman (2012) and Calonico et al. (2014b). We refer to these as the IK and CCT approaches, respectively. (Results using the CCT bandwidth and no bias correction are labeled "Conventional" in rdrobust output, which is why we adopt that terminology.) Table B1 presents the results for the same four DGPs we considered in Section 4, in all cases using a uniform kernel and nearest-neighbor standard errors.

Table B1: **Conventional Inference using CCT & IK Bandwidths, DGPs 1 to 4**

| $\rho = 0.8$ | CCT | | | | | | | IK | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ |
| 1 | 5.5% | 99.8% | | -0.003 | 0.177 | 32.04 | 508 | 6.0% | 96.5% | | -0.002 | 0.132 | 55.98 | 868 |
| 2 | 5.0% | 76.5% | | -0.004 | 0.356 | 63.17 | 498 | 6.0% | 69.3% | | -0.004 | 0.268 | 110.58 | 868 |
| 3 | 4.8% | 83.6% | | 0.005 | 0.380 | 49.59 | 451 | 5.4% | 77.2% | | 0.007 | 0.299 | 76.15 | 722 |
| 4 | 10.2% | 97.9% | | 0.185 | 0.415 | 40.00 | 357 | 11.3% | 98.0% | | 0.201 | 0.404 | 42.02 | 377 |
| $\rho = -0.8$ | CCT | | | | | | | IK | | | | | | |
| DGP | Reject | $\% < 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ | Reject | $\% < 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ |
| 1 | 6.0% | 99.7% | | -0.003 | 0.175 | 32.25 | 508 | 6.2% | 97.3% | | -0.002 | 0.131 | 56.45 | 868 |
| 2 | 4.9% | 79.4% | | -0.003 | 0.356 | 62.88 | 498 | 6.5% | 70.7% | | -0.004 | 0.268 | 110.34 | 868 |
| 3 | 5.0% | 83.1% | | 0.005 | 0.380 | 49.66 | 451 | 5.1% | 75.7% | | 0.007 | 0.301 | 75.82 | 722 |
| 4 | 4.9% | 43.6% | | 0.185 | 0.443 | 40.12 | 357 | 5.0% | 30.4% | | 0.202 | 0.433 | 42.34 | 377 |

*Notes:* Summary results from 10,000 artificial datasets of size $N = 2000$ each. CCT indicates results from the rdrobust package, using "Conventional" estimates and standard errors. IK indicates results from 2SLS estimation using the Imbens-Kalyanaraman bandwidth. The 4 rows reports results for the 4 DGPs discussed in Sections 4.1 to 4.5. We report the rate of rejecting $H_0$:$\beta = 0$, the fraction of these rejections that occur when $\hat{\beta} > 0$, and the medians of the estimate, estimated standard error, first-stage $F$, and effective observations.

For DGPs 1 to 3 the CCT results are very similar to the RBC results in Tables 1 and 2. This is because the local linear model is correctly specified in these cases, so there is no bias. The power asymmetry we see here is very similar to what we saw in the main text: The large majority of rejections of the true null $\beta = 0$ occur when $\hat{\beta}$ is shifted in the direction of the OLS bias, which is positive in the top panel ($\rho > 0$) and negative in the bottom panel ($\rho < 0$).

In DGP 4 the true CEF is quadratic, so the local linear estimator suffers from substantial positive bias. When the OLS bias is positive (top panel) the bias and power asymmetry reinforce each other,

so almost all rejections occur when $\hat{\beta} > 0$. When the OLS bias is negative (bottom panel) the bias and the power asymmetry counteract each other, so rejections are fairly evenly balanced. Of course this is a fortuitous coincidence – one should be more concerned with the worst case in the top panel.

In terms of the power asymmetry, results using the IK bandwidth algorithm are similar to the CCT results. But the IK results differ in one notable way: In DGPs 1 to 3, where the local linear estimator is correctly specified, the IK algorithm chooses substantially wider bandwidths than the CCP algorithm, leading to more effective observations and smaller standard errors. Interestingly, in DGP 4, where bias is present, the IK bandwidths shrink so they are only slightly larger than the CCT bandwidths, as the algorithm seeks to reduce bias.[24]

*Appendix B.1.* ***Undersmoothing to Reduce Bias***

Undersmoothing is a method that is often used in an attempt to reduce bias in conventional (i.e., not bias corrected) RD. The idea is to choose a bandwidth that is narrower than is MSE-optimal, as the bias of local linear estimators is increasing in bandwidth. Here, we examine the impact of undersmoothing on the power asymmetry. We consider the same CCT and IK bandwidths as in Table B1, except now we divide the bandwidth by 2. Table B2 presents the results.

The key message of Table B2 is that undersmoothing makes the power asymmetry problem worse. The vast majority of rejections of $H_0$:$\beta = 0$ occur when $\hat{\beta}$ is shifted in the direction of the OLS bias. In DGPs 1 to 3 this asymmetry in rejections is even more severe than in Table B1. The reason that undersmoothing makes the power asymmetry problem worse is simply that it reduces

[24]The reason the CCT bandwidths tend to be smaller than the IK bandwidths is as follows: Note that the CCT bandwidths $h_{CCT}$ satisfy the condition $Nh_{CCT}^5 \to 0$, whereas $Nh_{IK}^5 \to C > 0$. Both bandwidths can be written in the form $\hat{h} = \hat{C}^{1/5}N^{-1/5}$, where $\hat{C} = \hat{V}/(\hat{B} + \hat{R})$, where $\hat{V}, \hat{B}, \hat{R}$ are estimators of the variance and bias of the estimator and a regularization term, respectively (with the latter included so that the bandwidth is well-behaved when $\hat{B}$ is near-zero). If we set $\hat{B} = 0$, then the reason that the IK bandwidth is larger in finite samples can be explained with reference to the variance estimators $\hat{V}$. $\hat{V}_{IK} = \nu_K \cdot (\hat{\sigma}_+^2 + \hat{\sigma}_-^2)/\hat{d}$, where $\nu_K$ is a constant that depends on the kernel, $\hat{\sigma}_+^2 = Var(Y|0 \geq R < h_{pilot})$, $\hat{\sigma}_-^2 = Var(Y|0 > R \geq -h_{pilot})$, $\hat{d}$ is an estimator of the density of $R$ within the pilot bandwidth, and the pilot bandwidth $h_{pilot}$ is given by the Silverman rule of thumb, which is $1.84sd(R)N^{-1/5}$ for the Uniform kernel. From this expression it is clear that since each of the components of $\hat{C}_{IK} = \hat{V}_{IK}/\hat{R}_{IK}$ converge to a positive constant as $h \to 0$, $Nh_{IK}^5 \to C > 0$. By contrast, the variance estimator for the CCT bandwidth, $\hat{V}_{CCT}$, is designed to converge to zero as $h \to 0$: it is the sum of the estimated variances of the intercepts from separate local linear regressions either side of the cutoff, again using a pilot bandwidth (in the CCT case with a Uniform kernel, of $1.84 \min\{sd(R), IQR(R)/1.349\}N^{-1/5}$, which will coincide exactly with the IK pilot bandwidth when $R$ is uniformly distributed). It follows that $Nh_{CCT}^5 \to 0$, as claimed in Calonico et al. (2014b). This difference in limiting behavior translates into CCT bandwidths that are smaller than IK bandwidths, as is apparent in Table B1.

effective sample size, which mechanically reduces the first stage $F$. We have emphasized that the power asymmetry is a problem even when the first stage is quite strong, but it is also true that a weaker first stage makes the power asymmetry problem worse.

Table B2: **Conventional Inference with Undersmoothing, DGPs 1 to 4,** $\rho = 0.8$

| | CCT | | | | | | | IK | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ |
| 1 | 6.4% | 100% | | -.004 | .252 | 15.78 | 254 | 5.4% | 100% | | -.001 | .186 | 28.32 | 434 |
| 2 | 4.2% | 89.5% | | -.005 | .504 | 31.43 | 250 | 4.5% | 81.3% | | -.002 | .378 | 55.44 | 434 |
| 3 | 4.2% | 94.3% | | -.004 | .535 | 26.27 | 226 | 4.4% | 88.2% | | -.001 | .415 | 42.26 | 363 |
| 4 | 5.2% | 96.9% | | .045 | .604 | 20.46 | 179 | 5.2% | 96.9% | | .048 | .586 | 21.69 | 189 |

*Notes:* Summary results from 10,000 artificial datasets of size $N = 2000$ each. CCT and IK runs are identical to those in Table B1 except here we divide the bandwidth by 2. The degree of endogeneity, $\rho$, is set to 0.8 throughout. We report the rate of rejecting $H_0$:$\beta = 0$, the fraction of these rejections that occur when $\hat{\beta} > 0$, and the medians of the estimate, estimated standard error, first-stage $F$, and effective observations.

In DGP 4 the true CEF is quadratic, and, as we saw in Table B1, the local linear estimator suffers from substantial positive bias. In Table B2, row 4, we see that undersmoothing reduces this bias substantially (i.e., by about 3/4). But comparing these results to Table 1 in the main text (left panel), we see that undersmoothing is not as effective at removing bias as RBC. Furthermore, Undersmoothing comes at the cost of reducing effective sample size and reducing first-stage $F$. As we saw in Table 1, for RBC in DGP 4, 84% of rejections of $H_0$:$\beta = 0$ occur when $\hat{\beta} > 0$, and here it is 96.9%. So again undesrsmoothing makes the power asymmetry worse.

*Appendix B.2.* ***Local Quadratic Estimation***

The use of local quadratic regression (rather than local linear) is another attempt to reduce bias. A local quadratic regression adds controls for $R^2$ and $R^2 \cdot D$ to both the first-stage and outcome equations. We present results for conventional inference using local quadratic regression combined with either CCT or IK bandwidths in Table B3.

The key message of Table B3 is that local quadratic regression also tends to worsen the power asymmetry problem. We start by comparing the local linear results for DGPs 1 to 3 in Table B1 with the local quadratic results in Table B3. The reason the power asymmetry worsens in these cases is that the median first-stage $F$ statistics are about 40% smaller when we use local quadratic regression. The first-stage $F$ falls despite the fact that bandwidth and effective sample sizes increase

substantially. The reason the partial $F$ on the instrument $D$ falls is the collinearity between $D$ and $R^2$ in the first-stage regression.[25] These results provide an additional reason to avoid higher-order polynomials in Fuzzy RD regression, beyond the reason given by Gelman and Imbens (2019), who point out that more weight is placed on observations further from the cutoff the higher the degree of the estimating polynomial.

Table B3: **Conventional Inference with Local Quadratic Estimator, DGPs 1 to 4,** $\rho = 0.8$

| | CCT | | | | | | | IK | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DGP | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ | Reject | $\% > 0$ | Median: | $\hat{\beta}$ | $SE$ | $F$ | $N$ |
| 1 | 6.2% | 100% | | -.003 | .230 | 18.93 | 680 | 5.5% | 99.8% | | -.004 | .176 | 32.18 | 1131 |
| 2 | 5.1% | 85.8% | | -.005 | .460 | 37.53 | 681 | 4.8% | 74.4% | | -.007 | .352 | 64.27 | 1131 |
| 3 | 4.7% | 90.3% | | .000 | .485 | 30.71 | 623 | 4.7% | 84.3% | | .004 | .409 | 42.03 | 872 |
| 4 | 4.6% | 90.7% | | -.004 | .489 | 30.34 | 618 | 4.1% | 82.7% | | -.005 | .419 | 40.08 | 824 |

*Notes:* Summary results from 10,000 artificial datasets of size $N = 2000$ each. CCT and IK runs are identical to those in Table B1 except here we use local quadratic regression, adding controls for $R^2$ and $R^2 \cdot D$. The degree of endogeneity, $\rho$, is set to 0.8 throughout. We report the rate of rejecting $H_0$:$\beta = 0$, the fraction of these rejections that occur when $\hat{\beta} > 0$, and the medians of the estimate, estimated standard error, first-stage $F$, and effective observations.

The results for DGP 4 in the fourth row of Table B3 show how median bias vanishes when we use local quadratic regression. Of course, this is because the quadratic model is correctly specified. But the power asymmetry is severe, as 90.7% (82.7%) of rejections occur when $\hat{\beta} > 0$ for the CCT (IK) bandwidth. It is interesting to compare this to the RBC results in In Table B2, row 4, left panel, in the main text, where 84% of rejections occur when $\hat{\beta} > 0$.

[25] In the local quadratic case the first stage becomes $X = \pi_0 + \pi_1 D + \pi_2 R + \pi_3 RD + \pi_4 R^2 + \pi_5 R^2 D + e$. The first-stage $F$ can be written as $F = N\pi_1^2/[\sigma_e^2/(\sigma_D^2(1 - R_D^2))]$, where $\sigma_e^2$ is the variance of the first-stage error term, and $\sigma_D^2$ is the variance of the instrument, and $R_D^2$ is the $R$-squared we would obtain if we regressed $D$ on all the other variables on the right-hand side. Clearly, $R_D^2$ is expected to be large in the case of regression-discontinuity designs, as $D$ is a mechanical function of $R$ by construction. Adding terms in $R$ can be expected to raise $R_D^2$ and reduce $F$.

## Appendix C. Why the Triangular Kernel Should be Used for BSD Inference

In Section 4.1, we showed how the BSD bandwidth selection algorithm interacts with the power asymmetry to generate median bias in the BSD estimator towards OLS.[26] We noted this problem is rather severe when using a uniform kernel, but minor when using a triangular kernel. Here we present additional detail on how the bandwidth algorithm works.

Originally, Kolesár and Rothe (2018) proposed choosing bandwidth to minimize CI length, which depends on $se(\hat{\beta}_1)$ via $cv_{1-\alpha}(M) \times se(\hat{\beta}_1)$. This is an option in the RDHonest package, but according to the R code in Kolesár (2024), the default objective is "worst-case MSE," defined as $MSE_{wc} = Bias^2_{max} + se(\hat{\beta}_1)^2$, where $Bias_{max}$ is determined by the assumed upper bound for $f''(R)$. So $se(\hat{\beta}_1)$ enters through the second term. So both these objectives depend on $se(\hat{\beta}_1)$. By minimizing either function that depends on $se(\hat{\beta}_1)$, BSD tends to choose bandwidths that generate estimates close to $\hat{\beta}_{OLS}$. (In fact, in results not reported, we find median bias is worse if one uses the minimize CI length objective instead of the $MSE_{wc}$ criterion.)

Figure C1 plots the $MSE_{wc}$ objective function for artificial data set #1 (out of 10,000) from our Monte Carlo for DGP 4. The bandwidth ranges from 0.01 to max R. The solid and dotted lines plot the $MSE_{wc}$ objective for the uniform and triangular kernels, respectively. Notice that the objective function is very jagged for the uniform kernel, while it is smooth for the triangular. $Bias_{max}$ is a smoothly increasing function of $h$, so what drives the pattern is that $se(\hat{\beta}_1)$ is a jagged function of $h$ if using the uniform kernel, but it becomes a smooth if using a triangular kernel.[27] Thus, when using the uniform kernel, the $MSE_{wc}$ will have many local minima, making it very difficult to optimize. This is a second reason, in addition to the median bias problem, that the BSD method should be used in conjunction with the triangular rather than the uniform kernel.

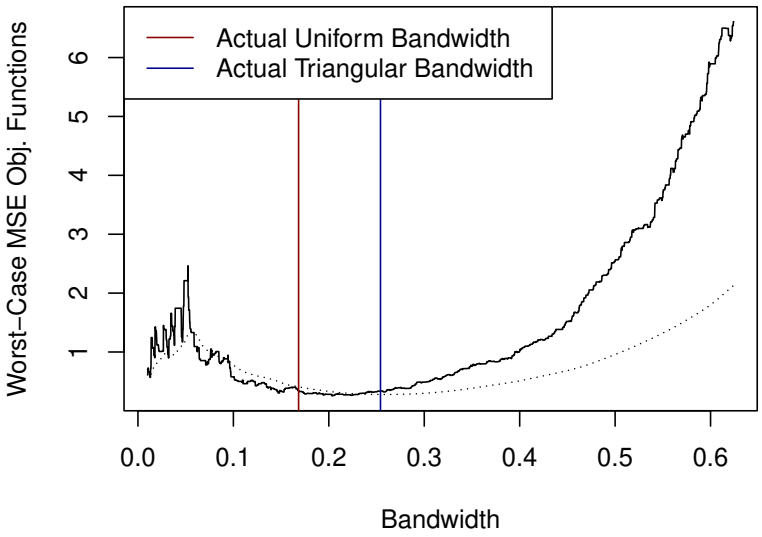
Examining the code for the RDHonest package (Kolesár, 2024), we see the optimization method used is "golden [section] search," which he notes requires a "unimodal" objective function to work well. He also note that it appears unimodal for the Triangular kernel (in line with our findings

[26] This is in addition to the bias due to misspecification of the local linear approximation.

[27] The uniform kernel weights all observations inside the bandwidth equally, and so newly included observations can change the standard error a great deal - hence the large swings in the Uniform kernel objective function in Figure C1. By contrast, the marginal contribution of observations far from the cutoff is close to zero when using the Triangular kernel, so the standard error changes slowly and smoothly.

above). In the case of the uniform kernel, to check if a "brute-force" method of optimization can improve upon the default at the cost of computation time, we perform a simple grid search over the intervals that comprise 99% of the bandwidths chosen by the RDHonest package for our Cases 1 and 4, with a fine grid step of .0005. In neither case do the resulting estimates improve substantially on the originals: Median bias actually *increases* for DGP 1, and is essentially unchanged for DGP 4, while false positive rates and mean bias unambiguously worsen in both cases. These results indicate that the power asymmetry combined with an attempt to minimize (a function of) $se(\hat{\beta}_1)$ inevitably generates median bias in the uniform kernel case, even if we improve the search algorithm.

Figure C1: **Worst-Case MSE Objective Functions: Triangular vs Uniform Kernel**




*Notes:* Based on the first draw of 10,000 with 2000 observations and the parameters set as in Case 4, with $\rho = -0.8$. *Solid black line:* Value of the worst-case MSE objective function evaluated on a grid of 10,000 bandwidths between 0.01 and $\max R$, using the uniform kernel. *Dashed black line:* Same, except using the triangular kernel. Red and blue vertical lines indicate the bandwidths actually selected using the default settings for the RDHonest package and the uniform and triangular kernels, respectively.