DONUT REGRESSION DISCONTINUITY DESIGNS

Claudia Noack Christoph Rothe

Abstract

Donut regression discontinuity (RD) designs exclude observations in a neighborhood around the cutoff. They are widely used in empirical work to address concerns about manipulation, sorting, or measurement error in the running variable. Despite their popularity, donut RD estimators are typically implemented heuristically, with limited formal guidance and unclear implications for bias, variance, and statistical inference. This paper provides a theoretical analysis of donut RD designs within the local linear estimation framework. We derive new results for point estimation, inference, and specification testing in donut RD designs. We show that donut trimming affects the bias and variance of the estimator and thus introduces substantial additional uncertainty relative to conventional RD estimators. We further show that recently developed bias-aware confidence intervals accurately account for this additional uncertainty without requiring modifications. Finally, we formalize specification tests based on comparing conventional and donut RD estimators to allow for valid statistical inference. The results are illustrated with two empirical applications.

1. INTRODUCTION

Regression discontinuity (RD) designs provide a framework for identifying causal effects from observational data under clear and interpretable assumptions. In empirical applications in economics and the social sciences, estimation and inference commonly rely on local linear regression, whose statistical properties have been extensively studied (e.g., Hahn et al., 2001; Imbens and Kalyanaraman, 2012; Calonico et al., 2014; Armstrong and Kolesár, 2018). In this paper, we study a widely used variant of the standard RD design that excludes observations near the treatment threshold. This approach is commonly referred to as a "donut" RD

First version: August 27, 2023. This version: January 26, 2026. Claudia Noack, Department of Economics, University of Bonn. E-mail: claudia.noack@uni-bonn.de. Website: https://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.

design (Barreca et al., 2011) for the gap it creates around the threshold in the distribution of the running variable.

Empirical researchers often use donut RD designs when there are concerns that observations closest to the cutoff may be atypical in ways that threaten identification.[1] A common motivation is the possibility of sorting or manipulation of the running variable, or administrative discretion in its measurement or recording, which can generate non-comparable units near the threshold. In other settings, heaping, rounding, or mass points in the running variable may induce irregular behavior near the cutoff that reflects reporting practices rather than the treatment assignment rule itself.

Depending on the context, researchers may view donut RD estimators either as more credible estimates of causal effects or as diagnostic tools. In the latter case, large differences between donut and conventional RD estimates are interpreted as evidence of violations of standard RD assumptions, without necessarily treating the donut estimate as a meaningful estimate of the causal parameter of interest. Conversely, small differences between donut and conventional RD estimates are often presented in robustness exercises as reassuring evidence that such identification concerns are limited.

Despite its popularity in empirical work, the donut RD approach is typically applied with little formal statistical guidance. In particular, it remains unclear how to interpret differences between conventional and donut RD estimates, or how large such differences must be to meaningfully support concerns about sorting, manipulation, or measurement problems at the cutoff. The consequences of trimming observations near the threshold for bias, variance, and confidence interval coverage are likewise poorly understood: excluding data close to the cutoff reduces local information and mechanically increases extrapolation, but the resulting trade-offs depend on trimming and bandwidth choices in ways that are rarely made explicit. Moreover, because conventional and donut estimators rely on largely overlapping samples and are therefore highly correlated, valid comparisons require careful treatment that is largely absent from applied practice.

To give a concrete example, Figure 1 shows average 1-year mortality rates for infants with birth weights around 1,500g, a very low birth weight threshold that often triggers additional medical interventions. A local linear RD regression, as in Almond et al. (2010), using a uniform kernel and a bandwidth of 85g yields a substantial and statistically significant RD estimate of 0.95%, with a standard error of 0.22%. Barreca et al. (2011) argue that

[1] As of January 2026, a Google Scholar search for "donut regression discontinuity design" returns almost 4,000 results.

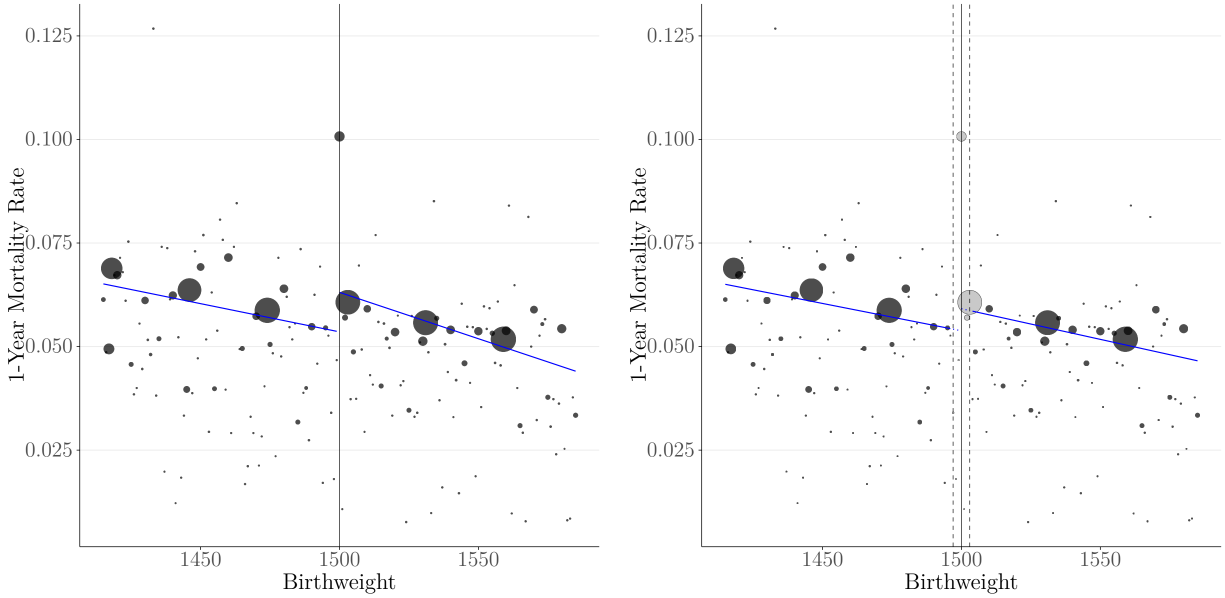

Figure 1: Average 1-year mortality rates of infants with birth weights around 1,500g. Size of dots is proportional to number of observations. Left panel shows local linear RD fit with uniform kernel and bandwidth of 85g. Right panel shows fit of same specification when data points with birth weight between 1,497g and 1,503g are excluded from the data.

these results may be driven by heaping at 1,500g and 1,503g. These values correspond to 100g and one-ounce multiples, respectively, and observations rounded to these points may be systematically different from nearby observations.[2] Indeed, when a donut RD design excludes observations within 3g of the 1,500g threshold, the point estimate falls to 0.16%, while the standard error increases to 0.28%. Although this estimate appears qualitatively different from the conventional one, there is no formal framework for assessing whether the difference is sufficiently large to be attributed to random sampling variation alone. It is also unclear whether the donut RD estimator delivers a reliable estimate of a causal effect, or whether its value is driven by additional variance or extrapolation bias. Finally, it is not known whether the standard errors and confidence intervals reported by standard software packages for donut RD estimators are valid.

This paper provides econometric tools to address these questions. First, we show that donut RD estimators exhibit substantially higher bias and variance than conventional RD estimators when the standard RD assumptions hold. For example, under local linear esti-

[2] For example, hospitals that round birthweights to the nearest 100g multiple, or have scales that measure birthweight in ounces rather than grams, may be systematically different from hospitals that report exact gram values.

mation, excluding observations within 10% of the main bandwidth around the treatment threshold increases bias by 41% to 63% and variance by 53% to 61%, depending on the kernel function. Second, we show that recently developed bias-aware confidence intervals (Armstrong and Kolesár, 2018, 2020; Kolesár and Rothe, 2018) remain valid in donut RD designs without requiring special adjustments. However, trimming observations near the threshold can substantially increase the length of these confidence intervals: excluding observations within 10% of the bandwidth increases asymptotic interval length by 26% to 33%, depending on the kernel. Third, we develop valid statistical tests for equality of conventional and donut RD estimands within a bias-aware framework that explicitly account for the dependence between the two estimators. Finally, we propose a new donut-based specification test with superior local power properties relative to existing approaches, at least against a class of plausible local alternatives.

The remainder of the paper is organized as follows. Section 2 describes the general setup of conventional and donut RD designs. Sections 3 and 4 analyze point estimation and confidence intervals, respectively. Section 5 studies donut RD methods as tools for specification testing. Section 6 discusses extensions of our results to fuzzy RD designs. Section 7 presents two empirical illustrations, and Section 8 concludes.

2. SETUP

2.1. **Conventional RD Designs.** Consider a conventional sharp regression discontinuity (RD) design in which the researcher is interested in the causal effect of a binary treatment.[3] 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. In a sharp RD design, 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 $T_i = \mathbf{1}\{X_i \geq 0\}$. 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 evaluated at zero, we also denote the jump in the conditional expectation of the observed outcome given the running variable at zero by

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

[3] We focus on sharp RD designs in this section to simplify the exposition. See Section 6 below for an extension of our results to fuzzy RD designs.

In a potential outcomes framework, the parameter τ coincides with the average treatment effect of units at the cutoff under certain continuity conditions. Local linear regression (Fan and Gijbels, 1996) has then arguably become the most commonly used estimation strategy to estimate this parameter, due to its attractive theoretical properties and easy implementation. Specifically, the local linear estimator of τ is given by

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

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. We note that by simple least squares algebra we can write the local linear RD estimator as weighted averages of the outcomes,

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

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

2.2. **Donut RD Estimator.** A donut RD estimator is similar to a conventional one, except that it excludes data points immediately surrounding the treatment threshold. Here we focus on the special case of a donut RD design in which the same bandwidth h and donut size $d \in [0, h)$ are used on either side of the threshold.[4] The corresponding 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 $[-h, -d] \cup [d, h]$. The donut estimator nests the conventional one as a special case, i.e., $\widehat{\tau}(h, 0) = \widehat{\tau}(h)$, and it can also be written as weighted averages of the

[4] We could easily accommodate asymmetric settings at the cost of a more involved notation. In particular, the donut could be one-sided, such that only observations above or below the threshold are excluded.

outcomes,

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

with weights that satisfy $w_i(h,0) = w_i(h)$ and depend on the data through the realizations $\mathcal{X}_n = \{X_1, \dots, X_n\}$ of the running variable only.

2.3. **Donut RD Estimand.** If the conditions that allow for identification of causal effects in a conventional RD design were to hold, the donut RD estimator would merely be a very inefficient estimator of the parameter τ, as the exclusion of data points around the cutoff mechanically increases both the bias (through additional extrapolation) and the variance (through the reduced number of data points) of the estimator. Donut RD approaches are therefore precisely motivated by concerns that the observations closest to the cutoff may be atypical in a way that threatens identification.

Depending on the empirical context, the use of donut RD methods typically takes one of two principal forms in empirical practice. Researchers might either use $\widehat{\tau}(h,d)$ as a more credible estimator of a causal effect, and also base inference on this estimator; or they might interpret the magnitude of the difference between $\widehat{\tau}(h,d)$ and $\widehat{\tau}(h)$ as a measure of the severity of the threats to the conventional RD design. Note that the latter perspective does not require one to take $\widehat{\tau}(h,d)$ as evidence about the causal parameter.

Both of these uses of donut RD designs are typically only motivated heuristically, and the formal statistical properties of the corresponding methods are largely unclear. To make progress, we consider a simple but general theoretical framework that is intended to be applicable in a wide range of empirical settings. Specifically, we postulate that there exists a hypothetical function $\mu^*(x)$ whose jump at zero is the target of the donut RD estimator. We denote this jump parameter by

$$\tau^* = \mu^*_+ - \mu^*_-.$$

We assume that the function $\mu^*(x)$ coincides with the conditional expectation function $\mu(x) = \mathbb{E}(Y_i|X_i = x)$ outside the donut, but not necessarily inside it. That is, we assume that $\mu^*(x) = \mu(x)$ for $|x| \geq d$, but that this equality does not (necessarily) hold for $|x| < d$. We also assume that the function $\mu^*(x)$ falls into the usual smoothness class of twice continuously differentiable functions (except at the threshold) with bounded second derivatives, which 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 $M > 0$ is a uniform smoothness bound specified by the analyst. This setup implies that the function $\mu^*(x)$ is identified from the distribution of observable quantities for values of x outside the "donut hole", and that the data are at least to some extent informative about the value of τ^* due to the shape restrictions imposed by the assumption that $\mu^* \in \mathcal{F}(M)$. Depending on the empirical circumstances, the goal of a donut RD design could then be interpreted as estimating τ^*, or as formally testing the null hypothesis that $\mu(x) = \mu^*(x)$ for all x by using a sample analogue $\tau^* - \tau$ as a test statistic.

Example 1 (Causal Setup)**.** Our generic setup maps directly into a model where the validity of a causal design is undermined by the presence of some atypical observations near the cutoff only. Suppose 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 $Y_i(1)$ and $Y_i(0)$ are a unit's potential outcomes with and without receiving the treatment, respectively, so that $Y_i = Y_i(T_i)$; and 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. 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.

2.4. **Asymptotic Frameworks.** For our large sample analysis of econometric methods in donut RD designs, we consider two different asymptotic regimes that present alternative perspectives on the size of the donut d relative to the bandwidth h. We note that practitioners are not meant to explicitly choose between the two regimes, and that most methods we present in this paper work under either regime.

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

We refer to the framework introduced in Assumption 1 as "small donut asymptotics". It models the size of the donut as proportional to the bandwidth, which in turn tends to zero at an appropriate rate. Small donut asymptotics are meant to deliver accurate approximations in commonly found empirical settings that remove observations from an area around the cutoff that is small relative to the bandwidth. Note that while this framework formally allows for $c \in [0, 1)$, the resulting asymptotic approximations only tend to be reliable for values of c closer to zero, say $c \in [0, .2)$.

Assumption 2 (Large Donut)**.** *The donut size 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.

2.5. **Further 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 density function f that is continuous on either side of the cutoff, bounded and bounded away from zero over an open interval that contains the donut hole.*

Continuity of the running variable is 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 analytic expressions for 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 additional algebra in some of the proofs. We also define the following functionals of the kernel:

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

for any constant $c \in [0,1)$.

3. POINT ESTIMATION

In this section, we study the properties of $\widehat{\tau}(h,d)$ as a point estimator of τ^*. Our main result shows that under small donut asymptotics the donut RD estimator stays consistent and asymptotically normal, but that the size of the donut can substantially affect the magnitudes of the asymptotic bias and variance. To state the result, we write the bias and variance of $\widehat{\tau}(h,d)$ conditional on the realizations $\mathcal{X}_n = \{X_1,\ldots,X_n\}$ of the running variable as $b(h,d) = \mathbb{E}(\widehat{\tau}(h,d)|\mathcal{X}_n) - \tau^*$ 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 Assumption 1 and Assumptions 3–5 hold. Then*

$$\frac{\widehat{\tau}(h,d) - b(h,d) - \tau^*}{s(h,d)} \xrightarrow{d} N(0,1), \text{ where}$$

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

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. The proof is straightforward and follows from observing that $\widehat{\tau}(h,d)$ is equal to a conventional RD estimator with a modified kernel function $K_c(u) = K(u)\mathbf{1}\{|u| > c\}$. The important insight is that $B_K(c)$ and $S_K(c)$ can differ quite substantially from their "no donut" counterparts

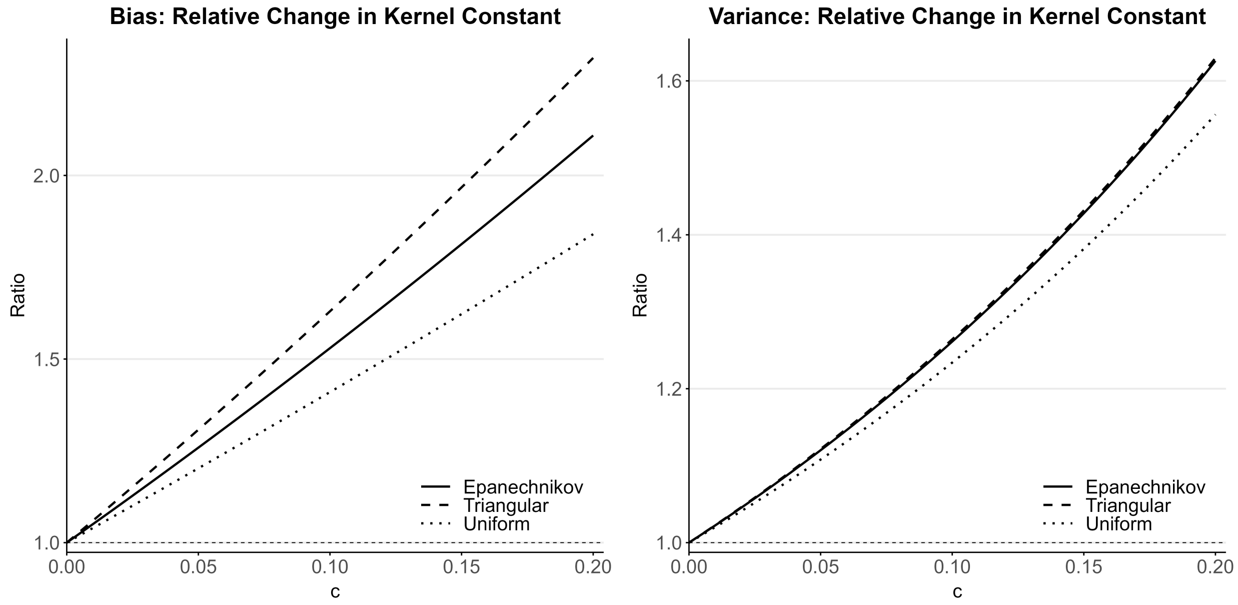

Figure 2: Ratios of kernel constants in the asymptotic bias and variance of the donut and the conventional RD estimator for uniform, triangular, and Epanechnikov kernels.

$B_K(0)$ and $S_K(0)$, even for moderate values of c. To illustrate this, Figure 2 plots the ratios $B_K(c)/B_K(0)$ and $S_K(c)/S_K(0)$ for $c \in (0, .2)$ and three commonly used kernel functions. We can see, for example, that even with moderate value like $c = .1$, 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%, and the variance increases by 53% to 61%, depending on the kernel function used.

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 $.1h$ on 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 $1 - (n.9h)^{-1}/(nh)^{-1} \approx 11\%$.

Remark 1 (Optimal Bandwidth)**.** Given the characterization of the asymptotic bias and variance in Theorem 1, the bandwidth that minimizes the asymptotic mean squared error (MSE) of the donut RD estimator for any particular donut size d is given by:

$$h_{\text{MSE}}(d) = \operatorname*{argmin}_{h>d} \left\{ h^4 B_K(d/h)^2 \left(\frac{\mu''_+ - \mu''_-}{2} \right)^2 + \frac{1}{nh} S_K(d/h) \left(\frac{\sigma^2_+}{f_+} + \frac{\sigma^2_-}{f_-} \right) \right\}.$$

It is generally challenging to characterize how $h^\star_{\text{MSE}}(d)$ changes with d as the bandwidth enters the MSE formula not only as a multiplicative factor preceding the kernel constants, but also within the constants themselves. Nonetheless, with somewhat tedious algebra, one can show that $h_{\text{MSE}}(d)$ is monotonically increasing in d for the three kernel functions considered above. This is not obvious a priori, as both the asymptotic bias and the asymptotic variance are increasing in d, and hence the direction of the change in mean squared error is not clear.

Proposition 1. *Suppose that $d = d_0 \times n^{-1/5}$ for some constant $d_0 > 0$, and that K is either the uniform, triangular, or Epanechnikov kernel. Then $h_{\text{MSE}}(d) = (d_0/c^*(d_0)) \times n^{-1/5}$, where*

$$c^*(d_0) = \operatorname*{argmin}_{c\in(0,1)} \left\{ \left(\frac{d_0}{c}\right)^4 B_K(c)^2 \left(\frac{\mu''_+ - \mu''_-}{2}\right)^2 + \frac{c}{d_0} S_K(c) \left(\frac{\sigma^2_+}{f_+} + \frac{\sigma^2_-}{f_-}\right) \right\}.$$

The term $(d_0/c^*(d_0))$ does not depend on n, is increasing in d_0, and such that

$$\lim_{d_0\to 0} h_{\text{MSE}}(d) = \left(\frac{S_K(0)\left(\frac{\sigma^2_+}{f_+} + \frac{\sigma^2_-}{f_-}\right)}{B_K(0)^2 \left(\mu''_+ - \mu''_-\right)^2} \right)^{\frac{1}{5}} \times n^{-1/5}.$$

Proposition 1 establishes that the MSE-optimal bandwidth $h_{\text{MSE}}(d)$ is increasing in the donut size for the uniform, triangular, and Epanechnikov kernels. This monotonicity result is not automatic for arbitrary kernel functions: it relies on verifying certain high-level conditions on the kernel-induced functions $B_K(c)$ and $S_K(c)$ which can be done analytically for the three kernel functions under consideration, but not for the class of all kernel functions that satisfy Assumption 5. We conjecture, however, that the proposition can also be verified on a case-by-case basis for other commonly used kernel functions.

Remark 2 (Large Donut Asymptotics)**.** The case of large donut asymptotics is less interesting for point estimation as in this framework the parameter τ^* cannot be consistently estimated: even if we were to observe the full population distribution of (Y_i, X_i) conditional on $|X_i| > d$, the shape restriction that $\mu^* \in \mathcal{F}(M)$ only yields that τ^* is partially identified and takes values in some identified set $T(M)$:

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

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

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

The estimator $\widehat{\tau}(h, d)$ simply converges in probability to the midpoint of the identified set, i.e. $\widehat{\tau}(h,d) = \tau_{\text{Lin}}(d) + o_P(1)$, but that midpoint has no particular causal interpretation.

4. CONFIDENCE INTERVALS

In this section, we consider 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 (Armstrong and Kolesár, 2018, 2020; Kolesár and Rothe, 2018) are asymptotically valid in donut RD designs under either large or small donut asymptotics without any special adjustment. The main insight is that the length of these confidence intervals can increase substantially with the size of the donut.

To describe "bias-aware" confidence intervals, note that it follows directly from Armstrong and Kolesár (2018) and Kolesár and Rothe (2018) that the conditional bias of the donut RD estimator is bounded uniformly over $\mathcal{F}(M)$ by a term $\bar{b}(h,d)$ that can be explicitly calculated from the data:

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

We can construct an estimator of the conditional variance of the donut RD estimator as

$$\widehat{s}^2(h,d) = \sum_{i=1}^{n} w_i(h,d)^2 \widehat{\sigma}_i^2,$$

where $\widehat{\sigma}_i^2$ is some suitable estimator of the conditional variance $\sigma_i^2 = \mathbb{V}(Y_i|X_i)$ such as the nearest-neighbor estimator studied by Abadie and Imbens (2006) and Abadie et al. (2014). We can then decompose the usual t-statistic as

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

into the sum of a term that is approximately standard normal in large samples and a term that can be bounded in absolute value by the observable quantity $\bar{b}(h,d)/\widehat{s}(h,d)$. This decomposition then leads to the bias-aware donut confidence interval with nominal confidence level $1-\alpha$,

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

where $\text{cv}_{1-\alpha}(r)$ is the $1-\alpha$ quantile of the $|N(r,1)|$ distribution, i.e., the distribution of the absolute value of a normal random variable with mean r and variance 1. The next theorem shows that this result has correct asymptotic coverage, uniformly over $\mu^* \in \mathcal{F}(M)$.

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*

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

It is instructive to compare the length of the donut confidence interval $C_n(d)$ to that of its conventional counterpart $C_n \equiv C_n(0)$ in the case where μ equals μ^*. Note that the presence of a donut generally increases the standard error, which in turn widens the confidence interval, but it also affects the "worst case bias to standard error" ratio that appears inside the critical value function. To obtain an analytical result, it is convenient to choose the bandwidth such that it minimizes the "worst case" asymptotic MSE of the conventional RD estimator, which means that one chooses the bandwidth

$$\bar{h}_{\text{MSE}} = \left(\frac{S_K(0)\left(\frac{\sigma_+^2}{f_+} + \frac{\sigma_-^2}{f_-}\right)}{B_K(0)^2 M^2} \right)^{\frac{1}{5}} \times n^{-1/5}.$$

This choice implies in particular that $\bar{b}(h,0)/\widehat{s}(h,0) = 1/2 + o_P(1)$. The presence of a donut then changes the length of the confidence interval by a factor of

$$\frac{\text{cv}_{1-\alpha}\left(\frac{\bar{b}(\bar{h}_{\text{MSE}},d)}{\widehat{s}(h,d)}\right)\widehat{s}(\bar{h}_{\text{MSE}},d)}{\text{cv}_{1-\alpha}\left(\frac{\bar{b}(\bar{h}_{\text{MSE}},0)}{\widehat{s}(\bar{h}_{\text{MSE}},0)}\right)\widehat{s}(\bar{h}_{\text{MSE}},0)} = \frac{\text{cv}_{1-\alpha}\left(\frac{1}{2}\frac{B_K(c)}{\sqrt{S_K(c)}}\frac{\sqrt{S_K(0)}}{B_K(0)}\right)}{\text{cv}_{1-\alpha}\left(\frac{1}{2}\right)} \cdot \sqrt{\frac{S_K(c)}{S_K(0)}} + o_P(1).$$

We plot the relative increase in asymptotic length (that is, the first term on the right-hand-side of the last equation) as a function of c for different kernels in Figure 3. 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 length of the confidence interval increases by 26% to 33%, depending on the kernel function. The majority of this change can be attributed to the increase in the standard deviation, but the change in the critical value function adds to the length as well.

Remark 3 (Robust Bias Correction)**.** Confidence intervals based on "robust bias correction" (Calonico et al., 2014) are a popular alternative to "bias-aware" confidence intervals in conventional RD settings. In Appendix A, we show that these confidence intervals only maintain correct coverage under small donut asymptotics, but not under large donut asymptotics. We also show that their length is generally even more affected by the size of the donut than that of bias-aware confidence intervals. This is because robust bias correction involves

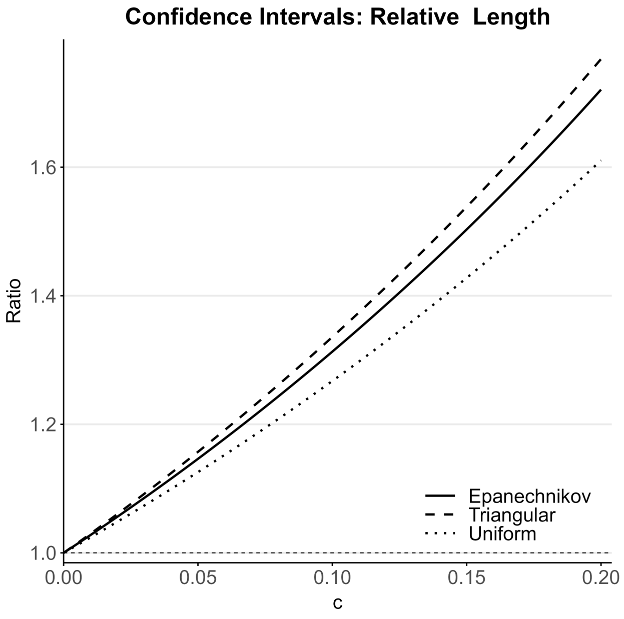

Figure 3: Ratio of asymptotic length of "donut" and "no donut" confidence intervals as a function of donut size for uniform, triangular, and Epanechnikov kernel functions.

higher-order local polynomial estimation, and these higher-order local polynomials turn out to be more sensitive to donut trimming than local linear regression is.

5. SPECIFICATION TESTING

Applied researchers often use donut RD designs as diagnostic tools to investigate the validity of conventional RD designs, taking "large" differences between the donut RD estimator and the conventional RD estimator as evidence of a violation of conditions that ensure identification of causal effects. This approach does not require that the donut RD estimator is by itself a reasonable estimator of a causal effect.

Here we derive two formal hypothesis tests (one based on the usual strategy and one based on a new approach) and compare their power properties over a class of reasonable local alternatives. We formalize the goal of testing for correct specification as 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 $\mu^* \in \mathcal{F}(M)$.[5] This formulation captures the idea that under the null hypothesis the observed data within the

[5] 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.

donut should be compatible with (i) the data outside the donut and (ii) the shape constraints imposed by the assumption that $\mu^* \in \mathcal{F}(M)$.

5.1. **Conventional Strategy: Donut vs. Standard RD Estimates.** We first formalize a testing strategy based on comparing donut RD estimates to conventional RD estimates. We denote the difference between the two estimators 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 by $b_\Delta(h,d) = \mathbb{E}(\widehat{\Delta}(h,d)|\mathcal{X}_n)$ and $s_\Delta^2(h,d) = \mathbb{V}(\widehat{\Delta}(h,d)|\mathcal{X}_n)$, respectively. These can be written as

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

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

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

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

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

with $\widehat{\sigma}_i^2$ again a nearest-neighbor estimate of $\sigma_i^2 = \mathbb{V}(Y_i|X_i)$. Following arguments analogous to those used to construct "bias-aware" confidence intervals, we use the decision rule to

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

is a t-statistic that 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. The following theorem shows that the resulting test has correct size.

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*:*

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

Remark 4 (Variance Structure)**.** The test statistic considered here is based on the difference of two generally highly correlated quantities. In particular, under small donut asymptotics the conditional variance of $\widehat{\Delta}(h,d)$ 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

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

captures the dependence between the conventional and the donut RD estimator.

5.2. **Alternative Strategy: Donut vs. "Within Donut" RD Estimates.** The test statistic in the previous subsection is the difference of two potentially highly correlated quantities. An obvious alternative for testing H_0 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

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

This approach compares two statistically independent quantities, as these are based on non-overlapping subsets of the data. The downside of this construction is that a large overall sample size is required to use this approach as the variance of $\widehat{\tau}(d,0)$ is large and its distribution may not be well-approximated by a central limit theorem without a sufficient number of data points within the donut.

We denote the respective conditional bias and variance by $b_\Gamma(h,d) = \mathbb{E}(\widehat{\Gamma}(h,d)|\mathcal{X}_n)$ and

$s_\Gamma^2(h,d) = \mathbb{V}(\widehat{\Gamma}(h,d)|\mathcal{X}_n)$, which can be written as

$$b_\Gamma(h,d) = \sum_{i=1}^n (w_i(h,d) - w_i(d,0))\mu(X_i) \text{ and}$$
$$s_\Gamma^2(h,d) = \sum_{i=1}^n \left(w_i(h,d)^2 + w_i(d,0)^2\right)\sigma_i^2,$$

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 $\sigma_i^2 = \mathbb{V}(Y_i|X_i)$. Following arguments analogous to those used to construct "bias-aware" confidence intervals, we use the decision rule to

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

is a t-statistic that 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. 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:*

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

5.3. **Power Comparisons.** We compare the power of the two tests in the previous subsections under small donut asymptotics[6] with $h = h_0 \times n^{-1/5}$ for some constant $h_0 > 0$ under sequences of local alternatives in which $\mu^*(x)$ differs from a continuous piecewise quadratic baseline function $\mu(x)$ inside the donut by a piecewise quadratic perturbation that is contin-

[6] The case of small donut asymptotics is the more interesting and illustrative one as under large donut asymptotics the parameter τ^* is only partially identified, and hence no hypothesis test has non-trivial asymptotic power against any alternative inside the identified set.

uous at ±d and discontinuous at the cutoff, and chosen such that

$$\tau_n^* - \tau = \lambda/\sqrt{nh}$$

for some local parameter $\lambda \in \mathbb{R}$. That is, local alternatives are chosen such that the jump at the cutoff, which is the parameter that is targeted by donut RD estimation, scales with the asymptotic standard deviation of the test statistics. The local alternatives considered are specifically of the form

$$H_{1,n} : \mu_n^*(x) = \mu(x) + \eta_n(x) \text{ for all } x \in \mathcal{X},$$

where $\mu(x) = -Mx^2 \text{sign}(x)/2$ is chosen such that the conventional RD estimator achieves its "worst case" bias, and $\eta_n(x) = \text{sign}(x)(\lambda/(2d^2\sqrt{nh}))\,(x - \text{sign}(x)d)^2\,\mathbf{1}\{|x| < d\}$.

Theorem 5. *Suppose that (i) Assumption 1 holds; (ii) Assumptions 3–5 hold; (iii) $\widehat{s}^2_\Delta(h,d) = s^2_\Delta(h,d)(1+o_P(1))$ and $\widehat{s}^2_\Gamma(h,d) = s^2_\Gamma(h,d)(1+o_P(1))$; and (iv) the bandwidth is $h = h_0 \times n^{-1/5}$. Then, under the local alternative $H_{1,n}$,*

$$t_\Delta \xrightarrow{d} N(\mathcal{D}_\Delta(c,h_0,K) - \lambda\,\mathcal{L}_\Delta(c,K),1) \quad \textit{and} \quad t_\Gamma \xrightarrow{d} N(\mathcal{D}_\Gamma(c,h_0,K) - \lambda\,\mathcal{L}_\Gamma(c,K),1)$$

for constants

$$\mathcal{D}_\Delta(c,h_0,K) = \frac{-Mh_0^{5/2}\big(B_K(c) - B_K(0)\big)}{\sigma_\tau\sqrt{S_\Delta(c)}}, \quad \mathcal{D}_\Gamma(c,h_0,K) = \frac{-Mh_0^{5/2}\big(B_K(c) - c^2 B_K(0)\big)}{\sigma_\tau\sqrt{S_\Gamma(c)}},$$
$$\mathcal{L}_\Delta(c,K) = \frac{1}{2}\frac{\int_0^c J_K(u,0)(u-c)^2 K(u)\,du}{c^2\sigma_\tau\sqrt{S_\Delta(c)}}, \quad \mathcal{L}_\Gamma(c,K) = \frac{1}{2}\frac{B_K(0)+1}{\sigma_\tau\sqrt{S_\Gamma(c)}},$$

with $\sigma_\tau^2 = \sigma_+^2/f_+ + \sigma_-^2/f_-$ and $S_\Delta(c) = S_K(c) + S_K(0) - 2\widetilde{S}_K(c)$ and $S_\Gamma(c) = S_K(c) + \frac{1}{c}S_K(0)$.

Note that the constants $\mathcal{D}_\Delta(c,h_0,K)$ and $\mathcal{D}_\Gamma(c,h_0,K)$ capture the drift induced by the curvature of the baseline function $\mu(\cdot)$. They are the probability limits of the respective "worst case bias to standard error" ratios, and would also be present under the null hypothesis (which corresponds to $\lambda = 0$). The constants $\mathcal{L}_\Delta(c,K)$ and $\mathcal{L}_\Gamma(c,K)$, on the other hand, capture the drift induced by the local alternative inside the donut.

Theorem 5 can be used to characterize the asymptotic power functions of the two tests under the local alternative $H_{1,n}$. It is convenient to choose h such that it minimizes "worst case" asymptotic MSE of numerator of the respective t-statistic, as in this case the "worst case bias to standard error" ratios become asymptotically equal to 1/2. With this choice of

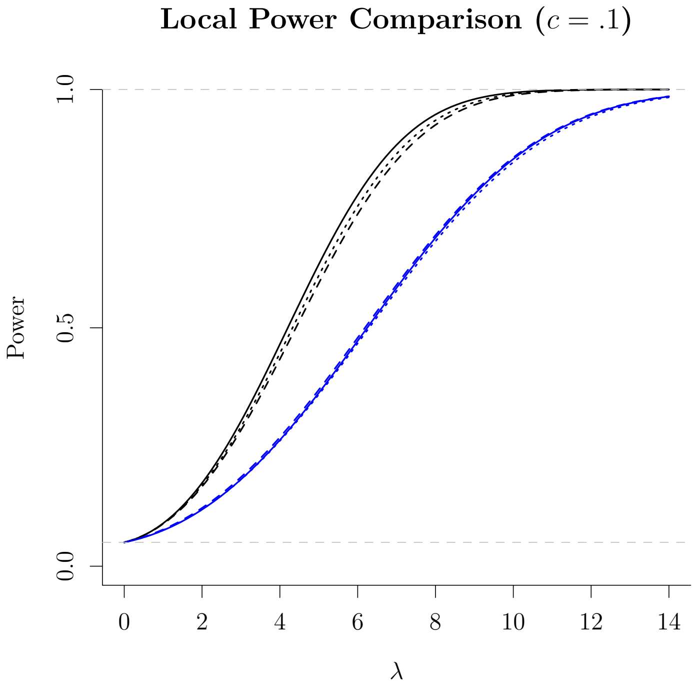

Figure 4: Asymptotic power of the test Δ (blue) and Γ (black) based on local alternatives indexed by λ for uniform, triangular, and Epanechnikov kernels and $\alpha = 0.05$.

bandwidth, we have that

$$\mathbb{P}\left(|t_\Delta| > \text{cv}_{1-\alpha}\left(\frac{\bar{b}_\Delta(h,d)}{s_\Delta(h,d)}\right)\Big| H_{1,n}\right) = \mathbb{P}\left(|Z - \lambda\mathcal{L}_\Delta(c,K)| > \text{cv}_{1-\alpha}\left(\frac{1}{2}\right)\right) + o_p(1),$$
$$\mathbb{P}\left(|t_\Gamma| > \text{cv}_{1-\alpha}\left(\frac{\bar{b}_\Gamma(h,d)}{s_\Gamma(h,d)}\right)\Big| H_{1,n}\right) = \mathbb{P}\left(|Z - \lambda\mathcal{L}_\Gamma(c,K)| > \text{cv}_{1-\alpha}\left(\frac{1}{2}\right)\right) + o_p(1),$$

where $Z \sim N(1/2, 1)$ is a generic normal random variable with mean 1/2 and variance 1. The first terms on the right-hand side of the last two displays are the asymptotic power functions of the two tests we consider. It is tedious to further characterize these functions analytically. We show their numerical values as a function of λ for various kernels and the special case $c = .1$ in Figure 4 (the overall shape of the power curves is qualitatively very similar for other values of c). This plot clearly shows that the test Γ, which is based on our new strategy to compare donut and “within donut” estimates, generally dominates the test Δ, which formalizes the comparison of donut and conventional estimates that is commonly carried out in the empirical literature. This dominance result is of course tied to the specific form of the local alternative.

6. FUZZY REGRESSION DISCONTINUITY DESIGNS

In this section, we briefly outline how our analysis extends to fuzzy regression discontinuity (RD) designs; additional details are provided in Appendix A. In a fuzzy RD design, treatment assignment is determined by whether the running variable exceeds a cutoff, but compliance with the assignment rule is imperfect. As a result, the probability of treatment exhibits a discontinuity at the cutoff, but it might not jump from zero to one.

The parameter of interest is typically defined as the ratio of two sharp RD parameters,

$$\theta = \frac{\tau_Y}{\tau_T}, \qquad \tau_Y = \mu_{Y+} - \mu_{Y-}, \quad \tau_T = \mu_{T+} - \mu_{T-},$$

where, for a generic random variable W_i, we write $\mu_W(x) = \mathbb{E}[W_i \mid X_i = x]$ for its conditional expectation given the running variable.[7] Under standard continuity assumptions (Hahn et al., 2001; Dong, 2018), θ corresponds to the average causal effect at the cutoff for units whose treatment status is affected by the assignment rule (i.e., compliers).

A fuzzy donut RD estimator can be constructed as the ratio of two sharp donut RD estimators,

$$\hat{\theta}(h,d) = \frac{\hat{\tau}_Y(h,d)}{\hat{\tau}_T(h,d)}, \qquad \hat{\tau}_Y(h,d) = \sum_{i=1}^{n} w_i(h,d)Y_i, \quad \hat{\tau}_T(h,d) = \sum_{i=1}^{n} w_i(h,d)T_i.$$

We then postulate that there are functions $\mu_Y^* \in \mathcal{F}(M_Y)$ and $\mu_T^* \in \mathcal{F}(M_T)$ that coincide with μ_Y and μ_T outside the donut, but possibly not inside it. The numerator and denominator of the fuzzy RD estimator are then interpreted as targeting the jumps $\tau_Y^* = \mu_{Y+}^* - \mu_{Y-}^*$ and $\tau_T^* = \mu_{T+}^* - \mu_{T-}^*$, respectively; and the fuzzy donut RD estimator itself is interpreted as targeting their ratio $\theta^* = \tau_Y^*/\tau_T^*$.

If the discontinuity in treatment probabilities is bounded away from zero outside the donut, then under small-donut asymptotics the fuzzy donut RD estimator is asymptotically equivalent to a sharp donut RD estimator with an appropriately defined outcome variable. In particular,

$$\hat{\theta}(h,d) - \theta \approx \sum_{i=1}^{n} w_i(h,d)U_i, \quad U_i = Y_i - \frac{\tau_Y}{\tau_T} - \frac{\tau_Y}{\tau_T^2}\big(T_i - \tau_T\big).$$

Under these conditions, all of our results for point estimation, confidence intervals, and specification testing apply directly to $\hat{\theta}(h,d)$ after replacing the outcome variable Y_i with

[7] Throughout this section, notation parallels that used earlier, with subscripts Y and T denoting outcomes and treatment indicators, respectively.

the auxiliary outcome U_i.

Under large-donut asymptotics, both $\hat{\tau}_Y(h,d)$ and $\hat{\tau}_T(h,d)$ are generally inconsistent for their respective target parameters, and consequently so is $\hat{\theta}(h,d)$. Nevertheless, it remains possible to construct valid confidence sets for the corresponding fuzzy RD estimand θ^* using the Anderson–Rubin–type procedure proposed by Noack and Rothe (2024). This approach is valid under both small- and large-donut asymptotics and does not require the treatment probability discontinuity to be bounded away from zero. For specification testing in fuzzy RD designs, we recommend conducting separate sharp RD specification tests for the numerator and denominator of the fuzzy RD estimator. The null hypothesis of correct specification is then rejected if the smaller of the two resulting p-values falls below the Bonferroni-adjusted significance level $\alpha/2$.

7. TWO EMPIRICAL APPLICATIONS

In this section, we present two empirical applications that illustrate our methods.

7.1. **Teacher contracts.** We first revisit the data from Bau and Das (2020), who study the effect of contract type of teachers, which can be either temporary or permanent, on various outcomes such as teachers' wages and teacher value added (TVA) in Pakistan. They exploit a policy shift in 2002 which led to teachers being hired almost exclusively on non-tenured temporary contracts. This policy shift was immediately preceded, however, by a budget crisis from 1998 to 2001, which was a uniquely long period of low hiring. Bau and Das (2020) conduct a (one-sided) fuzzy donut RD analysis with data from eight years before and after the reform, which excludes the four year period of low hiring.

For illustration, we focus here on teachers' log monthly wages as the outcome. Figure 5 shows the data together with the first stage and reduced form local linear regression fits in its left panel; and the data together with the first stage and reduced form donut local linear regression fits in its right panel. We can see that donut trimming has a particularly substantial effect on the first stage estimate of the jump in treatment probabilities.

As the donut size here is exactly half of the effective bandwidth, we consider this to be a setting where small donut asymptotics are less plausible and large donut asymptotics deliver more reasonable approximations. In Table 1, we report fuzzy donut RD point estimates as well as point estimates for the first stage and the reduced form, as well as p-values for tests that donut trimming is negligible in the first stage and the reduced form.

We can see that donut trimming substantially changes the estimate of the first stage parameter, the jump in treatment probabilities, from 0.14 to 0.94. The Δ specification test

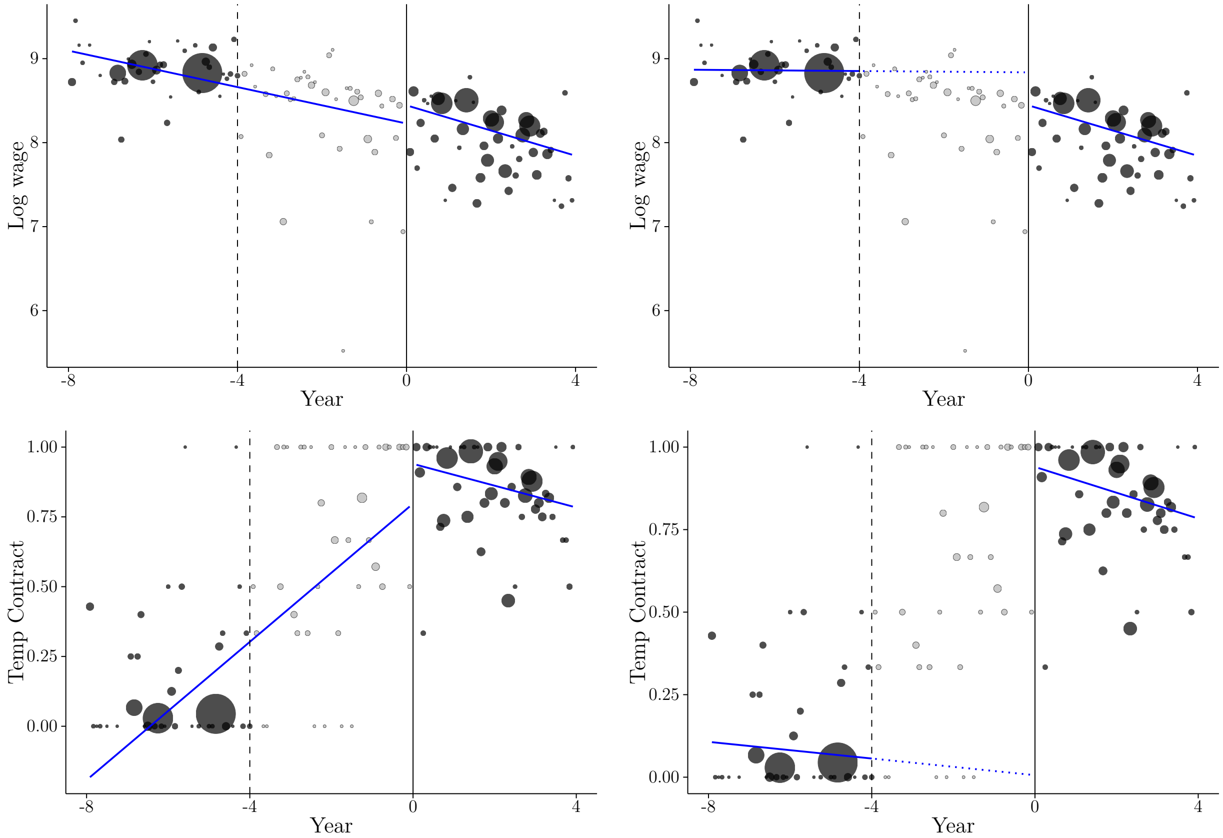

Figure 5: Average log salaries and fraction of temporary contracts around the discontinuity in the year of the reform. Size of dots is proportional to the number of observations. Left panel shows local linear RD fits with a uniform kernel using all available observations. Right panel shows the same specification when data points within four years below the cutoff are excluded (donut hole).

yields a p-value of 0.002, indicating that this difference is unlikely to be due to random sampling alone. Our new Γ test yields only a p-value of 0.088. The estimate of the reduced form parameter, the jump in log wages, is less affected by donut trimming, and both of our specification tests yield p-values well above commonly used significance levels. Taking evidence from both stages together, there is strong evidence against the conventional fuzzy RD design in the data.

7.2. **Low Birthweight.** We next revisit data from Almond et al. (2010), who study the effect of medical care on infant mortality of low birthweight babies.[8] Their study exploits that infants with a birthweight below 1,500g often receive additional medical treatment. Using a conventional regression discontinuity design with a bandwidth of 85g suggests that

[8] We obtained the data set from the NBER webpage: `http://data.nber.org/lbid/adkw.dta`.

Table 1: Estimates of the effect of contract type on (log) wages

	Estimates and confidence intervals				p-values	
	Full sample		Donut RD			
	Estimate	95% CI	Estimate	95% CI	Δ	Γ
τ_T (first stage)	0.14	(-0.31, 0.59)	0.94	(-0.11, 1.99)	0.002	0.088
τ_Y (reduced form)	0.22	(-0.36, 0.79)	-0.39	(-1.62, 0.84)	0.486	0.848
θ (main parameter)	1.52	–	-0.41	–	0.005	0.177

Notes: ...

1-year mortality is approximately 1 percentage point lower just below the 1,500g threshold, compared to 5.5% just above it. The left panel of Figure 1 in the introduction shows the data together with a local linear regression fit.

Barreca et al. (2011) argue that these results are driven by the heaping points at 1500g (rounding) and 1503g (corresponding to 53oz). We revisit their analysis and sequentially exclude observations within $\pm d$ grams of the 1,500g threshold, with d ranging from 0 to 4. All estimates use the original bandwidth of $h = 85$g and a uniform kernel. The right panel of Figure 1 in the introduction shows the data together with a local linear donut regression fit for the case that $d = 3$. We set the smoothness bound to $M = 10^{-6}$, which implies that the conditional mean function μ^* can differ by at most 0.1% from a straight line over the 85g window. The results show that even just including the observations with exactly 1500g yields a "marginally significant" p-value of 10% from our Δ test, and when excluding observations that deviate more than 3g from the treatment threshold, the p-values are below all commonly used significance levels.

Table 2: Estimates of the effect of crossing low birth weight threshold on 1-year infant mortality

d	Estimate (pp)	95% CI	Δ: p-value	Γ: p-value
–	0.95	(0.46, 1.44)	–	–
0	0.54	(0.04, 1.04)	0.102	–
1	0.55	(0.05, 1.05)	0.110	–
2	0.53	(0.03, 1.04)	0.096	0.659
3	0.16	(-0.47, 0.80)	0.001	0.078
4	0.18	(-0.46, 0.82)	0.001	0.008

Note: Alternative Γ specification test may not be appropriate here due to very few support points inside donut (there are d support points to the left and $d+1$ support points to the right of the cutoff here).

8. CONCLUSIONS

This paper provides a theoretical foundation for donut regression discontinuity designs, which are widely used in empirical practice but typically implemented without formal guidance. Our results clarify the trade-offs introduced by donut trimming and allow for valid statistical inference in both point estimation and diagnostic settings. An important open question concerns the choice of the donut size d. We conjecture that if the goal is valid causal inference, this choice cannot be made in a purely data-driven way and will have to rely on some subject-matter knowledge. In a specification testing context, however, there may be scope for a procedure that uses a range of values for d, without assuming that a particular one is the "correct" one.

A. ADDITIONAL THEORETICAL RESULTS

In this section, we collect further results on robust-bias correction confidence intervals and on donut methods for fuzzy RD designs.

...

B. PROOFS OF MAIN RESULTS

In this section, we collect the proofs of the main theoretical results in this paper.

...

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.

Bau, N. and J. Das (2020): "Teacher Value Added in a Low-Income Country," *American Economic Journal: Economic Policy*, 12, 62 – 96.

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

Dong, Y. (2018): "Alternative assumptions to identify LATE in fuzzy regression discontinuity designs," *Oxford Bulletin of Economics and Statistics*, 80, 1020–1027.

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

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

Noack, C. and C. Rothe (2024): "Bias-Aware Inference in Fuzzy Regression Discontinuity Designs," *Econometrica*, 92, 687–711.