DONUT REGRESSION DISCONTINUITY DESIGNS

Claudia Noack   Christoph Rothe

– Preliminary and Incomplete –

1. INTRODUCTION

Regression discontinuity (RD) designs (Hahn et al., 2001; Lee and Lemieux, 2010) allow for simple identification of treatment effects from observational data under transparent conditions. In empirical practice, researchers often carry out various robustness or falsification exercises to support the credibility of these conditions. One such exercise is the so-called "donut" RD design (Barreca et al., 2011), which involves repeating estimation and inference without the data points in some area around the treatment threshold. This approach is often motivated by concerns that possible systematic sorting of units or other data issues in some neighborhood of the treatment threshold might distort estimation and inference of RD treatment effects. The intuition is that if such concerns were unwarranted, then excluding data near the threshold should not change the empirical conclusions in a meaningful way.

While the donut approach is very popular in empirical practice, it is generally carried out in a heuristic fashion without much supporting statistical theory. For example, it is typically unclear how large the difference between conventional and donut RD has to be in order to justify concerns about the validity of the usual RD assumptions, or whether a donut RD approach yields more accurate estimates or confidence intervals than a conventional one. These issues are not trivial: if the usual RD assumptions hold, removing observations near the cutoff is easily seen to increase both the bias and variance of RD estimates through more extensive extrapolation of fitted functions and reduced effective sample size, respectively; and the magnitudes of these effects are not immediately clear. Conventional and donut

This Version: August 29, 2023. Financial support through European Research Council (ERC) grant SH1-77202 is gratefly acknowledged. Claudia Noack, Nuffield College and Department of Economics, University of Oxford, email: claudia.noack@economics.ox.ac.uk. Website: claudianoack.github.io. Christoph Rothe, Department of Economics, University of Mannheim, 68131 Mannheim, Germany. E-Mail: rothe@vwl.uni-mannheim.de. Website: www.christophrothe.net.

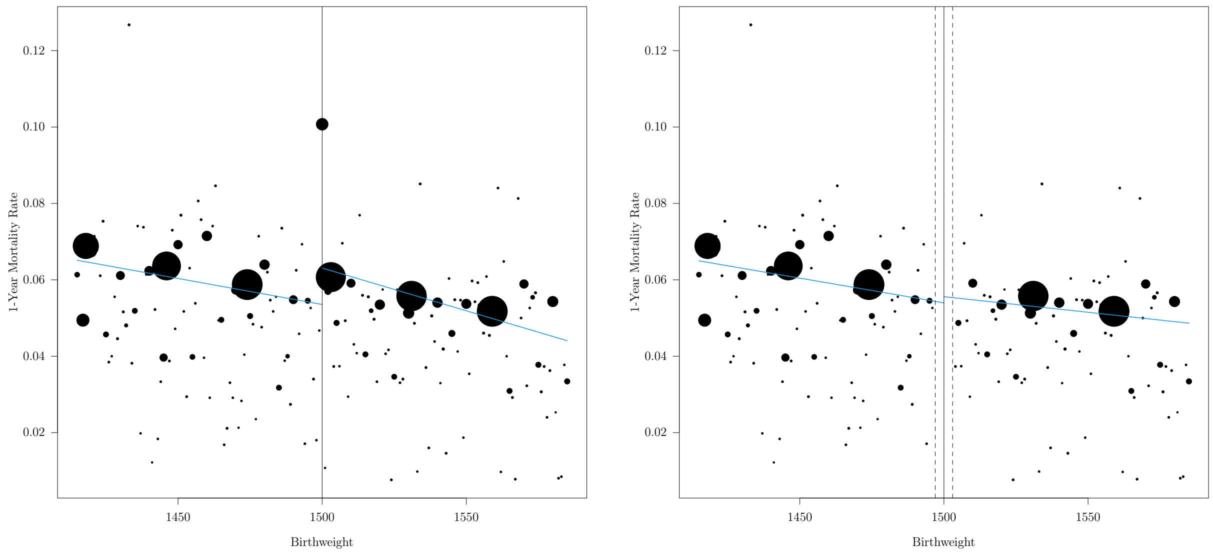

Figure 1: Average 1-year mortality rates of infants with birth weights around 1500g. Size of dots is proportional to number of observations. Left panel shows local linear RD fit with uniform kernel and bandwidth of 85g. Right panel shows fit of same specification when data points with birth weight between 1497g and 1503g are excluded from the data.

RD estimates are also highly correlated because they mostly use the same data, and thus statistically valid comparisons need to be executed carefully.

To give a concrete example, Figure 1 shows average 1-year mortality rates of infants with birth weights around 1500g, a "very low birth weight" threshold that often triggers additional medical interventions. A local linear RD regression, as in Almond et al. (2010), with uniform kernel and bandwidth of 85g implies a both empirically and statistically significant RD estimate of 0.95% with standard error 0.22%. Barreca et al. (2011) argue that this result could be driven by the heaping points at 1500g and 1503g (which correspond to 100g and 1oz multiples, respectively, and are possibly caused by non-random reporting errors). A donut RD that excludes observations within 3g of the 1500g threshold changes the point estimate to 0.16%, and its standard error to 0.28%. While this result appears qualitatively different from the conventional one at face value, there is no formal framework to judge, for example, whether the donut estimate could be driven by extrapolation bias, whether the donut standard error or the implied confidence interval are actually valid, or whether the difference between the conventional and the donut estimate is sufficiently large to be considered unlikely to be caused by random sampling alone.

This paper provides econometric tools to answer such questions. First, we show that donut RD estimates have substantially higher bias and variance than conventional estimates if the usual RD assumption hold. For example, we show that with local linear estimation excluding units that deviate at most 10% of the main bandwidth from the treatment threshold increases the bias by 41% to 63%, and the variance by 53% to 61%, depending on the type of kernel function used. Second, we show that recently developed bias aware confidence intervals (Armstrong and Kolesár, 2018, 2020; Kolesár and Rothe, 2018) remain valid in the context of donut RD designs without special adjustments. Excluding data near the threshold can increase the length of those confidence intervals substantially though: for example, excluding units that deviate from the treatment threshold by at most 10% of the bandwidth value increases the asymptotic length of the confidence interval by 22% to 28%, depending on the kernel. Third, we provide valid statistical tests for the equality of conventional and donut RD estimands in the bias-aware framework that takes the dependence patterns of the estimates into account. We also provide a new version of such a donut specification test that often has better power properties than the approach currently used in practice.

2. FRAMEWORK

2.1. **Conventional RD Design.** Consider a basic sharp regression discontinuity design in which the researcher is interested in the causal effect of a binary treatment. The data are an independent sample $\{(Y_i, X_i, T_i), i = 1, \ldots, n\}$ of size n from some large population. Here $Y_i \in \mathbb{R}$ is the outcome of interest, $X_i \in \mathbb{R}$ is the running variable, and $T_i \in \{0, 1\}$ is an indicator for the event that unit i receives the treatment. Units receive the treatment if and only if the running variable exceeds some known threshold, which we normalize to zero without loss of generality, so that $T_i = \mathbf{1}\{X_i \geq 0\}$. Writing $m_+ = \lim_{x \downarrow 0} m(x)$ and $m_- = \lim_{x \uparrow 0} m(x)$ for the right and left limit, respectively, of a generic continuous function m at zero, we also denote the jump in the conditional expectation of the observed outcome given the running variable at zero by

$$\tau = \mu_+ - \mu_-, \quad \mu(x) = \mathbb{E}(Y|X = x).$$

The core assumptions in a conventional RD analysis imply that the only systematic difference between units on either side of the treatment threshold is their treatment assignment. In this case, τ corresponds to the average treatment effect among units at the threshold. Specifically, if units have potential outcomes $Y_i(1)$ and $Y_i(0)$ with and without receiving the treatment, respectively, so that $Y_i = Y_i(T_i)$, and the conditional expectation of these potential outcomes given the running variable is smooth around the cutoff, then

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

2.2. **Conventional and Donut RD Estimator.** Due to its attractive theoretical properties and easy implementation, local linear regression (Fan and Gijbels, 1996) has arguably become the most commonly used estimation strategy in RD designs. Specifically, the local linear estimator of τ is given by

$$\widehat{\tau}(h) = e_1^\top \operatorname*{argmin}_\beta \sum_{i=1}^n K_h(X_i)(Y - (T_i, X_i, T_iX_i, 1)^\top \beta)^2$$

Here K is a kernel function with compact support, say $[-1, 1]$, $h > 0$ is a bandwidth, $K_h(x) = K(x/h)/h$, and $e_1 = (1, 0, 0, 0)^\top$ is the first unit vector, whose role in the above formula is simply to extract the appropriate coefficient from the right-hand side. Local linear regression thus proceeds by fitting linear specifications with different intercept and slope on either side of the threshold by weighted least squares, giving non-zero weights to units with running variable values $X_i \in [-h, h]$ only.

In a "donut RD" analysis, observations that are immediately surrounding the treatment threshold are excluded from the data. Practitioners generally motivate such an approach with concerns that due to possible systematic sorting of units in some neighborhood of the treatment threshold the parameter τ might not correspond to a meaningful causal effect; see below for details. With $d \in [0, h)$ a constant chosen by the researcher, the local linear donut RD estimator is

$$\widehat{\tau}(h, d) = e_1^\top \operatorname*{argmin}_\beta \sum_{i=1}^n K(X_i/h)(Y - (T_i, X_i, T_iX_i, 1)^\top \beta)^2 \mathbf{1}\{|X_i| \geq d\}.$$

This corresponds again to fitting linear specifications via weighted least squares on either side of the threshold, but now non-zero weights are only given to units with X_i taking values in the "donut-shaped" set $[-h, -d] \cup [d, h]$. This estimator of courses nests the conventional one as a special case, i.e., $\widehat{\tau}(h, 0) = \widehat{\tau}(h)$.[1] Also note that we will use throughout the paper that by simple least squares algebra we can write local linear RD estimators as weighted

[1] Our focus in this paper is on the special case in which the same bandwidth h and donut size d are used on either side of the threshold, but our analysis can easily accommodate asymmetric settings at the cost of a slightly more involved notation. For simplicity, we also focus on the case of sharp RD designs, but our results extend immediately to fuzzy RD designs as well.

averages of the outcomes,

$$\widehat{\tau}(h,d) = \sum_{i=1}^{n} w_i(h,d) Y_i, \qquad \widehat{\tau}(h) = \sum_{i=1}^{n} w_i(h,0) Y_i \equiv \sum_{i=1}^{n} w_i(h) Y_i,$$

with weights that depend on the data through the realizations $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ of the running variable only.

2.3. **Donut RD Estimand.** To study the properties of the donut RD estimator, we must define a corresponding target parameter that is potentially different from the target parameter of conventional RD. However, donut RD designs are often only heuristically motivated in the empirical literature by concerns about sorting or other data issues near the cutoff. We therefore introduce a simple but general model that is intended to be applicable in wide range of empirical settings.

Specifically, we assume that there is a hypothetical sample $\{(Y_i(1), Y_i(0), X_i^*, X_i, T_i), i = 1, \ldots, n\}$ of size n from a large population, where X_i^* is a "natural" running variable that would be observed in the absence of the possible data issues or the mechanism that induces sorting, such that

$$\mathbb{E}(Y_i(t)|X_i^* = x) \text{ is continuous for } t \in \{0,1\}.$$

All other variables are as described above. The observed running variable X_i is further assumed to be identical to X_i^* for those units whose realization of the latter falls outside the donut hole, and to fall into the donut if X_i^* does so as well. That is,

$$X_i = X_i^* \text{ if } |X_i^*| \geq d, \text{ and } |X_i| < d \text{ if } |X_i^*| < d,$$

Treatment assignment is based on the observed running variable, so that $T_i = \mathbf{1}\{X_i \geq 0\}$, and the observed outcome is $Y_i = Y_i(T_i)$. The parameter of interest is the average treatment effect among units whose "natural" value of the running variable is at the treatment threshold:

$$\tau^* = \mathbb{E}(Y_i(1) - Y_i(0)|X_i^* = 0) = \mu_+^* - \mu^*-, \quad \mu^*(x) = \mathbb{E}(Y_i|X_i^* = x),$$

which is generally different from $\tau = \mu_+ - \mu-$, the jump in the conditional expectation $\mathbb{E}(Y_i|X_i = x)$ of the observed outcome given the observed running variable at zero.

The setup implies that we can learn the function $\mathbb{E}(Y_i(t)|X_i^* = x), t \in \{0,1\}$, from the distribution of observable quantities for values of x outside the "donut hole":

$$\mu(x) = \begin{cases} \mathbb{E}(Y_i(1)|X_i^* = x) & \text{if } x \geq d, \\ \mathbb{E}(Y_i(0)|X_i^* = x) & \text{if } x \leq -d. \end{cases}$$

To make further progress, we assume that the function μ^* falls into the usual smoothness class of twice continuously differentiable functions 8except for the threshold) with bounded second derivatives, that is generally used to justify local linear regression approaches. That is, we assume that

$$\mu^* \in \mathcal{F}(M) = \{m_1(x)\mathbf{1}\{x \geq 0\} + m_0(x)\mathbf{1}\{x < 0\}, \|m_t''(\cdot)\|_\infty \leq M, t \in \{0,1\}\}, \tag{2.1}$$

where $M > 0$ that is a uniform smoothness bound that is assumed to be known by the analyst.

2.4. **Large Donut vs. Small Donut Asymptotics.** To study the large sample properties of donut RD estimation, one will need to take stand on the size of the donut d relative to the bandwidth h.

Assumption 1 (Small Donut). *$d = ch$ for some $c \in [0, 1)$, $h \to 0$ and $nh \to \infty$ as $n \to \infty$.*

Assumption 1 is our main framework, which we call "small donut asymptotics". It uses a theoretical device in which the donut size is modeled as proportional to the bandwidth that tends to zero at an appropriate rate. Small donut asymptotics model to the common empirical practice of removing only few observations close to the cutoff in a donut analysis. As we show below, under Assumption 1 the parameter τ^* is point identified and can be consistently estimated, but the size of the donut can have a substantial impact on the bias and variance properties.

Assumption 2 (Large Donut). *d is fixed, $h \to d$ and $n(d - h) \to \infty$ as $n \to \infty$.*

Assumption 2 is an alternative "large donut" asymptotic framework that treats d as fixed and the bandwidth as approaching d at an appropriate rate. This would be appropriate for settings in which the range of the support of the running variable that is used for estimation is much smaller than the gap created by the donut that needs to be extrapolated. Under such a framework the parameter τ^* is only partially identified, as even in large samples no data accumulates in small neighborhoods around the cutoff. Instead, we can only infer from the shape restriction $\mu^* \in \mathcal{F}(M)$ that

$$\tau^* \in T(M) \equiv \{m_+ - m_- : m \in \mathcal{F}(M) \text{ and } P(|m(X_i) - \mu(X_i)| \cdot \mathbf{1}\{|X_i| \geq d\} = 0) = 1\}.$$

With a continuously distributed running variable, the identified set $T(M)$ is easily seen to be an interval of length $2Md^2$ around the linear extrapolation of μ from the edge of the donut to the cutoff:

$$T(M) = [\tau_{\text{Lin}}(d) \pm Md^2], \qquad \tau_{\text{Lin}}(d) = \mu(d) - \mu(-d) - d(\mu'(d) + \mu'(-d)).$$

In this case consistent estimation of τ^* is clearly not possible, but this does of course not preclude valid inference.

2.5. **Regularity Conditions.** In addition to the two asymptotic frameworks above, we also introduce some further, and largely standard, regularity conditions.

Assumption 3. *The running variable X_i is continuously distributed with continuous density f_X that is bounded and bounded away from zero over an open interval that contains the donut hole.*

Continuity of the running variable often assumed in the RD literature, although it is not necessary for valid estimation and inference based on local linear regression (Armstrong and Kolesár, 2018; Kolesár and Rothe, 2018). We still maintain this assumption throughout the paper as it often simplifies the derivations of explicit asymptotic approximations.

Assumption 4. *(i) For all $x \in \mathcal{X}$ and some $q > 2$ $\mathbb{E}[(Y_i - \mathbb{E}[Y_i|X_i])^q|X_i = x]$ exists and is uniformly bounded; (ii) $\mathbb{V}[Y_i|X_i = x]$ is L-Lipschitz continuous for all $x \in \mathcal{X} \setminus \{0\}$ and uniformly bounded away from zero.*

These conditions are needed in order to apply a central limit theorem in various places.

Assumption 5. *The kernel function K is a bounded and symmetric density function function that is continuous on, and equal to zero outside of, some compact set, say $[-1, 1]$;*

This assumption is satisfied by most standard kernel functions, like the uniform, triangular or Epanechnikov kernel, for example. Kernel functions with unbounded support, like the Gaussian kernel, could be accommodated at the cost of algebra. Note that we use the notation that

$$B_K(c) = \int_c^1 J_K(u,c)K(u)u^2du, \quad S_K(c) = \int_c^1 J_K(u,c)^2K(u)^2du,$$
$$J_K(u,c) = e_1^\top \left(\int_c^1 (1,t)^\top(1,t)K(t)dt\right)^{-1} (1,u)^\top,$$

for any constant $c \in [0,1)$, for the remainder of this paper.

3. POINT ESTIMATION

In this section, we study the properties of $\widehat{\tau}(h,d)$ as a point estimator of τ^*. It is easy to see that under large donut asymptotics and our regularity conditions this estimators is inconsistent and instead converges in probability to the midpoint of the identified set, $\widehat{\tau}(h,d) = \tau_{\text{Lin}}(d) + o_P(1)$. We hence we focus on the more interesting case of small donut asymptotics. Denote the bias and variance of $\widehat{\tau}(h,d)$ conditional on the realizations $\mathcal{X}_n = \{X_1,\dots,X_n\}$ of the running variable by $b(h,d) = \mathbb{E}(\widehat{\tau}(h,d)|\mathcal{X}_n)$ and $s^2(h,d) = \mathbb{V}(\widehat{\tau}(h,d)|\mathcal{X}_n)$, respectively, and note that these terms can be written as

$$b(h,d) = \sum_{i=1}^n w_i(h,d)(\mu(X_i) - \tau^*) \text{ and } s^2(h,d) = \sum_{i=1}^n w_i(h,d)^2\sigma_i^2,$$

with $\sigma_i^2 = \mathbb{V}(Y_i|X_i)$ the conditional variances of outcomes given their corresponding running variable values. The following theorem gives asymptotic approximations to these quantities.

Theorem 1. *Suppose that (i) Assumption 1 holds; and (ii) Assumptions 3–5 hold. Then* $\widehat{\tau}(h,d) = \tau^* + o_P(1)$ *and*

$$b(h,d) = h^2B_K(c)\frac{\mu''_{Y+} - \mu''_{Y-}}{2} + o_P(h^2),$$
$$s^2(h,d) = \frac{1}{nh}S_K(c)\left(\frac{\sigma_+^2}{f_+} + \frac{\sigma_-^2}{f_-}\right) + o_P\left(\frac{1}{nh}\right).$$

The result shows that the size of the donut affects the asymptotic bias and variance of the donut RD estimator through the kernel constants $B_K(c)$ and $S_K(c)$ only. These constants can differ quite substantially from their "no donut" counterparts $B_K(0)$ and $S_K(0)$ even for moderate values of c. To illustrate this, Figure 2 shows the relative changes $B_K(c)/B_K(0)$ and $S_K(c)/S_K(0)$ for $c \in (0,.2)$ for three commonly used kernel functions. For example, with $c = .1$, which corresponds to a donut RD that removes the observations that differ by less than 10% of the chosen bandwidth from the cutoff value, the bias increases by 41% to 63%,

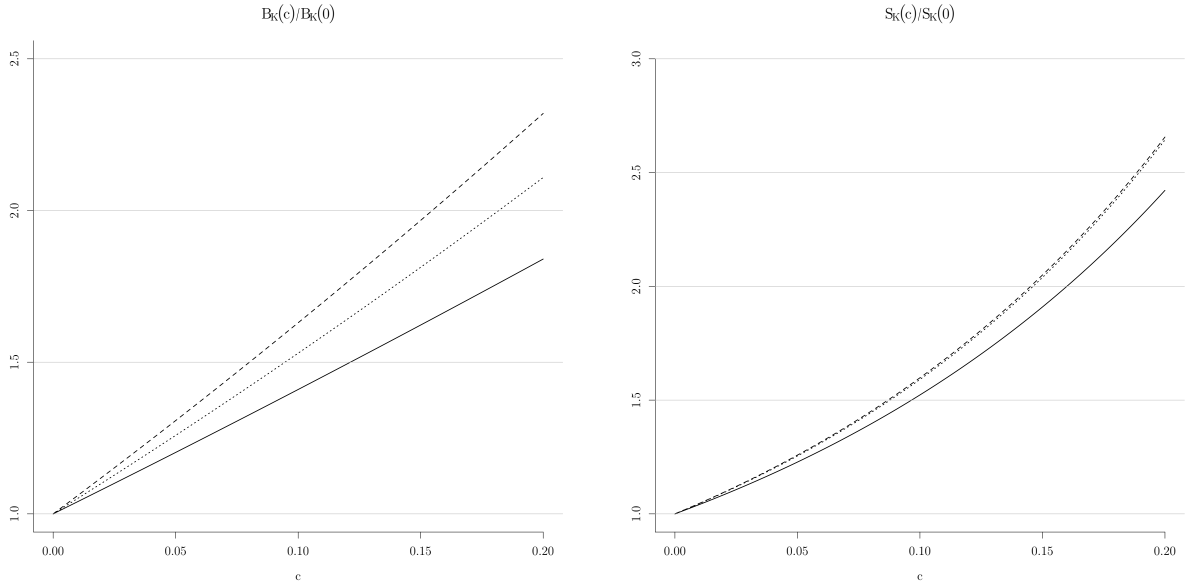

Figure 2: Ratio of "donut" and "no donut" asymptotic bias and variance kernel constants as a function of donut size for uniform (—), triangular (− − −) and Epanechnikov (· · ·) kernels.

and the variance increases by 53% to 61%, depending on the type of kernel function used.[2]

Remark 1 (Worst Case Bias)**.** Note that, following Armstrong and Kolesár (2018), the conditional bias $\bar{b}(h,d)$ can be bounded in finite samples uniformly over $\mathcal{F}(M)$ as

$$\sup_{\mu^*\in\mathcal{F}(M)} |b(h,d)| \equiv \bar{b}(h,d) = -\frac{M}{2}\sum_{i=1}^{n} w_i(h,d)X_i^2\text{sign}(X_i).$$

This "worst case" bias bound can be calculated explicitly from the data, and satisfies

$$\bar{b}(h,d) = -h^2 B_K(c)M + o_P(h^2)$$

under the conditions of the theorem.

4. CONFIDENCE INTERVALS

In this section, we study confidence intervals for τ^* based on the donut RD estimator $\widehat{\tau}(h,d)$. In particular, we argue that recently developed "bias-aware" confidence intervals

[2] To give a point of reference for these numbers, note that reducing the bandwidth of the conventional RD estimator by 10%, which removes observations within two slices of width $.1h$ at the outside rather than the inside of the estimation window, reduces the asymptotic bias by $1-(.9h)^2/h^2 = 19\%$, and increases the variance by only $1-(n.9h)^{-1}/(nh)^{-1} \approx 11\%$.

(Armstrong and Kolesár, 2018, 2020; Kolesár and Rothe, 2018) are valid in donut RD designs under either large or small donut asymptotics without particular adjustments. To explain the construction of these confidence intervals, recall that the "worst case" bias of the donut RD estimator is

$$\bar{b}(h,d) = -\frac{M}{2}\sum_{i=1}^{n} w_i(h,d)X_i^2 \text{sign}(X_i)$$

which can be computed from the data, and note that natural estimates of its conditional variance are of the form

$$\widehat{s}^2(h,d) \equiv \widehat{\mathbb{V}}(\widehat{\tau}_Y(h,d)|\mathcal{X}_n) = \sum_{i=1}^{n} w_i(h,d)^2 \widehat{\sigma}_i^2.$$

Here the $\widehat{\sigma}_i^2$ are suitable estimates of the conditional variances $\sigma_i^2 = \mathbb{V}(Y_i|X_i)$, such as nearest-neighbor estimators (Abadie and Imbens, 2006; Abadie et al., 2014). We can then decompose the usual t-statistic

$$\frac{\widehat{\tau}(h,d)-\tau^*}{\widehat{s}(h,d)} = \frac{\widehat{\tau}(h,d)-\tau^*-b(h,d)}{\widehat{s}(h,d)} + \frac{b(h,d)}{\widehat{s}(h,d)}.$$

as the sum of a term that is approximately standard normal in large samples, and a term that can be bounded in absolute value by $\bar{b}(h,d)/\widehat{s}(h,d)$. This decomposition that motivates the bias-aware confidence interval:

$$C_n(d) = \left[\widehat{\tau}(h,d) \pm \text{cv}_{1-\alpha}\left(\frac{\bar{b}(h,d)}{\widehat{s}(h,d)}\right)\widehat{s}(h,d)\right]$$

with $\text{cv}_{1-\alpha}(r)$ the $1-\alpha$ quantile of $|N(r,1)|$.

Theorem 2. *Suppose that (i) either Assumption 1 or 2 holds; (ii) Assumptions 3–5 hold; and (iii)* $\widehat{s}^2(h,d) = s^2(h,d)(1+o_P(1))$. *Then*

$$\liminf_{n\to\infty} \inf_{\mu^*\in\mathcal{F}_H(M)} P(\tau^* \in C_n(d)) \geq 1-\alpha.$$

The donut confidence interval $C_n(d)$ is generally wider than its conventional counterpart $C_n \equiv C_n(0)$ if $\mu = \mu^*$ because the donut generally increases both the standard error and the bias to standard error ratio. Under small donut asymptotics, if h is chosen to minimize the "worst case" asymptotic MSE (which implies that $\bar{b}(h)/\widehat{s}(h) = 1/2 + o_P(1)$) , the length

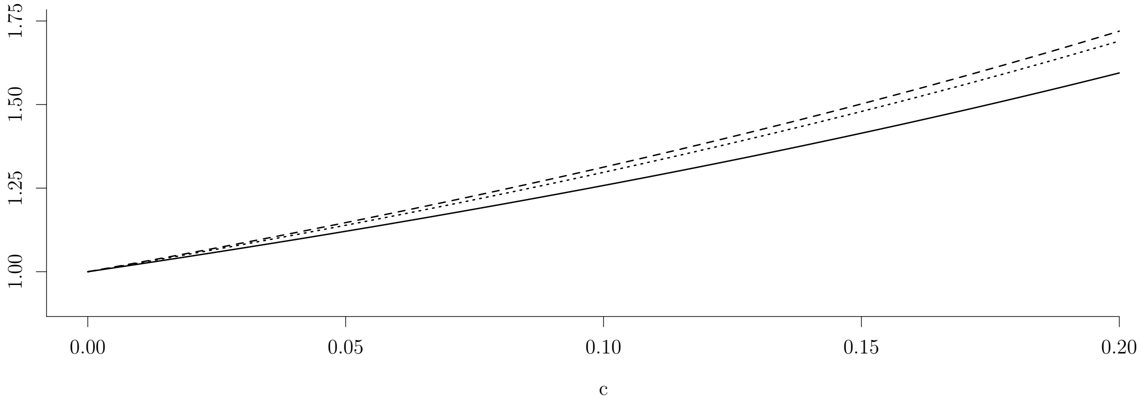

Figure 3: Ratio of asymptotic length of bias-aware “donut” and conventional confidence intervals (cf. equation (4.1)) under small donut asymptotics as a function of donut size for uniform (—), triangular (− − −) and Epanechnikov (· · ·) kernels.

increases by a factor of

$$\frac{\mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}(h,d)}{\widehat{s}(h,d)}\right)\widehat{s}(h,d)}{\mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}(h,0)}{\widehat{s}(h,0)}\right)\widehat{s}(h,0)} = \frac{\mathrm{cv}_{1-\alpha}\left(\frac{1}{2}\frac{B(S)}{\sqrt{S(c)}}\frac{\sqrt{S(0)}}{B(0)}\right)}{\mathrm{cv}_{1-\alpha}\left(\frac{1}{2}\right)} \cdot \sqrt{\frac{S(c)}{S(0)}} + o_P(1). \tag{4.1}$$

We plot the relative increase in asymptotic length, that is, the first term on the right-hand-side of the last equation, for different kernel functions in Figure 3.

Remark 2 (Robust Bias Correction)**.** One can show that the “robust bias correction” confidence intervals of Calonico et al. (2014) also maintain correct coverage under small donut asymptotics, but not under large donut asymptotics. Their length is also substantially affected by the size of the donut, and they generally tend to be longer than bias-aware confidence intervals.

5. SPECIFICATION TESTING

5.1. **Comparing Donut and Conventional RD Estimates.** In this section, we consider a way of determining whether conventional and donut RD estimates are “significantly” different in statistical sense. We consider (an appropriately studentized version of) the difference between the two estimates as a test statistic for the null hypothesis:

$$H_0 : \mu^*(x) = \mu(x) \text{ for all } |x| < d,$$

under our maintained conditions that $\mu^*(x) = \mu(x)$ for all $|x| > d$ and $\mu^* \in \mathcal{F}(M)$.[3] We denote the difference between conventional and donut RD estimates by

$$\widehat{\Delta}(h, d) \equiv \widehat{\tau}(h, d) - \widehat{\tau}(h, 0) = \sum_{i=1}^{n}(w_i(h, d) - w_i(h, 0))Y_i,$$

and the respective conditional bias and variance as $b_\Delta(h, d) = \mathbb{E}(\widehat{\tau}_\Delta(h, d)|\mathcal{X}_n)$ and $s^2_\Delta(h, d) = \mathbb{V}(\widehat{\tau}_\Delta(h, d)|\mathcal{X}_n)$, which can be written as

$$\begin{aligned} b_\Delta(h, d) &= \sum_{i=1}^{n}(w_i(h, d) - w_i(h, 0))\mu(X_i) \text{ and} \\ s^2_\Delta(h, d) &= \sum_{i=1}^{n}\left(w_i(h, d)^2 + w_i(h, 0)^2 - 2w_i(h, d)w_i(h, 0)\right)\sigma^2_i, \end{aligned}$$

respectively. We then prove the new result that under our null hypothesis the conditional bias term can be bounded in finite samples uniformly over $\mathcal{F}(M)$ by

$$\sup_{\mu \in \mathcal{F}(M)} |b_\Delta(h, d)| \equiv \bar{b}_\Delta(h, d) = -\frac{M}{2}\sum_{i=1}^{n}(w_i(h, d) - w_i(h, 0))X_i^2 \text{sign}(X_i).$$

On the other hand, a natural estimate of the conditional variance is given by

$$\widehat{s}^2_\Delta(h, d) = \sum_{i=1}^{n}\left(w_i(h, d)^2 + w_i(h, 0)^2 - 2w_i(h, d)w_i(h, 0)\right)\widehat{\sigma}^2_i,$$

with $\widehat{\sigma}^2_i$ again a nearest-neighbor estimate of $\sigma^2_i = \mathbb{V}(Y_i|X_i)$. A statistical test can then be based on the t-statistic

$$t_\Delta = \frac{\widehat{\Delta}(h, d)}{\widehat{s}_\Delta(h, d)},$$

which is approximately normal in large samples with unit variance and mean bounded in absolute value by $\bar{b}_\Delta(h, d)/\widehat{s}_\Delta(h, d)$ under the null hypothesis. This motivates the following decision rule:

$$\text{Reject } H_0 \text{ if } |t_\Delta| > \text{cv}_{1-\alpha}\left(\frac{\bar{b}_\Delta(h, d)}{\widehat{s}_\Delta(h, d)}\right)$$

The following theorem shows that the resulting test has correct size.

[3] Note that it does not suffice to state the null hypothesis as $H_0 : \tau = \tau^*$, as some control of μ within the donut is necessary to derive the properties of the conventional RD estimator.

Theorem 3. *Suppose that (i) either Assumption 1 or 2 holds; (ii) Assumptions 3–5 hold; and (iii) $\widehat{s}^2_\Delta(h,d) = s^2_\Delta(h,d)(1+o_P(1))$. Then, under H_0:*

$$\liminf_{n\to\infty} \inf_{\mu^*\in\mathcal{F}_H(M)} P\left(|t_\Delta| \geq \mathrm{cv}_{1-\alpha}\left(\frac{\bar{b}_\Delta(h,d)}{\widehat{s}_\Delta(h,d)}\right)\right) \leq \alpha.$$

The formal power properties of this test depend on the asymptotic framework and the type of (local) alternative under consideration.

Remark 3 (Variance Structure)**.** Note that under small donut asymptotics the variance that appears in the definition of the t-statistic t_Δ satisfies:

$$\widehat{s}^2_\Delta(h,d) = \frac{1}{nh}(S_K(c) + S_K(0) - 2\widetilde{S}_K(c))\left(\frac{\sigma^2_+}{f_+} + \frac{\sigma^2_-}{f_-}\right) + o_P\left(\frac{1}{nh}\right)$$

where $S_K(c)$ is as defined above and

$$\widetilde{S}_K(c) = \int_c^1 J_K(u,c)J_K(u,0)K(u)^2du.$$

The term $\widetilde{S}_K(c)$ captures the dependence between conventional and donut RD estimates.

5.2. **Comparing Donut and "Within Donut" RD Estimates.** The test described in the previous subsection is based on the comparison of two potentially highly correlated quantities. An obvious alternative for testing our H_0, which effectively state evaluating the function μ inside and outside the donut is consistent with the same treatment effect, is to compare the donut RD estimate to a conventional RD estimator with bandwidth d. We refer to this latter estimator $\widehat{\tau}(d,0) = \widehat{\tau}(d)$ as the "within donut" RD estimator, as it only uses data within the donut hole. That is, one can base a test of H_0 on (an appropriately studentized version of) the difference

$$\widehat{\Gamma}(h,d) \equiv \widehat{\tau}(h,d) - \widehat{\tau}(d,0) = \sum_{i=1}^n (w_i(h,d) - w_i(d,0))Y_i.$$

We denote the respective conditional bias and variance by $b_\Gamma(h,d) = \mathbb{E}(\widehat{\Gamma}(h,d)|\mathcal{X}_n)$ and $s^2_\Gamma(h,d) = \mathbb{V}(\widehat{\Gamma}(h,d)|\mathcal{X}_n)$, which can be written as

$$b_\Gamma(h,d) = \sum_{i=1}^n (w_i(h,d) - w_i(d,0))\mu(X_i) \text{ and}$$

$$s^2_\Gamma(h,d) = \sum_{i=1}^n \left(w_i(h,d)^2 + w_i(d,0)^2\right)\sigma_i^2,$$

We then show that we can bound the bias in finite samples uniformly over $\mathcal{F}(M)$ under the null hypothesis by

$$\sup_{\mu^*\in\mathcal{F}(M)} |b_\Gamma(h,d)| = \bar{b}_\Gamma(h,d) = -\frac{M}{2}\sum_{i=1}^{n}(w_i(h,d)-w_i(d,0))X_i^2\text{sign}(X_i).$$

On the other hand, a natural estimate of the conditional variance is given by

$$\widehat{s}_\Gamma^2(h,d) = \sum_{i=1}^{n}\left(w_i(h,d)^2+w_i(d,0)^2\right)\widehat{\sigma}_i^2,$$

with $\widehat{\sigma}_i^2$ again a nearest neighbor estimate of $\sigma_i^2 = \mathbb{V}(Y_i|X_i)$. A statistical test can then be based on the t-statistic. We then define the t-statistic

$$t_\Gamma = \frac{\widehat{\Gamma}(h,d)}{\widehat{s}_\Gamma(h,d)},$$

which is approximately normal in large samples with unit variance and mean bounded in absolute value by $\bar{b}_\Gamma(h,d)/\widehat{s}_\Gamma(h,d)$ under the null hypothesis. This motivates the following decision rule:

$$\text{Reject } H_0 \text{ if } |t_\Gamma| > \text{cv}_{1-\alpha}\left(\frac{\bar{b}_\Gamma(h,d)}{\widehat{s}_\Gamma(h,d)}\right).$$

The following theorem shows that the resulting test has correct size.

Theorem 4. *Suppose that (i) either Assumption 1 or 2 holds; (ii) Assumptions 3–5 hold; and (iii)* $\widehat{s}_\Gamma^2(h,d) = s_\Gamma^2(h,d)(1+o_P(1))$*. Then, under* H_0*:*

$$\liminf_{n\to\infty}\inf_{\mu^*\in\mathcal{F}_H(M)} P\left(|t_\Gamma| \geq \text{cv}_{1-\alpha}\left(\frac{\bar{b}_\Gamma(h,d)}{\widehat{s}_\Gamma(h,d)}\right)\right) \leq \alpha.$$

The formal power properties of this test depend on the asymptotic framework and the type of (local) alternative under consideration. One can show that this test has better local power properties against certain natural types of alternatives than the conventional test. The practical downside is that a large overall sample size is needed to use this test as we need a sufficient number of data points within the donut in order for the distribution of $\widehat{\tau}(d,0)$ to be well-approximated by a central limit theorem.

6. NUMERICAL RESULTS

6.1. **Simulations.** In this section, we report the results of a simple simulation study. We generate the running variable as $X_i \sim U(-1,1)$ and the outcome as $Y_i = \mu_L(X_i) + \varepsilon_i$, where $\varepsilon_i \sim N(0, 0.5)$ independent of X_i and $\mu_L(x) = \text{sign}(x)x^2 - L\text{sign}(x)((x - 0.1 \times \text{sign}(x))^2 - 0.1^2 \times \text{sign}(x))\mathbf{1}\{|x| < 0.1\}$ for $L \in \{0, 10, \dots, 40\}$. We define $\mu^*(x) = \mu_0(x)$ so that $\tau^* = 0$. The case $L = 0$ then corresponds to a setting where donut RD estimation would not be necessary, whereas for larger values of L we have settings in which the conditional expectation of the observed data differs more and more extremely from its hypothetical counterpart over the area $(-0.1, 0.1)$. We then compute conventional and donut RD estimates, corresponding bias-aware confidence intervals, and our two specification tests with a triangular kernel, $d = 0.1$ and $M = 2$.[4] We use $n = 1,000$ as the sample size and set the number of replications to 10,000. Our results can be summarized as follows.

Point Estimation. Table 1 shows empirical the bias, standard deviation and root mean squared error (RMSE) of the conventional and the donut RD estimator for the various values of L. Note that the data-driven bandwidth selector for the conventional RD estimator chooses an average value of $h = 0.49$. The results in Table 1 can thus be compared with theoretical predictions under small donut asymptotics with $c \approx 0.19$. Focusing first on the case $L = 0$ we see that the ratios of bias and standard deviation (and thus RMSE) between donut and conventional RD estimators are as predicted by theory. The properties of the donut RD estimator are then unaffected by L as expected, whereas the bias, but not the standard deviation, of the conventional RD estimator increases with L. Note however only with the extreme value $L = 40$ the RMSE of the donut estimator becomes smaller than that of the conventional one. This illustrates that extreme deviations from the usual assumptions are necessary for donut RD estimation to become the dominant point estimator.

Confidence Intervals. Table 2 shows the empirical coverage rates and average lengths of the conventional and donut RD confidence intervals for the various values of L. Donut CIs have correct coverage for all values of L as expected, and their average length remains the same across scenarios as well. Conventional CIs have correct coverage for $L = 0$, which then deteriorates for larger values of L. The ratio of average lengths of the two CI types for $L = 0$ is as predicted by our small donut asymptotic theory.

Specification Testing. Table 3 shows the empirical rejection rates of the two specification tests we considered for the various values of L and the usual nominal level $\alpha = 0.05$. Both

[4] Computation are carried out in `R` with the package `RDHonest`, using the latter's default setting for bandwidth selection.

Table 1: Simulation Results: Point Estimation

	Bias		Std. Dev.		RMSE	
L	Regular	Donut	Regular	Donut	Regular	Donut
0	0.043	0.115	0.099	0.162	0.108	0.198
10	-0.026	0.115	0.099	0.161	0.102	0.198
20	-0.094	0.115	0.100	0.162	0.137	0.199
30	-0.161	0.117	0.101	0.162	0.190	0.200
40	-0.229	0.116	0.102	0.162	0.251	0.199

Table 2: Simulation Results: Confidence Intervals

	CI Coverage		CI Length	
L	Regular	Donut	Regular	Donut
0	0.954	0.948	0.430	0.764
10	0.962	0.949	0.431	0.764
20	0.884	0.946	0.430	0.764
30	0.699	0.946	0.430	0.763
40	0.445	0.947	0.430	0.763

Table 3: Simulation Results: Specification Testing

	Rejection Frequency	
L	$\widehat{\Delta}$	$\widehat{\Gamma}$
0	0.053	0.052
10	0.119	0.152
20	0.232	0.345
30	0.396	0.593
40	0.565	0.803

tests are seen to have correct size if $L = 0$ and thus the null hypothesis holds. Both tests also exhibit increasing rejections rates in L as expected. However, the alternative test based on $\widehat{\Gamma}$ exhibits strictly greater power than the conventional one based on $\widehat{\Delta}$ under for the type of conditional expectation functions considered here.

6.2. **Empirical Application.** To be completed.

7. CONCLUSIONS

This preliminary draft of the paper was created as companion for a conference presentation and shows our main results regarding the theoretical properties of donut RD estimation and inference. Further details, proofs and an empirical application will be added soon.

REFERENCES

Abadie, A. and G. W. Imbens (2006): "Large Sample Properties of Matching Estimators for Average Treatment Effects," *Econometrica*, 74, 235–267.

Abadie, A., G. W. Imbens, and F. Zheng (2014): "Inference for misspecified models with fixed regressors," *Journal of the American Statistical Association*, 109, 1601–1614.

Almond, D., J. J. Doyle Jr, A. E. Kowalski, and H. Williams (2010): "Estimating marginal returns to medical care: Evidence from at-risk newborns," *Quarterly journal of economics*, 125, 591–634.

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

——— (2020): "Simple and honest confidence intervals in nonparametric regression," *Quantitative Economics.*

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

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

Fan, J. and I. Gijbels (1996): *Local polynomial modelling and its applications*, Chapman & Hall/CRC.

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

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

Kolesár, M. and C. Rothe (2018): "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review*, 108, 2277—-2304.

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