Regression to the Mean in Regression Discontinuity Design: Bias and Sensitivity Analysis

Abstract. When making causal inferences from observational data, researchers must consider the effects of confounding. In a regression discontinuity design (RDD), individuals receive a treatment based on whether they score below or above a threshold value measured on a continuous variable. By assuming continuous regression lines for the potential outcomes at the threshold, RDD methods remove the confounding bias in estimating the treatment effect at the threshold. This effect is estimated by the jump in the regression line for the observed outcome at the threshold. Although RDD methods have gained deserved attention in economics, social sciences and epidemiology, we show that inferences from RDDs using the local and global linear regression estimators are prone to regression to the mean bias in certain situations. A common situation is when a running variable has a normal distribution and the cutoff is relatively far from the mean of this distribution. We derive the expression for the limiting bias in this case. In general, the bias occurs when some units receive (or do not receive) treatment when their running variable values are extreme relative to the typical value of the running variable. Through simulations, we show that the regression to the mean bias can lead to inflated type I error rates and bias toward the null in typical settings. Simulations show that the RTM effect can be different for different estimators. We develop a novel method to correct this bias and provide valid inferences. We verify our correction method in simulations and apply it to a real-life example of the incumbency advantage in U.S. House elections.

MSC 2020 classification: 62D20, 92D30

Keywords: Regression to the mean; running variable; sharp discontinuity in treatment assignment; threshold; local regression; local treatment effect.

1 Introduction

A regression discontinuity design is concerned with the effect of a treatment that is determined fully or in part by an observed continuous 'running variable' exceeding a threshold. This paper focuses on sharp regression discontinuity designs where study units are assigned to the treatment exactly when the running variable crosses the threshold. Although it originated in the works of Thistlethwaite and Campbell (1960), the regression discontinuity design (RDD) only now has attained its peak in popularity among empirical researchers driven by new theoretical clarity (Hahn et al., 2001; Lee, 2008), estimation methods (Porter, 2003; Imbens and Kalyanaraman, 2011; Calonico et al., 2014), methods for testing validity of the design using observed data (McCrary, 2008; Smith et al., 2016), and a growing number of applications of the design in diverse fields, e.g., education (Jacob and Lefgren, 2004; Banks and Mazzonna, 2012), housing (Rischard et al., 2021), healthcare (Zuckerman et al., 2006) and policy evaluation (Bakolis et al., 2016). RDD has also been extended to more complex situations, e.g., multiple running variables (Keele and Titiunik, 2015), geographical discontinuity (Keele and Titiunik, 2015) and ordinal running variable (Suk et al., 2022). Additionally, researchers, e.g., Cattaneo et al. (2015), Branson et al. (2019) and Sales and Hansen (2020), have provided new methods for inference using RDD.

Basing on the original framework of Hahn et al. (2001), this paper shows that even when the identification assumptions in an RDD is satisfied, inference from the design can be severely biased by regression to the mean bias.

Regression to the mean (RTM) is among the oldest statistical phenomena which typically refers to the fact that extreme measurements of a random measurement tend to have a value nearer to the expected value in the next measurement. Regression

to the mean has been known to induce bias in many statistical procedures. Our interest is in the investigation of the RTM's effect on RDD. Particularly, if there are extreme values of the running variable that influences whether the corresponding unit is assigned treatment or not then regression to the mean value may affect the unit's outcomes. Imagine, the running variable and the outcome are positively correlated. If the threshold is a high number compared to typical values of the running variable, units crossing the threshold will tend to see their outcomes regressed downwards. We investigate how this phenomena affects the RDD estimates for commonly used estimators.

In earlier literature, Trochim (Trochim, 1984, Chapter 5) notes, "When groups represent distinct populations, measurement error and regression to the mean can operate separately within the groups. Thus, even in the absence of the program, within-group attenuation of slope will lead to pseudo-effects." Although, Trochim provides no elaboration or illustration of the bias.

But recent writings in epidemiology seem to suggest either that regression to the mean does not affect estimation in an RDD or it works in favor of the design. van Leeuwen et al. (2016) write "A high baseline measurement will on expectation regress down to a lower value and a low baseline measurement will on expectation regress up to a higher value. However, as this will occur equally on both sides of the [threshold], the measurement error in the end will be irrelevant for the correct estimation of the treatment effect." While Vandenbroucke and Le Cessie (2014) write "The beauty of the regression discontinuity design, however, is that it exploits the regression-to-the-mean phenomenon by estimating the regression-to-the-mean line from a group of persons on one side of the intervention threshold and then by extrapolating the line at the other side of the threshold to represents the 'expected outcomes'."

In this paper we first provide a brief overview of the basic structure and analysis

of an RDD. Next, we show the data structure that leads to the RTM bias. Then, using simulation experiments, we illustrate the impact of the regression to the mean bias in an RDD. We complement these illustrations by demonstrating the bias also in synthetic data created from data from the Current Population Survey. We then propose a method for correcting the bias. We test this method in simulated data sets and then in a real data exercise based on the work of Lee (2008). We use the R package "rdd" (Dimmery, 2022) for our estimation and standard error calculations in the local linear regression-based strategies.

2 ROLE OF THE RUNNING VARIABLE IN AN RDD

There are n units sampled from a population. For the unit i we observe, a continuous running variable R_i, outcome Y_i and treatment indicator Z_i. The threshold r_0 determines whether the unit i receives the treatment, i.e., $Z_i = 1$ if and only if $R_i \geq r_0$. In the following, we define the estimand, and provide the identification strategy and estimation methods. A more complete discussion can be found in Hahn et al. (2001); Lee (2008); Lee and Lemieux (2014); Bor et al. (2014).

Let $Y_i(1)$ and $Y_i(0)$ denote the two potential outcomes of unit i depending on if it received or did not receive the treatment. Then, our observed outcome $Y_i = Y_i(1)Z_i + Y_i(0)(1 - Z_i)$. The treatment effect of interest is the expected difference in the potential outcomes at the threshold, i.e., $\tau = E(Y_i(1) - Y_i(0) \mid R_i = r_0)$. Unlike the commonly used average treatment effect, this effect is local only for $R_i = r_0$.

Let $\mu_1(r) = E(Y_i(1) \mid R_i = r)$ and $\mu_0(r) = E(Y_i(0) \mid R_i = r)$. If both these functions are continuous at the point $r = r_0$ then

$$\tau = \lim_{s \downarrow 0} \mu_1(r_0 + s) - \lim_{s \downarrow 0} \mu_0(r_0 - s).$$

The two parts of the above equation can be estimated using the observed data by regressions that estimate the functions μ_1 and μ_0 near the threshold. If the functional forms of the μ's can be speculated up to a small number of parameters, then we can estimate the μ's using data on each side of the threshold. However, lacking the knowledge of the functional form, the literature suggest a better strategy for estimating μ_0 near the threshold as fitting a linear model of Y_i on R_i using points with $r_0 - h \leq R_i < r_0$. Thus, the linearity is used as a local approximation of the unknown regression function $\mu_0(r)$ near the threshold. Similarly, we can estimate μ_1 near the threshold by a linear regression model inside $r_0 < R_i \leq r_0 + h$ which approximates the unknown regression function $\mu_1(r)$ near the threshold.

Operationally, this involves fitting the following local linear regression model restricted to the data points i's with $r_0 - h \leq R_i \leq r_0 + h$:

$$Y_i \sim \alpha_0 + \alpha_1 Z_i + \beta_0(R_i - r_0) + \beta_1 Z_i(R_i - r_0). \tag{1}$$

An estimate $\hat{\beta}_1$ in (1) gives an estimate of τ. When the probability of observing R_i near r_0 is positive, as the sample size increases and we appropriately decrease the bandwidth h, $\hat{\beta}_1$ will converge to τ (Hahn et al., 2001). When $h = \infty$, we have a global regression model for the RD. If the functional forms of the regression functions are non-linear, a global regression fit is likely to The standard error of the estimator is calculated as the typical Huber-White robust standard error of this linear model (Lee and Lemieux, 2014).

We now remark on the role of the distribution of R_i in RDD analysis. First, the estimand does not change if the distribution of R_i were different. Second, the exact distribution of R_i is ancillary to the above estimation method. The identification of the parameter only requires that R_i has a positive probability in a neighborhood of

r_0. Third, the distribution of R_i is relevant for determining a good choice of h when the functional forms of the μ's are unknown. If there are only a few points near $R_i = r_0$ then a wider bandwidth might be needed to get a good approximation of the functions in terms of lower mean squared error (if μ_j is non-linear near the threshold, taking a wider bandwidth will increase the bias, while taking a narrower bandwidth will increase variance). Still, with a hypothetical or actual large sample, the density of R_i is also irrelevant in the above estimation method so long as its probability near r_0 is positive.

Yet, contrary to the above, we demonstrate in the following sections that certain distributions of the running variable can bias the estimator even in a large sample and even when the identification assumptions are satisfied. Specifically, this can happen when some units receive the treatment because of larger (or smaller) than a typical value of R_i and/or do not receive the treatment because of smaller (or larger) than a typical value of R_i. This bias is thus because of regression to the mean. The RTM artifact may affect specific estimators differently, while the identification of the local treatment effect under the continuity assumption is unaffected. We detail the limiting bias for the global linear regression estimator in the following section. Next, we demonstrate the bias in simulations where it affects both the Type-I error and power adversely. Then, we show the bias in a synthetic example where a drastic change in the inference is seen by changing the distribution of the running variable and nothing else.

3 JUSTIFICATION OF THE RTM BIAS IN SHARP RDD

We present a technical discussion of the source of RTM bias in sharp RDD. Consider the population model where R_i and $Y_i(0)$ are jointly normally distribution with means

$(r_0+\Delta, 0)$ and variances σ_r^2, σ_y^2 and correlation ρ. In other words, $R_i \sim N(r_0, \sigma_r^2)$ and

$$Y_i(0) = \rho\frac{\sigma_y}{\sigma_r}(R_i - r_0 - \Delta) + \sqrt{1-\rho^2}\sigma_y\epsilon_i,$$

where ϵ_i are standard normal random variables and independent of the running variable R_i. Also, let $Y_i(1) = Y_i(0) (= Y_i)$. The treatment effect at r_0 is $\tau = E(Y_i(1) - Y_i(0) \mid R_i = r_0) = E(Y_i(0) - Y_i(0) \mid R_i = r_0) = 0$, i.e., a null effect. The estimand is $\lim_{s\downarrow 0} E(Y_i \mid R_i = r_0 + s) - \lim_{s\downarrow 0} E(Y_i \mid R_i = r_0 - s)$.

The calculations below show the effect of RTM on the bias of the global linear regression estimator when the threshold r_0 is extreme relative to the distribution of R_i. Thus, notice the role of Δ – the distance of the mean of the running variable from the threshold – relative to σ_r – the standard deviation of the running variable. When $\Delta > 0$ and σ_r is smaller relative to Δ, the threshold point is far from the distribution of the running variable. Thus, the regression to the mean bias will become pronounced when Δ/σ_r is large and ρ is large positive or negative. The explicit expression of the limiting bias is derived below.

For a concrete discussion, fix $\rho > 0$ and $\Delta > 0$. Consider the global regression model to estimate the treatment effect where we fit two regressions of Y_i on R_i for the data to the right of r_0 and to the left of r_0, respectively. Estimate τ by the difference of the fitted values of these regressions at r_0. By standard results of least squares regression, this estimator consistently estimates

$$\{E(Y_i \mid R_i > r_0) + \frac{\sigma_{ry|r_0+}}{\sigma^2_{r|r_0+}}(r_0 - E(R_i \mid R_i > r_0))\} - \{E(Y_i \mid R_i < r_0) + \frac{\sigma_{ry|r_0-}}{\sigma_{ry|r_0-}}(r_0 - E(R_i \mid R_i < r_0))\}, \tag{2}$$

where the first two terms are for the regression to the right of r_0 and the second two terms are for the regression to the left of r_0; $\sigma_{ry|r_0+} = cov(R_i, Y_i \mid R_i > r_0)$ and

$\sigma^2_{r|r_0+} = var(R_i, R_i \mid R_i > r_0)$ and $\sigma_{ry|r_0-}$ and $\sigma^2_{r|r_0-}$ are defined similarly with the conditioning event changed to $R_i < r_0$.

Focus for the moment on the regression fit to the right of r_0. Some calculations show, $E(Y_i \mid R_i > r_0) + \frac{\sigma_{ry|r_0+}}{\sigma^2_{r|r_0+}}(r_0 - E(R_i \mid R_i > 0))$ is

$$-\rho\sigma_y \frac{\phi(-\Delta/\sigma_r)}{1-\Phi(-\Delta/\sigma_r)} + \frac{\sigma_{ry|r_0+}}{\sigma^2_{r|r_0+}} \left\{ -\Delta + \sigma_r \frac{\phi(-\Delta/\sigma_r)}{1-\Phi(-\Delta/\sigma_r)} \right\},$$

where ϕ and Φ are the density and distribution functions of a standard normal random variable, respectively. Consequently, the limiting bias of this estimator for estimating $\lim_{s\downarrow 0} E(Y_i \mid R_i = r_0 + s) = E(Y_i(1) \mid R_i = r_0)$ is

$$\rho\frac{\sigma_y}{\sigma_r}\Delta - \rho\sigma_y \frac{\phi(-\Delta/\sigma_r)}{1-\Phi(-\Delta/\sigma_r)} + \frac{\sigma_{ry|r_0+}}{\sigma^2_{r|r_0+}} \left\{ -\Delta + \sigma_r \frac{\phi(-\Delta/\sigma_r)}{1-\Phi(-\Delta/\sigma_r)} \right\}$$

Rearranging,

$$\sigma_r \left\{ -\Delta/\sigma_r + \frac{\phi(-\Delta/\sigma_r)}{1-\Phi(-\Delta/\sigma_r)} \right\} \left\{ \frac{\sigma_{ry|r_0+}}{\sigma^2_{r|r_0+}} - \rho\frac{\sigma_y}{\sigma_r} \right\}.$$

The term $\frac{\sigma_{ry|r_0+}}{\sigma^2_{r|r_0+}} - \rho\frac{\sigma_y}{\sigma_r}$, inside the second parenthesis, is the difference of the limits of two regression slopes: (i) the first is for the regression of Y_i on R_i using the data $R_i > r_0$, and (ii) the second is for the regression fit of Y_i on R_i using all the data. For $\Delta > 0$ and $\rho > 0$, the regression to the mean effect manifests as a less steep slope for the regression (i) compared to (ii); thus, $\frac{\sigma_{ry|r_0+}}{\sigma^2_{r|r_0+}} - \rho\frac{\sigma_y}{\sigma_r} < 0$. This is because, as $r_0 < E(R_i)$ and $\rho > 0$, the outcomes Y_i tend to be regressed upwards near r_0. For moderately large to large Δ/σ_r, the term $-\Delta/\sigma_r + \frac{\phi(-\Delta/\sigma_r)}{1-\Phi(-\Delta/\sigma_r)} < 0$; resulting in a positive limiting bias.

Similarly, the limiting bias in estimating $\lim_{s\downarrow 0} E(Y_i \mid R_i = r_0 - s) = E(Y_i(0) \mid$

$R_i = r_0$) by the last two terms of (2) is

$$-\sigma_r \left\{ \Delta/\sigma_r + \frac{\phi(-\Delta/\sigma_r)}{\Phi(-\Delta/\sigma_r)} \right\} \left\{ \frac{\sigma_{ry|r_0-}}{\sigma^2_{r|r_0-}} - \rho \frac{\sigma_y}{\sigma_r} \right\}.$$

This time, because of RTM, $\frac{\sigma_{ry|r_0-}}{\sigma^2_{r|r_0-}} - \rho\frac{\sigma_y}{\sigma_r} > 0$, which results in a negative limiting bias for estimating $\lim_{s\downarrow 0} E(Y_i \mid R_i = r_0 - s)$.

Subtracting the two limiting biases, we get an overall positive limiting bias for the global regression based estimator for estimating τ. Notably, the above calculations show the terms that add to the regression to the mean bias, σ_r and Δ/σ_r. When both are large, the RTM bias is large.

In the above, the running variable R_i is a single normal distribution. The data structure above is seen in many RDD studies. For example, Seaver and Quarton (1976) study the Dean's list effect using grade point average as the running variable. Clearly, the threshold for getting into the Dean's list is extreme relative to an average student's GPA. In a second example, Chen et al. (2018) use daily max air quality index as a running variable to study the effect of the air quality alert system in Toronto, Canada on health outcomes. Alert is rung for larger than typical values of the air quality index; see their Figure S1.

In our simulations below the running variable's distribution is a mixture of two normals. The liming bias for the global regression based estimator under this model is obtained using similar calculations as above and the result shows the same effect of the RTM bias. The bias in that case depends on a number of parameters, e.g., the means of the mixtures, their standard deviations and correlation.

We focus primarily on RDD methods using linear regression based estimators as they are popular in practice. It is worth noting that the bias is relative to the estimator in use. Different estimators may have different biases; this is also shown

in our simulations for global and local linear regression estimators and for different bandwidth selection methods. Note that other methods have been proposed for RDD based inference, such as Cattaneo et al. (2015), Branson et al. (2019) and Sales and Hansen (2020). It will be interesting to investigate the possibility of the regression to the mean bias in these methods. We leave this for future research.

4 ILLUSTRATIONS OF THE BIAS

4.1 TYPE-I ERROR RATE

Consider the model

$$Y_i(0) = 0.5R_i + v_i$$

and $Y_i(1) = Y_i(0)$. Therefore, the local treatment effect $\tau = 0$. We generate the running variable using $R_i = r_i^\star + u_i$, where the noise parts (u_i, v_i) are jointly drawn i.i.d. from a bi-variate normal distribution with means $(0, 0)$, variances 0.1 each and correlation ρ. Finally, the $r_i^\star$'s are drawn i.i.d. from the density shown in Figure 1, which is an equal mixture of two normal densities with the same variance 0.5^2 but centered at -1 and 0.5, respectively. Let the threshold $r_0 = 0$.

The parameter ρ is the correlation between the noise terms in the running variable and the potential outcome. Along with the sample size n, we vary the correlation parameter ρ in our simulations. As the previous section suggests, a larger correlation leads to a larger RTM bias in the global regression. Notice that, in the mixture normal distribution, the mean of the right normal component is closer to the threshold compared to that of the left normal component. This feature is not necessary for the RTM effect, neither is a mixture distribution. The RTM effect occurs when some units receive (or do not receive) treatment when their running variable values are

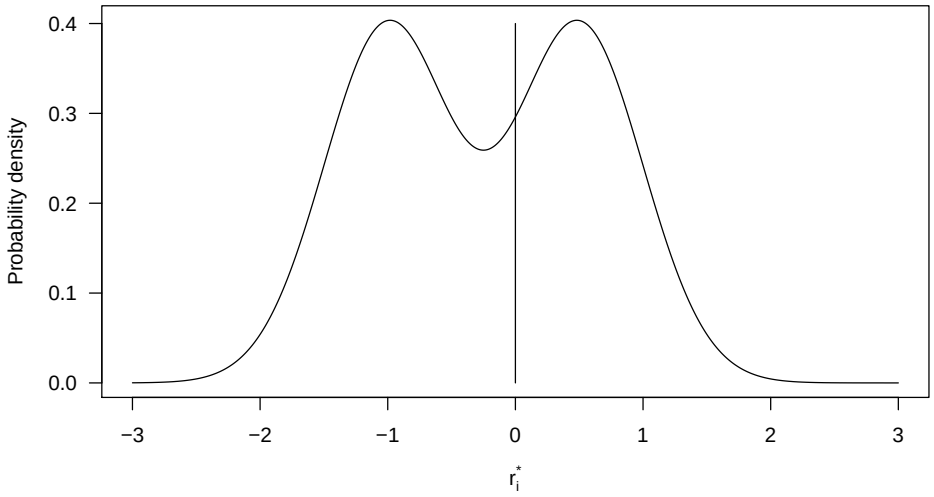

Figure 1: Probability density function of the running variable without noise in the simulation example. The vertical line shows the threshold at 0.

extreme relative to the typical value of the running variable.

Table 1: Type-I error rates at 5% significance level for testing against $H_1 : \tau < 0$ using a global linear regression model

	Sample size		
ρ	500	1000	1500
0.2	0.14	0.23	0.28
0.4	0.35	0.56	0.71
0.6	0.62	0.86	0.96
0.8	0.81	0.98	0.99

We calculated Table 1 of type-I error rates for the linear model (1) on all the data, i.e., $h = \infty$ for a global linear regression based estimator. This model is correctly specified for the above data generating process since Y_i and R_i are linearly related. Still, the type-I error rates for testing $H_0 : \tau = 0$ against $H_1 : \tau < 0$ are all beyond the nominal level 0.05 (5%).

Table 2: Type-I error rates at 5% significance level for testing against $H_1 : \tau < 0$ using local linear regression models

Imbens-Kalyanaraman Bandwidth

	Sample size		
ρ	500	1000	1500
0.2	0.10	0.15	0.18
0.4	0.20	0.31	0.38
0.6	0.33	0.47	0.52
0.8	0.48	0.57	0.60

Cross-Validation Bandwidth

	Sample size		
ρ	500	1000	1500
0.2	0.10	0.15	0.18
0.4	0.22	0.31	0.36
0.6	0.32	0.45	0.47
0.8	0.43	0.49	0.49

As we showed in the previous section, the bias observed in the table occurs because of a regression to the mean. Some of the points with R_i larger than the threshold are too extreme relative to the 'left' normal component in Figure 1. Since R_i has a positive correlation with Y_i, the outcomes for these points tend to be regressed downward. Similar things happen for the points left to the threshold, making Y_i values regress upward for points just below the threshold. As the RDD estimates the treatment effect by the difference in the outcomes just above and below the threshold, we observe a spurious negative effect of the treatment. In Table 1, this bias gets larger with larger values of ρ. Additionally, Table 2 shows that the type-I errors using local linear regression models with two different popular methods for selecting bandwidth are also inflated; see Imbens and Kalyanaraman (2011) for the Imbens-Kalyanaraman optimal bandwidth selection method and Imbens and Lemieux (2008) for the cross-validation based bandwidth selection method.

We explore the bias by further plotting the point estimate and standard errors for the cross-validated bandwidth selection method. Figure 2 shows larger negative biases with larger ρ and also slighty larger standard errors for larger ρ. The point estimate stabilizes with larger sample size. On the other hand, the standard error decreases with larger sample size. Consequently, a larger sample and a larger correlation both increase the type-I error.

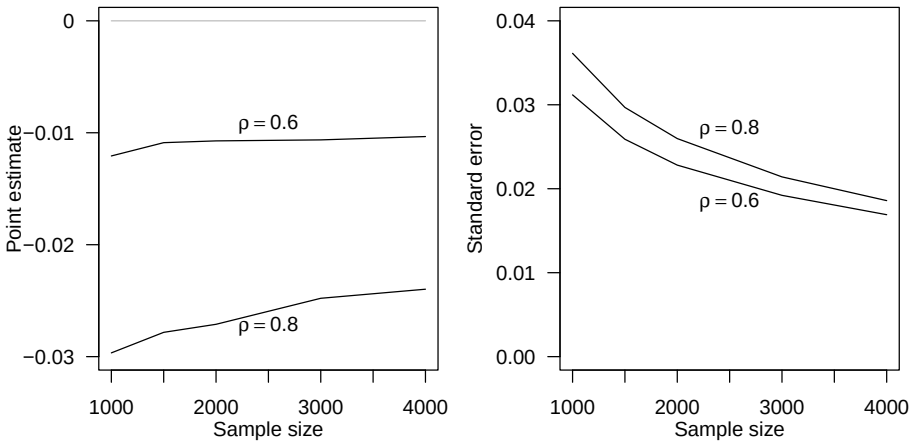

Figure 2: Point estimate and standard error for the cross-validated bandwidth selection. True $\tau = 0$. The point estimates are more negatively biased for $\rho = 0.8$ than $\rho = 0.6$, while the standard errors are slightly larger for $\rho = 0.8$ than $\rho = 0.6$

We next vary the standard deviation of the running variable's 'left' normal component. As seen in Section 3, the effect of the standard deviation is not monotone on the bias. In the notation of Section 3, increasing σ_r decreases Δ/σ_r; however, when both are large, the RTM bias is large. Yet, for a large enough standard deviation, the bias is expected to be lowered as points near the threshold are no longer extreme relative to the running variable's distribution. This is seen in Figure 3 where the density plot of the t-statistics for standard deviation 0.5 is shifted to the left, indicating a larger bias, than the density plot of the t-statistics for standard deviation 0.9.

While the true data generating model is linear on both sides of the threshold, one may still choose to over-specify the regression function as a higher-order polynomial. Then, a polynomial model can capture some of the effect of regression to the mean near the threshold. At the same time, Gelman and Imbens (2019) have demonstrated several disadvantages, including noisy estimates and poor coverage, of using higher order polynomials. Their suggestion is to use local linear or quadratic regression. Similar to local linear regression above, a local quadratic polynomial is

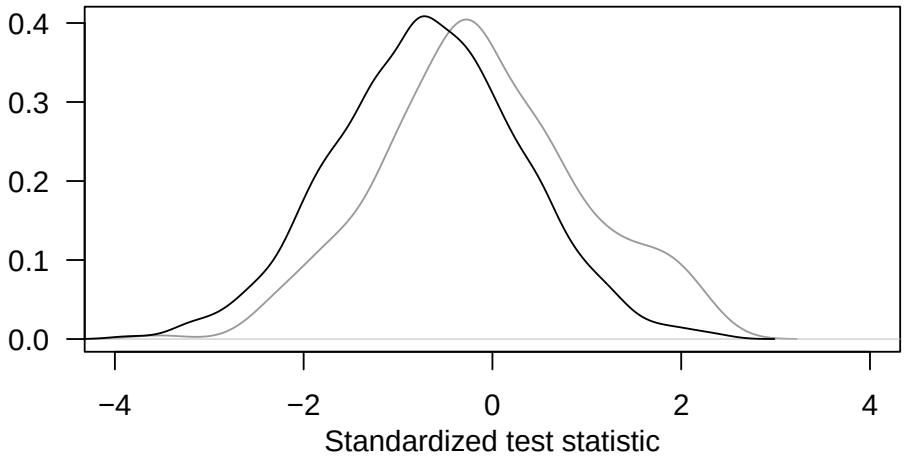

Figure 3: Density of the standardized t-statistic for the estimator with the cross-validated bandwidth for two different standard deviations of the 'left' normal component of mixture normal. The black and grey curves are for standard deviations .5 and .9 respectively.

Table 3: Type-I error rates at 5% significance level for testing against $H_1 : \tau < 0$ using local quadratic regression model and cross-validated bandwidth

	Sample size		
ρ	4000	6000	8000
0.2	0.08	0.08	0.08
0.4	0.11	0.12	0.13
0.6	0.15	0.15	0.17

also influenced by the regression to the mean. Table 3 reports the type-I error rates of a local quadratic model with cross-validated bandwidth. Type-I error is still inflated; although it is smaller compared to the results in the right table of Table 2. In other words, the limiting bias is smaller is magnitude for this method.

Thus, as noted before, the choice of the estimator, e.g., bandwidth selection method and the degree of polynomial, affects the RTM bias. We further investigated the point estimates and standard errors of this local quadratic regression based estimator across different ρ and sample size n. The patterns are the same as for the local linear regression models, shown in Figure 2.

4.2 BIAS TOWARD THE NULL FOR A NON-NULL TREATMENT EFFECT

The previous simulation showed the bias in an RDD analysis that finds a spurious treatment effect. The regression to the mean bias can also result in a bias toward the null when there is a treatment effect. To illustrate this, in the previous simulation model, consider a constant additive treatment effect τ. The observed outcome under this model is $Y_i = 0.5R_i + \tau + u_i$ if $r_i \geq 0$, and $Y_i = 0.5R_i + u_i$ otherwise. When τ is positive and R_i and Y_i have a positive correlation near r_0, the bias will work in the opposite direction to the treatment effect which will lower the rejection rate of the null hypothesis of no treatment effect against the upper sided alternative $H_1 : \tau > 0$. For sample size 1000, Figure 4 provides the empirical rejection rates for this testing problem using a local linear model with optimal bandwidth of Imbens and Kalyanaraman (2011).

Figure 4 shows that as the regression to the mean effect increases with an increase in the ρ since the power decreases for all effect sizes. These results and those from the previous section show that depending on the direction of the treatment effect and the magnitude and direction of the correlation, the RDD method can result in either a bias toward the null or a bias away from the null for regression to the mean bias.

4.3 ILLUSTRATION IN A SYNTHETIC DATA SET

We now illustrate the bias using information on yearly earnings in 1974 and 1975 for 15,992 individuals from the Current Population Survey. Following Gelman and Imbens (2019), we consider the earnings in 1975 in thousands of dollars as the outcome and the earnings in 1974 in thousands of dollars as the running variable.

In this exercise we pretend that the threshold for the yearly earning is 12 (thousands dollars). Earning \$12000 in 1974 does not have any special significance. There

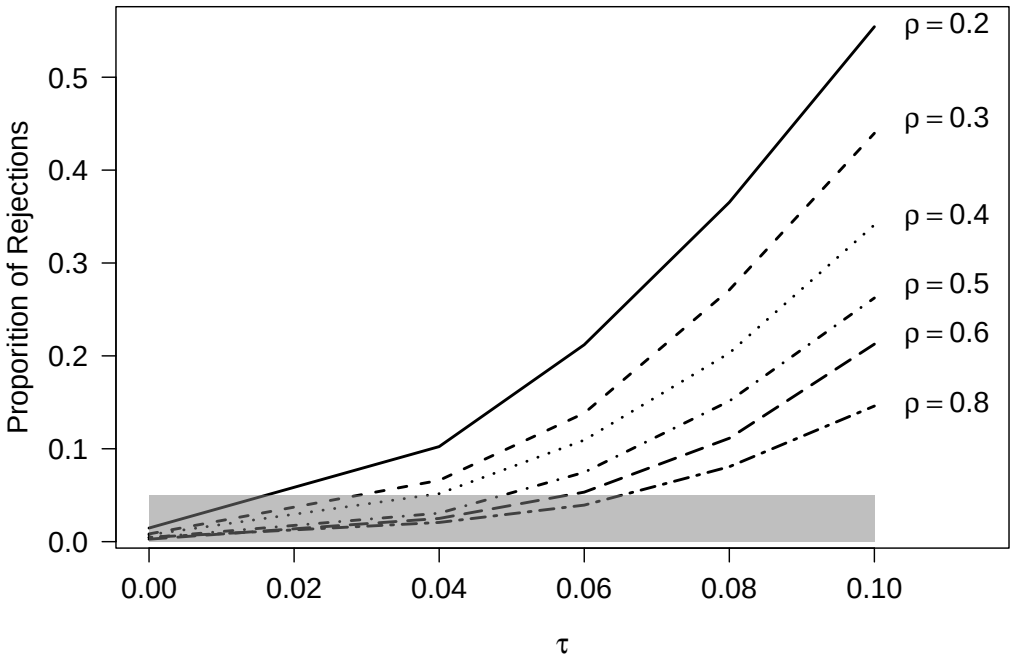

Figure 4: Empirical proportion of rejects of the null hypothesis of $H_0 : \tau = 0$ against $H_1 : \tau > 0$ at 5% significance level for different ρ's. The estimated power is reduced for a large positive correlation.

is no reason to expect a sharp jump in 1975 earnings for people who earned just above this threshold in 1974 from the people who earned just below this threshold. Thus, we would expect that the treatment effect is nearly zero.

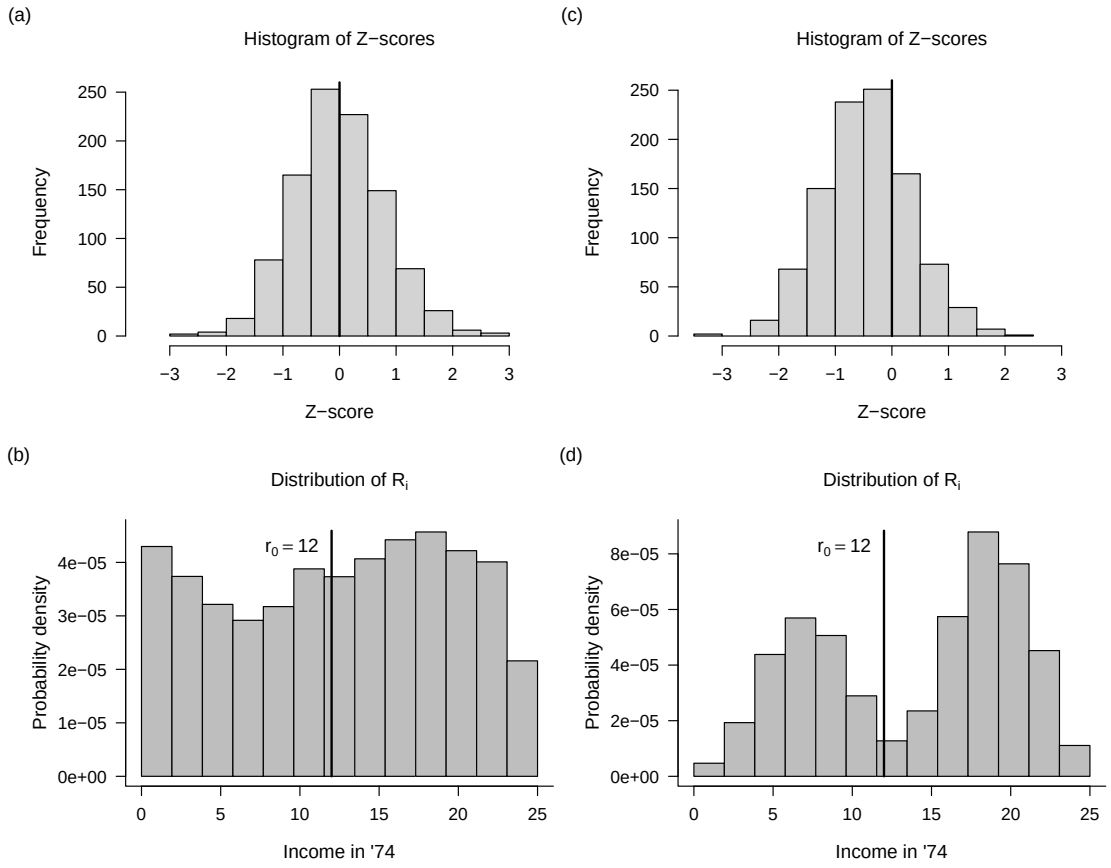

Figure 5: Distribution of the running variable based on income in 1974 (in thousands of dollars) and the histogram of the Z-scores. In (a) and (b) data are simulated from the original data set with replacement using equal weights; in (c) and (d) data are simulated with replacement using unequal weights that vary with the value of the running variable.

We consider two synthetic data sets from this data which differ only in the distributions of the running variable. The first data set is created by drawing a simple

random sample of size 5,000 with replacement from the 15,992 individuals. The second data set, also of size 5,000, is drawn with replacement from the 15,992 individuals using a probability sampling, where individuals with income in 1974 closer to either \$7,000 or \$19,000 get larger sampling weights than other individuals. Figures 5(b) and (d) show the distributions of the running variable in these two synthetic data generating models.

We use the RDD method on these two synthetic data sets with $r_0 = 12$. This process is repeated several times, each time sampling two data sets of sizes 5,000s and applying the RDD method. Figures 5(a) and (c) show the distribution of the Z-scores calculated from the RDD methods for the two data generating models, respectively.

These figures illustrate the regression to the mean bias. The histogram in Figure 5(a) is reasonably close to being symmetric around zero as would be expected under no treatment effect. But, the histogram in Figure 5(c) is skewed to the left because of the high positive correlation between the running variable and the outcome. These results demonstrate the significant effect of the distribution of the running variable on an RDD inference.

5 BIAS CORRECTION AND ANALYSIS USING LEE'S US HOUSE ELECTIONS STUDY

5.1 CORRECTION

To attempt a correction for the bias, we model the distribution of the running variable using the flexible family of finite Gaussian mixture models. We describe our bias correction method using our earlier simulation model for ease of explanation.

In our simulation model, for a reasonable sample size, the finite Gaussian mixture model will estimate two Gaussian components for the running variable (cf. Figure 1).

Call the corresponding proportions p_1 and p_2, centers μ_1 and μ_2, and variances σ_1^2 and σ_2^2, respectively. In our simulation, these two components will be on the opposite sides of the threshold $r_0 = 0$. Assume, $\mu_1 < r_0 < \mu_2$. For a point R_i, using the above values, we can calculate w_{i1}, the probability that it is from the left component.

Before proceeding, recall the following result from probability theory. Consider bi-variate normal random variables W and V with means (μ, ν), variances (σ^2, θ^2) and correlation ρ. Then, for any k

$$E(V \mid W > k) = \nu - \rho\theta \times \frac{\phi((k-\mu)/\sigma)}{1 - \Phi((k-\mu)/\sigma)}, \tag{3}$$

and

$$E(V \mid W < k) = \nu + \rho\theta \times \frac{\phi((k-\mu)/\sigma)}{\Phi((k-\mu)/\sigma)}. \tag{4}$$

Here ϕ and Φ denote the standard normal density and distribution function, respectively.

The regression to the mean effect appears when r_0 is far relative to one of the two or both components. Fix a cutoff value κ so that we say that r_0 is far from the first component if $1 - \Phi((r_0 - \mu_1)/\sigma_1) < \kappa$, and similarly r_0 is far from the second component if $\Phi((r_0 - \mu_2)/\sigma_2) < \kappa$. If r_0 is not far relative to either component, we do not make any correction.

If r_0 is far relative to the first component, we define the corrected value of Y_i with $R_i > r_0$ as

$$\widetilde{Y}_i = Y_i + w_{i1} \times \rho_H \theta_H \frac{\phi((r_0 - \mu_1)/\sigma_1)}{1 - \Phi((r_0 - \mu_1)/\sigma_1)}.$$

This correction uses formula (3) and adjusts the outcomes upward for those points. ρ_H and θ_H correspond to the correlation and variance term for the points with their running variable higher than r_0. In case r_0 is not judged to be far relative to the first

component we do not make any corrections and define $\widetilde{Y}_i = Y_i$.

If r_0 is far relative to the second component, we define the corrected outcome for the points lower than r_0. as

$$\widetilde{Y}_i = Y_i - (1 - w_{i1}) \times \rho_L \theta_L \frac{\phi((r_0 - \mu_1)/\sigma_1)}{\Phi((r_0 - \mu_1)/\sigma_1)}.$$

This time, we use equation (4), and ρ_L and θ_L are the correlation and variance respectively for the points with their running variable lower than r_0. Again, if r_0 is not far from the second component, we set $\widetilde{Y}_i = Y_i$.

The details of our two step method follow. We first set κ as our sensitivity parameter. Our method requires estimates of the unknown quantities ρ_L, ρ_H, θ_L and θ_H. To keep this estimation process separate from the process where we make inference regarding τ we split our data randomly into two parts of sizes one-third and two-thirds of the total sample, respectively. We use the cross-validated bandwidth h (Imbens and Lemieux, 2008) calculated from the (R_i, Y_i) values to estimate these unknown quantities as follows. Use the points in the first data split with their running variables in $[r_0 - h, r_0)$ to calculate ρ_L as the correlation between R_i and Y_i, and θ_L as the variance of the residual from the regression of Y_i on R_i. Calculate ρ_H and θ_H similarly based on the points in $(r_0, r_0 + h]$.

Next, using these estimates, calculate $\widetilde{Y}_i$'s for the second data split as defined above. Finally, we fit model (1) for this corrected set of data points $(R_i, \widetilde{Y}_i)$'s using a cross-validated bandwidth $\tilde{h}$ calculated from the corrected data set. We suggest a different bandwidth for this step because (i) the sample size is smaller for this data set and (ii) the correction can affect the functional relation between the running variable and outcome.

We easily generalize the method to more than two mixture components for the

running variables by modifying the definition of $\widetilde{Y}_i$. Here Y_i for a point to the right of r_0 is corrected for each component to the left of the threshold, and similarly for the points to the left of r_0.

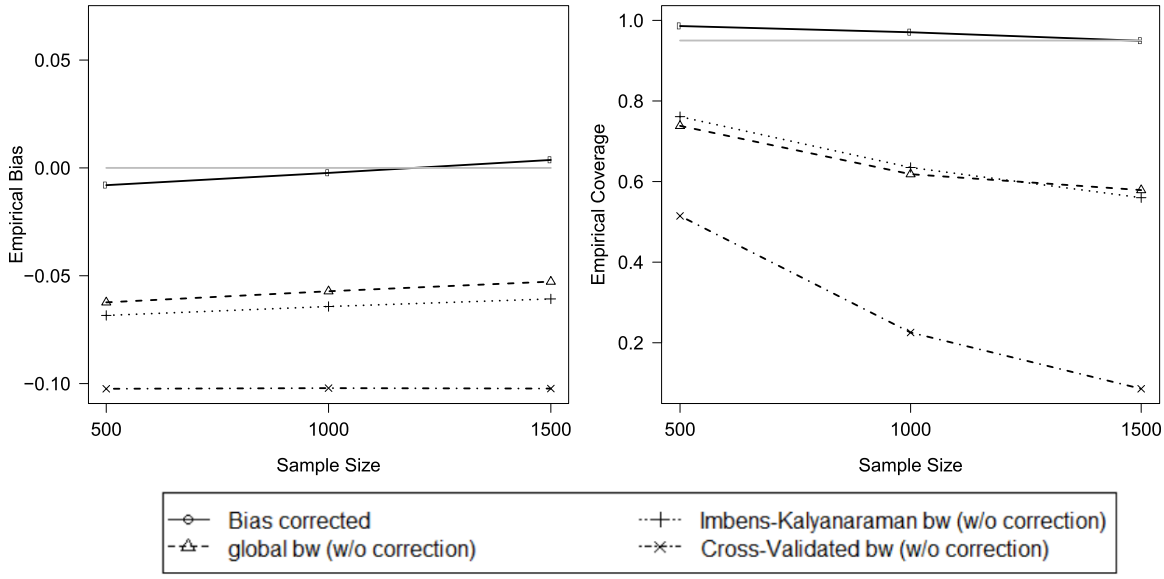

Figure 6: Empirical bias and coverage for inference using the standard local linear regression model. The solid line shows the bias corrected method with cross-validation based bandwidth selection. In these simulation models $\rho = 0.6$.

Figure 6 plots the empirical bias and coverage of this method applied to our simulated data sets. The negative biases of the other estimators are prominent in the left plot, which result in poor coverage of the corresponding confidence intervals in the right plot. We fit the finite Gaussian mixture models using the `Mclust` function from `R` (R Core Team, 2021) package `mclust` with its default settings and use $\kappa = 0.10$. The confidence intervals are from nonparametric bootstrap. Each bootstrap estimation applies the above two step process on a with replacement draw of the same size as the original data set. Figure 6 shows that this method reduces the bias of the estimator and provides valid 95% coverage of the corresponding confidence intervals.

5.2 US HOUSE ELECTIONS STUDY

Lee (2008) used an RDD to estimate the effect of an incumbency advantage in US house elections. Since the majority vote wins, there is a discontinuity in a party being an incumbent in an election based on the difference in the parties' vote shares in the past election. Following Lee Lee (2008), let the running variable be the difference in the vote shares between the Democratic and Republican parties in the last election, with the threshold 0, and let the outcome variable be the democratic vote share in the current election. We use the same data set as Lee (2008) and similarly create the data set from the original 6,558 observations (districts) by discarding 653 observations with past vote share differences greater than 0.99 or less than -0.99.

A Gaussian mixture model estimates two components this running variable with roughly equal proportions 0.505 and 0.495, centers -0.18 and 0.31, and variances 0.0634 and 0.0804, respectively. The probability of a variate from the first component being larger than $r_0 = 0$ is 24%, and the probability of a variate from the second component being smaller than $r_0 = 0$ is 14%. These numbers do not seem to be too small to suggest that there will be a regression to the mean bias. This is also clear from Figure 7 of the density of the running variable.

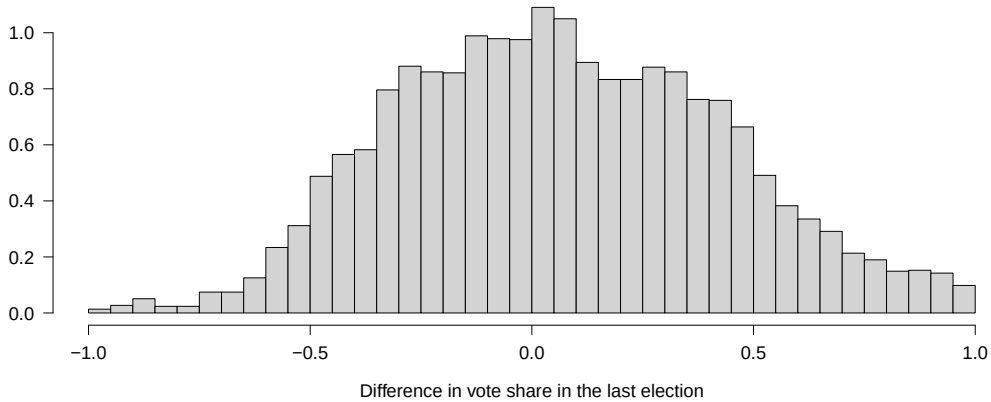

Figure 7: Density for the running variable for the Lee data.

Thus, previously reported analyses of this data set are not likely influenced by regression to the mean bias. We redid the analysis using the local linear model where we selected the bandwidth using cross-validation. The estimate of the effect was 0.0817; see Lee (2008) for further interpretation and discussion.

We created synthetic data sets from this data to test our correction method. We sampled from the 5,905 observations 3,000 observations without replacement where the probability of selecting observation i was proportional to $.5 \times \phi((R_i + 0.18)/\sqrt{0.05}) + .5 \times \phi((R_i + 0.31)/\sqrt{0.05})$. Thus, we use the same centers of the Gaussian mixture as the original data but reduce the variances of the two components. If these two components are known exactly, the probability for a variate from the first component being larger than $r_0 = 0$ is 21%, and the probability for a variate from the second component being smaller than $r_0 = 0$ is 8%. We would not know these components in our analysis.We repeat this process to create 5,000 such synthetic data sets.

Figure 8 shows the boxplots of two sets of estimates of the treatment effect, one without correcting for the bias and the other with a correction for the bias, on these synthetic data sets. Note the longer lower tail in the first boxplot which shows the left skewness of the effect estimates due to the regression to mean bias. In contrast, the estimates calculated after corrections for the bias have a normal boxplot. This boxplot has a larger variance than its counterpart since the inference process leaves out one-third of the observations to estimate the required quantities.

The 95% confidence intervals built using the method that does not correct for the bias have an average length of 0.036 but provide only an empirical coverage of about 90% for the estimate 0.0817 from the original analysis. On the other hand, the 95% confidence intervals built after correcting for the bias have an average length of 0.048, and provide an empirical coverage of about 96%. Although, this does not guarantee

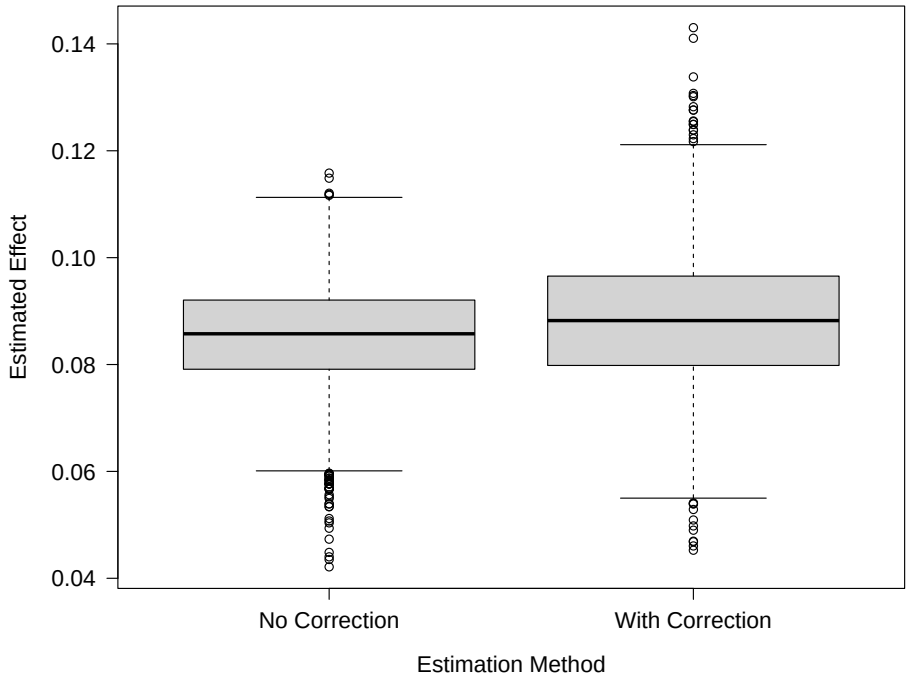

Figure 8: Effect estimates by RDD with (right) and without (left) correction for the regression to the mean bias for synthetic data sets based on the Lee data.

that the proposed method provides the correct coverage of the true incumbency effect at the threshold, which is unknown.

6 DISCUSSION

In this paper we have illustrated the regression to the mean bias in regression discontinuity designs. This bias is different in how it affects the inference from recent works by Daw and Hatfield (2018) and Illenberger et al. (2020) who show the effect of regression to the mean bias on matching estimators in difference-in-differences and synthetic controls analysis, respectively. The bias is induced by the distributional pattern of the running variable where the points near the threshold value are extreme relative to the masses of the running variable on either or both sides of the threshold. Because the existing methods of an RDD do not model the running variable, those methods are susceptible to the regression to the mean bias. We have proposed a method to correct for the bias by first modeling the running variable and then adjusting the outcomes near the threshold when the bias seems possible based on the first model.

The possibility and direction of the bias may be detected from plots of the density of the running variable and of the running variable and outcome. These two plots are commonly produced in an RDD analysis (Lee and Lemieux, 2014; Bor et al., 2014). However, as an RDD analysis focuses on a neighborhood around the threshold, there is also a tendency to create these plots only for such a neighborhood. We suggest plotting the density of the running variable on the complete range of the running variable.

There are several limitations of our proposed method of bias-corrected analysis. First, it is important to set the value of κ, the cutoff on the tail probabilities above or

below the threshold relative to the components for the running variable, appropriately. If this value is set too large, there may be over-adjustment, resulting in a bias in the opposite direction. In our experiments, we have found that $\kappa = .10$ works well since this choice removes the empirical bias and gives appropriate coverage of the confidence interval. Another option is to use synthetic data analyses where we fix the running variable's distribution as in the study and simulate the outcome for different correlations and known effects. Then, these synthetic data analysis results can guide the choice of κ that gives the smallest mean squared error of the bias corrected estimator and a desirable empirical coverage of the corresponding confidence intervals using κ. Further research is needed to determine if a systematic choice of κ is possible. Currently, we suggest treating κ as a sensitivity parameter, varying it around 0.10, and observing its influence on the inference. If the statistical inference is unchanged for κ values in the pre-determined range, e.g., $(.07, .12)$, then the inference is robust to the effect of the regression to the mean. Additional diagnostic checks might be needed in this sensitivity analysis as it assumes a bivariate normal model for the outcome and the running variable in each component of the running variable. We suggest using checks for normality by looking at the qq-plots of the residual from a regression of the outcome on the running variable in each component. If there are significant concerns of non-normality from the diagnostic plots, one should try suitable transformations of the outcome variables.

Second, the bootstrapping process for calculating the confidence interval is computationally expansive relative to standard methods which calculate the confidence intervals analytically. But this process is highly parallelizable. As a whole, we suggest that researchers analyzing an RDD should visually investigate the possibility of the regression to the mean bias in their analysis and, as needed, provide evidence of the robustness of their inference to the regression to the mean, which can be done using

our proposed correction method.

Funding information:

Authors were partly funded by the National Science Foundation.

Conflict of interest:

Authors state no conflict of interest.

Data and Code Availability:

The datasets analyzed during the current study, and the code to reproduce all the results in the paper are available as an online supplement to the paper.

References

Bakolis, I., Kelly, R., Fecht, D., Best, N., Millett, C., Garwood, K., Elliott, P., Hansell, A. L., and Hodgson, S. (2016). Protective effects of smoke-free legislation on birth outcomes in England–A regression discontinuity design. *Epidemiology*, 27(6):810–818.

Banks, J. and Mazzonna, F. (2012). The effect of education on old age cognitive abilities: Evidence from a regression discontinuity design. *The Economic Journal*, 122(560):418–448.

Bor, J., Moscoe, E., Mutevedzi, P., Newell, M.-L., and Bärnighausen, T. (2014). Regression discontinuity designs in epidemiology: Causal inference without randomized trials. *Epidemiology*, 25(5):729–737.

Branson, Z., Rischard, M., Bornn, L., and Miratrix, L. W. (2019). A nonparametric Bayesian methodology for regression discontinuity designs. *Journal of Statistical Planning and Inference*, 202:14–30.

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

Cattaneo, M. D., Frandsen, B. R., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference*, 3(1):1–24.

Chen, H., Li, Q., Kaufman, J. S., Wang, J., Copes, R., Su, Y., and Benmarhnia, T. (2018). Effect of air quality alerts on human health: a regression discontinuity analysis in toronto, canada. *The Lancet Planetary Health*, 2(1):e19–e26.

Daw, J. R. and Hatfield, L. A. (2018). Matching and regression to the mean in difference-in-differences analysis. *Health Services Research*, 53(6):4138–4156.

Dimmery, D. (2022). *rdd: Regression Discontinuity Estimation.* R package version 0.57.

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

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

Illenberger, N. A., Small, D. S., and Shaw, P. A. (2020). Impact of regression to the mean on the synthetic control method : Bias and sensitivity analysis. *Epidemiology*, 31(6):815–822.

Imbens, G. and Kalyanaraman, K. (2011). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959.

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

Jacob, B. A. and Lefgren, L. (2004). Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, 86(1):226–244.

Keele, L. J. and Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, 23(1):127–155.

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

Lee, D. S. and Lemieux, T. (2014). Regression discontinuity designs in social sciences. *Regression Analysis and Causal Inference, H. Best and C. Wolf (eds.), Sage*.

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

Porter, J. (2003). Estimation in the regression discontinuity model. *Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison*.

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rischard, M., Branson, Z., Miratrix, L., and Bornn, L. (2021). Do school districts affect NYC house prices? Identifying border differences using a bayesian nonparametric approach to geographic regression discontinuity designs. *Journal of the American Statistical Association*, 116(534):619–631.

Sales, A. C. and Hansen, B. B. (2020). Limitless regression discontinuity. *Journal of Educational and Behavioral Statistics*, 45(2):143–174.

Seaver, W. B. and Quarton, R. J. (1976). Regression discontinuity analysis of dean's list effects. *Journal of Educational Psychology*, 68(4):459.

Smith, L. M., Lévesque, L. E., Kaufman, J. S., and Strumpf, E. C. (2016). Strategies for evaluating the assumptions of the regression discontinuity design: A case study using a human papillomavirus vaccination programme. *International Journal of Epidemiology*, 46(3):939–949.

Suk, Y., Steiner, P. M., Kim, J.-S., and Kang, H. (2022). Regression discontinuity designs with an ordinal running variable: Evaluating the effects of extended time accommodations for English-language learners. *Journal of Educational and Behavioral Statistics*, 47(4):459–484.

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

Trochim, W. M. (1984). *Research design for program evaluation: The regression-discontinuity approach*, volume 6. SAGE Publications, Incorporated.

van Leeuwen, N., Lingsma, H. F., de Craen, A. J., Nieboer, D., Mooijaart, S. P., Richard, E., and Steyerberg, E. W. (2016). Regression discontinuity design. *Epidemiology*, 27(4):503–511.

Vandenbroucke, J. P. and Le Cessie, S. (2014). Commentary: Regression discontinuity design, let's give it a try to evaluate medical and public health interventions. *Epidemiology*, 25(5):738–741.

Zuckerman, I. H., Lee, E., Wutoh, A. K., Xue, Z., and Stuart, B. (2006). Applica-

tion of regression-discontinuity analysis in pharmaceutical health services research. *Health Services Research*, 41(2):550–563.