Optimal Data-Driven Regression Discontinuity Plots

Sebastian CALONICO, Matias D. CATTANEO, and Rocío TITIUNIK

Exploratory data analysis plays a central role in applied statistics and econometrics. In the popular regression-discontinuity (RD) design, the use of graphical analysis has been strongly advocated because it provides both easy presentation and transparent validation of the design. RD plots are nowadays widely used in applications, despite its formal properties being unknown: these plots are typically presented employing ad hoc choices of tuning parameters, which makes these procedures less automatic and more subjective. In this article, we formally study the most common RD plot based on an evenly spaced binning of the data, and propose several (optimal) data-driven choices for the number of bins depending on the goal of the researcher. These RD plots are constructed either to approximate the underlying unknown regression functions without imposing smoothness in the estimator, or to approximate the underlying variability of the raw data while smoothing out the otherwise uninformative scatterplot of the data. In addition, we introduce an alternative RD plot based on quantile spaced binning, study its formal properties, and propose similar (optimal) data-driven choices for the number of bins. The main proposed data-driven selectors employ spacings estimators, which are simple and easy to implement in applications because they do not require additional choices of tuning parameters. Altogether, our results offer an array of alternative RD plots that are objective and automatic when implemented, providing a reliable benchmark for graphical analysis in RD designs. We illustrate the performance of our automatic RD plots using several empirical examples and a Monte Carlo study. All results are readily available in R and STATA using the software packages described in Calonico, Cattaneo, and Titiunik. Supplementary materials for this article are available online.

KEY WORDS: Binning; Partitioning; RD plots; Tuning parameter selection.

1. INTRODUCTION

The regression discontinuity (RD) design, originally introduced by Thistlethwaite and Campbell (1960), is among the most popular quasi-experimental empirical strategies to estimate (local) causal treatment effects in economics, political science, and many other social, behavioral, and natural sciences. In this research design, for each unit $i = 1, 2, \ldots, n$, researchers observe an outcome variable Y_i and a continuous covariate X_i, and units are assigned to treatment or control depending on whether their observed covariate exceeds a known cutoff. Provided the units of analysis cannot systematically sort around the cutoff, the RD design employs observations just below and just above the cutoff as control and treatment groups to conduct inference on the (local) causal effect of the treatment. The underlying idea, and crucial assumption, is that units around the cutoff do not systematically differ in their unobservable characteristics, thereby offering valid counterfactual comparisons between control and treatment groups. For recent reviews on the RD design, including references to a large number of empirical applications employing RD designs, see, for example, Cook (2008), Imbens and Lemieux (2008), and Lee and Lemieux (2010).

A key feature of the RD design is its simplicity and transparency. The empirical analysis relies on simple and easy-to-interpret identifying assumptions to study the effect of a policy or intervention for units near the threshold, involving only a univariate outcome Y_i and a univariate continuous covariate X_i (which determines treatment assignment). Estimation and inference of RD treatment effects is usually conducted using local polynomial estimators, and great attention has been devoted to these estimators in the recent methodological RD literature (see Hahn, Todd, and van der Klaauw 2001; Porter 2003; Imbens and Kalyanaraman 2012; Calonico, Cattaneo, and Titiunik 2014b, and references therein). Other approaches are also possible, such as those employing randomization inference methods (Cattaneo, Frandsen, and Titiunik 2015).

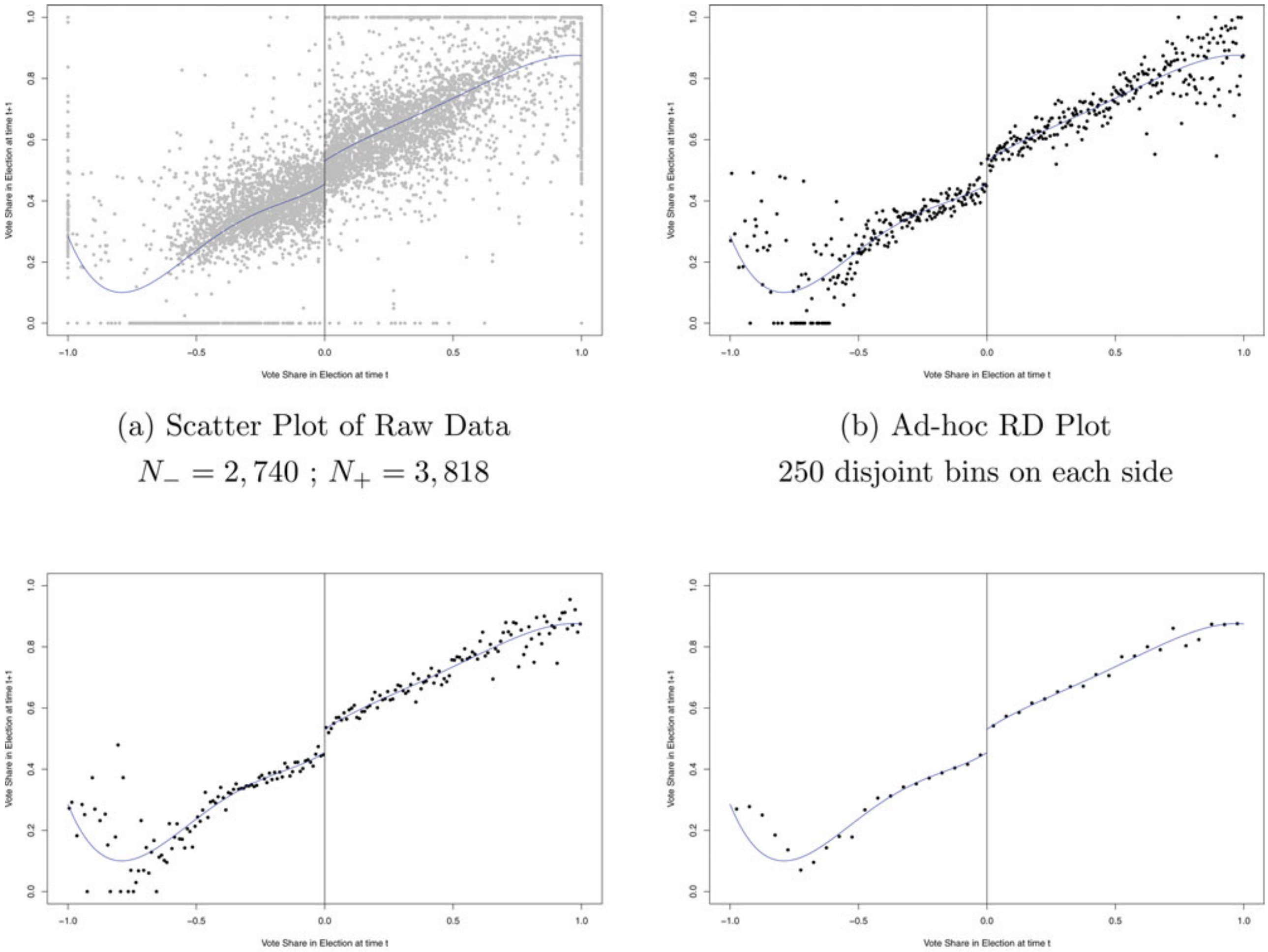
No matter the inference approach employed, graphical exploratory analysis and graphical falsification tests are essential when employing RD designs. These methods have been strongly advocated in the literature because they play an important role in both the presentation and validation of RD research designs—see, for example, Imbens and Lemieux (2008, sec. 3) and Lee and Lemieux (2010, sec. 4.1). The most common graphical representation of RD designs is a plot that contains two main ingredients. The first shows two smooth polynomial approximations of the underlying conditional expectations of the outcome variable Y_i given the observed covariate X_i, for control and treatment units separately. The second ingredient is a collection of local sample means of the outcome variable constructed by partitioning the support of the covariate X_i into disjoint bins for control and treatment units separately, and computing sample averages of the outcome variable Y_i for each bin using only observations whose value of the covariate X_i falls within that bin.

Figure 1 gives three examples of these RD plots using the data of Lee (2008), who studied the vote share advantage enjoyed by the incumbent party in U.S. House of Representatives electoral races. This figure also includes the scatterplot of the raw data for comparison. In this empirical example, the identification strategy is based on the discontinuity generated by the rule that assigns electoral victory to the party that obtains the most votes. The forcing variable (X_i) is the Democratic margin of

Sebastian Calonico is Assistant Professor, Department of Economics, University of Miami, Coral Gables, FL 33124 (E-mail: scalonico@bus.miami.edu). Matias D. Cattaneo is Associate Professor, Department of Economics, University of Michigan, Ann Arbor, MI 48109 (E-mail: cattaneo@umich.edu). Rocío Titiunik is Assistant Professor, Department of Political Science, University of Michigan, Ann Arbor, MI 48109 (E-mail: titiunik@umich.edu). This article has benefited from the insightful suggestions of the co-editor, David Ruppert, an associate editor, and three reviewers. The authors also thank Andreas Hagemann, Guido Imbens, Michael Jansson, Zhuan Pei, and Andres Santos for their comments. Financial support from the National Science Foundation (SES 1357561) is gratefully acknowledged.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/jasa.

Journal of the American Statistical Association
December 2015, Vol. 110, No. 512, Theory and Methods
DOI: 10.1080/01621459.2015.1017578



(a) Scatter Plot of Raw Data
$N_- = 2,740$; $N_+ = 3,818$

(b) Ad-hoc RD Plot
250 disjoint bins on each side

(c) Ad-hoc RD Plot
100 disjoint bins on each side

(d) Ad-hoc RD Plot
20 disjoint bins on each side

Figure 1. Scatterplot and ad hoc RD plots for U.S. House elections data. (a) Scatterplot of raw data $N_- = 2740$; $N_+ = 3818$ (b) ad hoc RD plot 250 disjoint bins on each side (c) ad hoc RD plot 100 disjoint bins on each side (d) ad hoc RD plot 20 disjoint bins on each side. Notes: (i) sample size is $n = 6558$; (ii) N_- and N_+ denote the sample sizes for control and treatment units, respectively; (iii) solid blue lines depict fourth-order polynomial fits using control and treated units separately.

victory in a given election—the difference in vote share between the Democratic candidate and her strongest opponent—and the normalized threshold is $\bar{x} = 0$, since the party wins the election when its margin of victory is positive and loses otherwise. The outcome variable (Y_i) is the Democratic vote share in the following U.S. House election. (We further discuss this empirical application in Section 6.) Each plot in Figure 1 includes fourth-order polynomial fits for control and treatment units separately, and the gray dots in Figure 1(a) represent a raw observation while the black dots in Figure 1(b)–1(d) represent the sample average for each disjoint bin.

The two ingredients of the RD plots serve different goals. The polynomial fits seek to represent the behavior of the underlying conditional expectations in a smooth fashion and from a global perspective. On the other hand, the local sample means have the general goal of providing a visual representation of the design without relying on parametric assumptions regarding the underlying regression functions, while also capturing the local behavior of the data. In particular, local means may serve two purposes:

1. *Detection of discontinuities*. Local means can provide important information regarding the validity of the key identifying assumption of the design—the continuity of the conditional expectations at the cutoff $\bar{x}$. By providing a plot of the underlying regression function that is by construction discontinuous, the local means can highlight the presence of potential discontinuities in the conditional expectations away from the cutoff, which would cast doubt on the key identifying assumption of the design. From this perspective, the binning structure of the RD plot is fundamental: while the global polynomial fits will typically hide the presence of such discontinuities, the local sample means will not. In other words, constructing two distinct estimators of the underlying regression functions, one smooth (the global polynomial fit) and the other discontinuous (the local means), is particularly useful when the goal is to identify the presence of potential discontinuities.
2. *Representation of variability*. A second, equally important, goal of the local sample means is to provide a

> disciplined representation of the overall variability of the data. As shown in Figure 1, a scatterplot of the raw data is uninformative and not particularly revealing of the features of the RD design. In this case, the local sample means play a different role: they are used to construct a somewhat smoothed, yet variable scatterplot of the data by averaging the observations within each of the disjoint bins. From this perspective, the goal is not to "trace out" the underlying regression functions but rather to construct an undersmoothed estimate that highlights visually the overall variability of the data.

Despite general agreement around the purpose of RD plots and their widespread use, their formal properties remain unknown. In particular, these plots are constructed using an ad hoc choice of the partitions' size (i.e., the number of bins used to construct the local sample means), making the procedure less automatic and more subjective than is ideal for a tool whose main role is to provide objective evidence about the plausibility of the research design's main assumptions. Given the absence of concrete guidance on these choices, practitioners typically experiment and select an arbitrary number of bins, which may misrepresent the actual behavior of the data. Figure 1 illustrates some of the potential problems underlying ad hoc RD plots. First, Figure 1(a) shows that the scatterplot of the raw data is highly uninformative: despite the highly significant nonzero RD treatment effect at the cutoff, no discontinuities are noticeable when looking only at the cloud of points—a problem that is exacerbated for binary outcomes. Second, Figures 1(b) through 1(d) show that by choosing the number of bins in an ad hoc way, the type of information conveyed by RD plots can vary widely, which implies that different ad hoc RD plots may give very different representations of the underlying data and, by implication, the validity of the design.

We address these concerns in the construction of RD plots by proposing automatic, data-driven procedures to select the number of bins in RD plots specifically tailored to each of the two goals mentioned above: detection of discontinuities and representation of variability. To provide a plot that is well suited to detecting potential discontinuities in the underlying regression functions, we optimize the number of bins used to compute the local sample means so that the integrated mean square error (IMSE) of the resulting (discontinuous, binned) estimator of the underlying regression functions is minimized. To provide a plot that is appropriate to represent the underlying variability in the data, we develop a bin selector that employs more bins than the optimal number selected by the IMSE-minimization strategy. We formalize this second approach in two distinct ways. First, we propose a different optimal choice of the number of bins based on a weighted integrated mean square error (WIMSE), which gives a formal justification for undersmoothing (i.e., selecting a larger number of bins than the IMSE-optimal choice). Second, we propose another choice of the number of bins, which generates local sample means with an asymptotic variability mimicking the overall variability of the data. Both of these choices lead to undersmoothed tuning parameter selectors relative to the IMSE-optimal choices, thereby generating more variability in the local sample means depicted in the RD plots.

We derive all these optimal choices of the number of bins for two distinct types of RD plots. First, we study the properties of the most common RD plot used in the literature, one that employs an evenly spaced (ES) binning of the data. Second, we introduce an alternative RD plot based on quantile spaced (QS) binning. The latter approach forces each bin to have approximately the same number of observations, a feature that may be appealing when the data are sparse: this partitioning scheme may be interpreted as covariate design adaptive. For each type of RD plot, we derive formally the optimal number of bins selectors mentioned above, and develop data-driven nonparametric consistent implementations thereof. Our main implementations employ spacings estimation techniques to construct the data-driven optimal partition size choices because these estimators do not require additional tuning parameter choices, and thus may be seen as more robust in applications. However, this technique requires continuity of the outcome variable, and hence is not applicable in all possible empirical settings (e.g., binary outcomes). To handle noncontinuous outcomes, we also propose and formally analyze partition size data-driven selectors employing nonparametric polynomial estimators. For this case, the underlying tuning parameter for implementation (i.e., the polynomial power) may be chosen using cross-validation or related methods; see, for example, Ruppert, Wand, and Carroll (2009) for further discussion.

Finally, we also analyze the performance of our automatic RD plots visually and numerically. First, we apply our results to the incumbency advantage example already introduced, and find that our optimal data-driven RD plots perform well when using real data. We also offer a similar analysis of three other empirical applications in the supplemental appendix. Second, we study the finite-sample properties of our results in a Monte Carlo experiment employing several data-generating processes, and find that our RD plots tuning parameter selectors perform extremely well. Third, we compare numerically the two RD plotting alternatives analyzed in this article: ES versus QS. Our results highlight the fact that neither approach dominates the other in general, because features of the underlying (unknown) data-generating process (i.e., distribution of X_i and shapes of the conditional expectation and conditional heteroscedasticity) ultimately determine which RD plot may be preferred.

The rest of the article is organized as follows. Section 2 introduces the RD design, reviews basic results and concepts, and presents a formal description of RD plots. Section 3 introduces the popular ES RD plot, derives formal asymptotic expansions for the variance and bias of the underlying estimator, and employs these results to develop several number of bins selectors depending on the researcher's goal. Section 4 proceeds analogously but for the alternative RD plot based on QS bins. Section 5 presents data-driven, fully automatic implementations of our tuning parameter selectors for RD plots and establishes their consistency properties. Section 6 showcases how our data-driven RD plots perform visually and numerically using both real and simulated data, and briefly compares the quantile and ES approaches. Section 7 discusses two simple extensions, and Section 8 concludes. The supplemental appendix contains the proofs of our main theorems, additional methodological and technical results, detailed simulation evidence, and further empirical illustrations not included here to conserve space. Companion R and

STATA software packages are described in Calonico, Cattaneo, and Titiunik (2014a, 2015).

2. SETUP

In the regression discontinuity design, the observed data are a random sample $(Y_i, X_i)'$, $i = 1, 2, \ldots, n$, from a large population, with X_i a continuous random variable with (possibly restricted) support $[x_l, x_u]$ and continuous density $f(x)$. All units with a value of the observed "score" or "forcing" variable X_i greater than a known threshold $\bar{x}$ are assigned to the treatment group ($T_i = 1$), while all units with $X_i < \bar{x}$ are assigned to the control group ($T_i = 0$). Thus, under perfect compliance, treatment received is defined as $T_i = \mathbb{1}(X_i \geq \bar{x})$ with $\mathbb{1}(\cdot)$ denoting the indicator function. As is common in the program evaluation literature (e.g., Imbens and Wooldridge 2009), we employ potential outcomes notation to characterize the two underlying counterfactual states (control or treatment). Letting $Y_i(1)$ and $Y_i(0)$ denote the potential outcome with and without treatment, respectively, the observed outcome is

$$Y_i = Y_i(0) \cdot (1 - T_i) + Y_i(1) \cdot T_i = \begin{cases} Y_i(0) & \text{if } T_i = 0 \\ Y_i(1) & \text{if } T_i = 1 \end{cases}.$$

The most popular parameter of interest is the average treatment effect at the threshold, given by $\tau_{\text{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = \bar{x}]$. This parameter is nonparametrically identifiable under a mild continuity condition (Hahn, Todd, and van der Klaauw 2001), and RD estimators employing local polynomial techniques have become the default choice in the literature (see Porter 2003; Imbens and Kalyanaraman 2012; Calonico, Cattaneo, and Titiunik 2014b, and references therein). In the so-called sharp RD design, T_i is a deterministic function of treatment assignment (perfect compliance), while in the so-called fuzzy RD design treatment take-up and treatment assignment may differ. This distinction, however, is mostly irrelevant for our purposes because we do not focus on estimation and inference for RD treatment effects, but rather on the RD plots commonly encountered in empirical work. These plots may be used for presentation and falsification of both sharp and fuzzy RD research designs. See Section 7 for a brief discussion of how our results may be applied to fuzzy RD designs or extended to allow for other covariates entering the analysis.

We set

$$\mu_-(x) = \mathbb{E}[Y_i(0)|X_i = x], \qquad \mu_-^{(1)}(x) = \frac{\partial}{\partial x}\mu_-(x),$$
$$\sigma_-^2(x) = \mathbb{V}[Y_i(0)|X_i = x],$$
$$\mu_+(x) = \mathbb{E}[Y_i(1)|X_i = x], \qquad \mu_+^{(1)}(x) = \frac{\partial}{\partial x}\mu_+(x),$$
$$\sigma_+^2(x) = \mathbb{V}[Y_i(1)|X_i = x],$$

and impose the following assumption through the article.

Assumption 1. For $x_l, x_u \in \mathbb{R}$ with $x_l < \bar{x} < x_u$, and all $x \in [x_l, x_u]$:

a. $\mathbb{E}[Y_i^4|X_i]$ is bounded, and $f(x)$ is continuous and bounded away from zero.
b. $\mu_-(x)$ and $\mu_+(x)$ are S times continuously differentiable ($S \geq 1$).
c. $\sigma_-^2(x)$ and $\sigma_+^2(x)$ are continuous and bounded away from zero.

Part (a) in Assumption 1 imposes existence of moments and requires that the running variable X_i be continuously distributed. Part (b) imposes smoothness on the underlying regression functions, while part (c) requires that the conditional variance be continuous; all these functions may be different at either side of the threshold. Notice that $\mu_-(x) = \mathbb{E}[Y_i|X_i = x]$ for all $x < \bar{x}$ and $\mu_+(x) = \mathbb{E}[Y_i|X_i = x]$ for all $x \geq \bar{x}$, enabling (consistent) estimation of these conditional expectations for control and treatment units, respectively.

2.1 RD Plots

The main features of an RD design are easily summarized employing RD plots. As mentioned previously, these plots include two main ingredients: (i) smooth polynomial estimation, and (ii) disjoint local sample means estimation. We now formalize the underlying estimation approaches used to construct the RD plots, which provides the basis for our analysis. Our main focus is on tuning parameter selection for the construction of the collection of local sample means under two distinct partitioning schemes: ES and QS partitions of $[x_l, \bar{x})$ and $[\bar{x}, x_u]$, that is, of the observations to the left and right of the cutoff.

2.1.1 Global Polynomial Estimation. In the RD plots, the unknown functions $\mu_-(x) = \mathbb{E}[Y_i(0)|X_i = x]$ and $\mu_+(x) = \mathbb{E}[Y_i(1)|X_i = x]$ are estimated using global polynomials for control and treatment observations separately. To formalize this approach, let $k \in \mathbb{Z}_+$ and $\mathbf{r}_k(x) = (1, x, x^2, \ldots, x^k)'$, and define

$$\hat{\mu}_{-,k}(x) = \mathbf{r}_k(x)'\hat{\boldsymbol{\beta}}_{-,k},$$
$$\hat{\boldsymbol{\beta}}_{-,k} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \sum_{i=1}^{n} \mathbb{1}(X_i < \bar{x})(Y_i - \mathbf{r}_k(x)'\boldsymbol{\beta})^2,$$
$$\hat{\mu}_{+,k}(x) = \mathbf{r}_k(x)'\hat{\boldsymbol{\beta}}_{+,k},$$
$$\hat{\boldsymbol{\beta}}_{+,k} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{k+1}} \sum_{i=1}^{n} \mathbb{1}(X_i \geq \bar{x})(Y_i - \mathbf{r}_k(x)'\boldsymbol{\beta})^2.$$

In words, $\hat{\mu}_{-,k}(x)$ and $\hat{\mu}_{+,k}(x)$ are kth-order polynomial fits of Y_i on X_i employing only control and treatment units, respectively.

These polynomial regressions may be viewed as a nonparametric approach, usually called series or (linear) sieve estimation, for the approximation of the underlying population conditional expectations when $k = k_n \to \infty$ as $n \to \infty$ (see, e.g., Newey 1997; Chen 2007; Ruppert, Wand, and Carroll 2009; Belloni et al. 2015 for reviews). Below we will exploit this interpretation explicitly to construct consistent plug-in rules for the optimal tuning parameter choices. Employing results from the nonparametrics literature, it is possible to select k_n using some data-driven approach such as (plug-in) IMSE minimization or cross-validation. In practice, however, $k = 4$ or $k = 5$ are almost always the preferred choices. Either way, we do not discuss further the choice of k for RD plots because this is a well-understood problem. Instead, our main focus is on choosing the partition size for the local means, a result that is not currently available in the literature.

Global polynomial approximations may not perform well in RD applications and, more generally, in approximating regression functions locally. These polynomial approximations for regression functions tend to (i) generate counterintuitive weighting schemes (Gelman and Imbens 2014), (ii) have erratic behavior near the boundaries of the support (usually known as the Runge's phenomenon in approximation theory), and (iii) oversmooth (by construction) potential discontinuities in the interior of the support. Thus, the other crucial ingredient of RD plots is a collection of disjoint local sample means, which we describe formally next.

2.1.2 Local Sample Mean Estimation. The second ingredient in the RD plot is a collection of local sample means of the outcome variable computed over a disjoint partition of the support of the running variable, for control and treatment units separately. To describe this construction formally, we employ ideas from the nonparametric literature on partitioning estimators (for further details, see Cattaneo and Farrell 2013, and references therein).

We define $\mathcal{P}_{-,n} = \{P_{-,j} : j = 1, 2, \ldots, J_{-,n}\}$ and $\mathcal{P}_{+,n} = \{P_{+,j} : j = 1, 2, \ldots, J_{+,n}\}$, two generic disjoint partitions of the support of the running variable X_i to the left and right of the cutoff, which vary with the sample size n. More precisely,

$$[x_l, \bar{x}) = \bigcup_{j=1}^{J_{-,n}} P_{-,j},$$

$$P_{-,j} = \begin{cases} [x_l, p_{-,1}) & j = 1 \\ [p_{-,j-1}, p_{-,j}) & j = 2, \ldots, J_{-,n} - 1 \\ [p_{-,J_{-,n}-1}, \bar{x}) & j = J_{-,n} \end{cases}$$

and

$$[\bar{x}, x_u] = \bigcup_{j=1}^{J_{+,n}} P_{+,j},$$

$$P_{+,j} = \begin{cases} [\bar{x}, p_{+,1}) & j = 1 \\ [p_{+,j-1}, p_{+,j}) & j = 2, \ldots, J_{+,n} - 1 \\ [p_{+,J_{+,n}-1}, x_u] & j = J_{+,n} \end{cases}$$

with $J_{-,n}, J_{+,n} \in \mathbb{Z}_{++}$ denoting the partition sizes for control and treatment groups, respectively. As an example, in the incumbency advantage illustration we introduced above $x_u = 100$, $\bar{x} = 0$, and $x_l = -100$, so a partition to the right of the cutoff in 20-percentage-point increments would be $[\bar{x}, x_u] = [0, 20) \cup [20, 40) \cup [40, 60) \cup [60, 80) \cup [80, 100]$.

We set $\mathbb{1}_A(x) = \mathbb{1}(x \in A)$ to save notation. The partitioning estimators (of order 1), sometimes called binning estimators or local constant regression estimators, are formally described as follows:

$$\hat{\mu}_-(x; J_{-,n}) = \sum_{j=1}^{J_{-,n}} \mathbb{1}_{P_{-,j}}(x)\bar{Y}_{-,j},$$

$$\bar{Y}_{-,j} = \frac{\mathbb{1}(N_{-,j} > 0)}{N_{-,j}} \sum_{i=1}^{n} \mathbb{1}_{P_{-,j}}(X_i)Y_i$$

$$\hat{\mu}_+(x; J_{+,n}) = \sum_{j=1}^{J_{+,n}} \mathbb{1}_{P_{+,j}}(x)\bar{Y}_{+,j},$$

$$\bar{Y}_{+,j} = \frac{\mathbb{1}(N_{+,j} > 0)}{N_{+,j}} \sum_{i=1}^{n} \mathbb{1}_{P_{+,j}}(X_i)Y_i$$

with

$$N_{-,j} = \sum_{i=1}^{n} \mathbb{1}_{P_{-,j}}(X_i), \quad N_- = \sum_{j=1}^{J_{-,n}} N_{-,j},$$

$$N_{+,j} = \sum_{i=1}^{n} \mathbb{1}_{P_{+,j}}(X_i), \quad N_+ = \sum_{j=1}^{J_{+,n}} N_{+,j}.$$

The estimators $\hat{\mu}_-(x; J_{-,n})$ and $\hat{\mu}_+(x; J_{+,n})$ collect the sample means of the outcomes Y_i for observations with covariate X_i taking values within each bin in the partitions $\mathcal{P}_{-,n}$ and $\mathcal{P}_{+,n}$, and may be interpreted as nonparametric estimators of $\mu_-(x)$ and $\mu_+(x)$, respectively. Like other nonparametric procedures, these binning-type estimators involve a choice of tuning and smoothing parameters. In this case, $(J_{-,n}, J_{+,n})$ may be regarded as the tuning parameters (e.g., similar to a bandwidth for conventional kernel estimators) and $(\mathcal{P}_{-,n}, \mathcal{P}_{+,n})$ may be viewed as the smoothing parameters (e.g., similar to the shape of kernel function for conventional kernel estimators). Under Assumption 1, and provided a well-behaved partitioning scheme is used, it is not difficult to show that $\hat{\mu}_-(x; J_{-,n}) \to_{\mathbb{P}} \mu_-(x)$ and $\hat{\mu}_+(x; J_{+,n}) \to_{\mathbb{P}} \mu_+(x)$, provided that $J_{-,n} \to \infty$ and $J_{+,n} \to \infty$ as $n \to \infty$ and some regularity conditions hold. Throughout the article all limits are taken as $n \to \infty$ unless otherwise stated.

The behavior of these estimators is dependent on how the partitions are constructed and, as mentioned above, this article considers two approaches for choosing the partitions: ES partitions and QS partitions. Given a chosen partitioning scheme, the parameters $J_{-,n}$ and $J_{+,n}$ control the rate of approximation of the partitioning estimators, capturing the usual variance and bias trade-off: larger $(J_{-,n}, J_{+,n})$ imply more variance but less bias (more, smaller bins), while smaller $(J_{-,n}, J_{+,n})$ imply less variance but more bias (fewer, larger bins). The main contribution of this article is to formalize these ideas for each of the two partitioning schemes, to derive several (optimal) choices of $(J_{-,n}, J_{+,n})$ explicitly capturing the specific objective of the RD plot (i.e., tracing out the regression function or capturing the underlying variability of the data), and to develop consistent data-driven implementations thereof.

3. EVENLY SPACED RD PLOTS

In this section, we consider ES bins for the construction of the partitioning scheme underlying the RD plots. Thus, we set

$$p_{-,j} = x_l + j \cdot \frac{\bar{x} - x_l}{J_{-,n}} \qquad \text{and} \qquad p_{+,j} = \bar{x} + j \cdot \frac{x_u - \bar{x}}{J_{+,n}},$$

leading to the ES partitioning estimators denoted by $\hat{\mu}_{\mathrm{ES},-}(x; J_{-,n})$ and $\hat{\mu}_{\mathrm{ES},+}(x; J_{+,n})$, with nonrandom partitioning schemes denoted by $\mathcal{P}_{\mathrm{ES},-,n}$ and $\mathcal{P}_{\mathrm{ES},+,n}$, respectively.

The use of ES bins is the most common strategy in the ad hoc construction of RD plots. For example, the original incumbency advantage plots in Lee (2008) present local means in fixed-length bins that are 0.5 percentage points wide. Using the notation just introduced, this translates into $J_{-,n} = J_{+,n} = 200$, since there are 200 bins of length 0.5 on either side of the cutoff between $x_l = -100$ and $x_u = 100$, and the two bins closest to the cutoff on either side of it are

$$p_{-,199} = -100 + 199 \cdot \frac{100}{200} = -0.5$$

and

$$p_{+,1} = 0 + 1 \cdot \frac{100}{200} = 0.5.$$

3.1 Variance and Bias Properties

To study formally the properties of the ES RD plots, we begin by developing formal asymptotic expansions for the integrated variance and bias of the underlying partitioning estimators. Let $w(x)$ denote a weighting function, formally introduced in Theorem 1, and set $\mathbf{X}_n = (X_1, X_2, \ldots, X_n)'$ to save notation. The integrated variance of the estimators, for control and treatment groups, $\hat{\mu}_{\mathrm{ES},-}(x; J_{-,n})$ and $\hat{\mu}_{\mathrm{ES},+}(x; J_{+,n})$, are

$$\mathrm{var}_{\mathrm{ES},-}(J_{-,n}) = \int_{x_l}^{\bar{x}} \mathbb{V}\left[\hat{\mu}_{\mathrm{ES},-}(x; J_{-,n}) \middle| \mathbf{X}_n\right] w(x)dx$$
$$\mathrm{var}_{\mathrm{ES},+}(J_{+,n}) = \int_{\bar{x}}^{x_u} \mathbb{V}\left[\hat{\mu}_{\mathrm{ES},+}(x; J_{+,n}) \middle| \mathbf{X}_n\right] w(x)dx.$$

Similarly, the integrated squared bias for these estimators is

$$\mathrm{Bias}_{\mathrm{ES},-}(J_{-,n}) = \int_{x_l}^{\bar{x}} (\mathbb{E}\left[\hat{\mu}_{\mathrm{ES},-}(x; J_{-,n}) \middle| \mathbf{X}_n\right] - \mu_-(x))^2 w(x)dx$$
$$\mathrm{Bias}_{\mathrm{ES},+}(J_{+,n}) = \int_{\bar{x}}^{x_u} (\mathbb{E}\left[\hat{\mu}_{\mathrm{ES},+}(x; J_{+,n}) \middle| \mathbf{X}_n\right] - \mu_+(x))^2 w(x)dx.$$

Since variability plays a crucial role in the construction of the RD plots, all of our selectors will use the variance quantities $\mathrm{var}_{\mathrm{ES},-}(J_{-,n})$ and $\mathrm{var}_{\mathrm{ES},+}(J_{+,n})$. In some cases, we will also employ the bias quantities $\mathrm{Bias}_{\mathrm{ES},-}(J_{-,n})$ and $\mathrm{Bias}_{\mathrm{ES},+}(J_{+,n})$ to construct choices of number of bins $J_{-,n}$ and $J_{+,n}$. The next result gives a formal first-order nonparametric approximation to the integrated variance and squared bias of the estimators.

Theorem 1. Suppose Assumption 1 holds with $S \geq 2$, and $w : [x_l, x_u] \mapsto \mathbb{R}_+$ is continuous.

$(-)$ If $J_{-,n} \log(J_{-,n})/n \to 0$ and $J_{-,n} \to \infty$, then

$$\mathrm{var}_{\mathrm{ES},-}(J_{-,n}) = \frac{J_{-,n}}{n} \mathscr{V}_{\mathrm{ES},-}\{1 + o_{\mathbb{P}}(1)\},$$
$$\mathscr{V}_{\mathrm{ES},-} = \frac{1}{\bar{x} - x_l} \int_{x_l}^{\bar{x}} \frac{\sigma_-^2(x)}{f(x)} w(x)dx,$$
$$\mathrm{Bias}_{\mathrm{ES},-}(J_{-,n}) = \frac{1}{J_{-,n}^2} \mathscr{B}_{\mathrm{ES},-}\{1 + o_{\mathbb{P}}(1)\},$$
$$\mathscr{B}_{\mathrm{ES},-} = \frac{(\bar{x} - x_l)^2}{12} \int_{x_l}^{\bar{x}} \left(\mu_-^{(1)}(x)\right)^2 w(x)dx.$$

$(+)$ If $J_{+,n} \log(J_{+,n})/n \to 0$ and $J_{+,n} \to \infty$, then

$$\mathrm{var}_{\mathrm{ES},+}(J_{+,n}) = \frac{J_{+,n}}{n} \mathscr{V}_{\mathrm{ES},+}\{1 + o_{\mathbb{P}}(1)\},$$
$$\mathscr{V}_{\mathrm{ES},+} = \frac{1}{x_u - \bar{x}} \int_{\bar{x}}^{x_u} \frac{\sigma_+^2(x)}{f(x)} w(x)dx,$$
$$\mathrm{Bias}_{\mathrm{ES},+}(J_{+,n}) = \frac{1}{J_{+,n}^2} \mathscr{B}_{\mathrm{ES},+}\{1 + o_{\mathbb{P}}(1)\},$$
$$\mathscr{B}_{\mathrm{ES},+} = \frac{(x_u - \bar{x})^2}{12} \int_{\bar{x}}^{x_u} \left(\mu_+^{(1)}(x)\right)^2 w(x)dx.$$

All the results presented in this article remain valid if $w(x) = w_+(x)\mathbb{1}(x \geq \bar{x}) + w_-(x)\mathbb{1}(x < \bar{x})$, thus allowing for $w(x)$ to be discontinuous at $\bar{x}$. Theorem 1 captures formally the natural trade-off between variability and bias in approximating the underlying regression function using local sample means computed using disjoint ES partitions of size $J \in \{J_{-,n}, J_{+,n}\}$: the larger the J, the smaller the bias because each bin is smaller and hence the sample mean approximates the underlying function better, while the larger the J, the larger the variance because each bin is small and hence has only a few observations. In what follows, we use this intuition explicitly to develop different tuning parameter selectors depending on the explicit goal in mind.

3.2 Approximating the Underlying Regression Functions

As a first goal, we consider optimal choices of the number of bins $J_{-,n}$ and $J_{+,n}$ with the explicit goal of approximating the underlying regression function in an IMSE sense. As discussed previously, the resulting selectors are important and useful for empirical work because they validate the otherwise possibly oversmoothed polynomial approximations to the underlying regression functions. Thus, we recommend these selectors to construct RD plots to visually check for potential discontinuities in the regression functions. Once potential discontinuities have been identified, formal hypothesis tests ("placebo" tests) may be conducted using robust inference procedures from the RD literature (e.g., Calonico, Cattaneo, and Titiunik 2014b).

Under the conditions of Theorem 1, the IMSE loss function of the estimators underlying the ES RD plots satisfies

$$\mathrm{IMSE}_{\mathrm{ES},-}(J_{-,n}) = \int_{x_l}^{\bar{x}} \mathbb{E}[(\hat{\mu}_{\mathrm{ES},-}(x; J_{-,n}) - \mu_-(x))^2 | \mathbf{X}_n] w(x)dx = \frac{J_{-,n}}{n} \mathscr{V}_{\mathrm{ES},-}\{1 + o_{\mathbb{P}}(1)\} + \frac{1}{J_{-,n}^2} \mathscr{B}_{\mathrm{ES},-}\{1 + o_{\mathbb{P}}(1)\}$$

and

$$\mathrm{IMSE}_{\mathrm{ES},+}(J_{+,n}) = \int_{\bar{x}}^{x_u} \mathbb{E}\left[(\hat{\mu}_{\mathrm{ES},+}(x; J_{+,n}) - \mu_+(x))^2 \middle| \mathbf{X}_n\right] w(x)dx = \frac{J_{+,n}}{n} \mathscr{V}_{\mathrm{ES},+}\{1 + o_{\mathbb{P}}(1)\} + \frac{1}{J_{+,n}^2} \mathscr{B}_{\mathrm{ES},+}\{1 + o_{\mathbb{P}}(1)\}.$$

These results give an approximation to a family of IMSE loss functions, depending on the choice of weight function $w(x)$. In general, assuming that $\mathscr{B}_{\mathrm{ES},-} \neq 0$ and $\mathscr{B}_{\mathrm{ES},+} \neq 0$, the expansions of $\mathrm{IMSE}_{\mathrm{ES},-}(J_{n,-})$ and $\mathrm{IMSE}_{\mathrm{ES},+}(J_{n,+})$ give the optimal choices of number of bins:

$$J_{\mathrm{ES}\text{-}\mu,-,n} = \left\lceil \left(\frac{2\mathscr{B}_{\mathrm{ES},-}}{\mathscr{V}_{\mathrm{ES},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$J_{\mathrm{ES}\text{-}\mu,+,n} = \left\lceil \left(\frac{2\mathscr{B}_{\mathrm{ES},+}}{\mathscr{V}_{\mathrm{ES},+}}\right)^{1/3} n^{1/3} \right\rceil \qquad (1)$$

with $y = \lceil x \rceil \in \mathbb{N}$, $x \in \mathbb{R}_{++}$, denoting the smallest integer y such that $x \leq y$.

3.3 Approximating the Underlying Variability of the Data

In addition to developing optimal choices of tuning parameters for approximating the underlying regression functions using local sample means, we propose two distinct approaches specifically developed to represent the overall variability of the data. As discussed in Section 5, the resulting implementations are objective, fully automatic, and easy-to-implement, yet disciplined, thus providing useful benchmarks for smoothing out the scatterplot of the raw data in empirical applications.

The first approach employs a natural loss function and leads to a formal justification for "manual" increases in the variability of the ES RD plots (i.e., undersmoothing by increasing the number of bins employed), which is commonly done in practice. The second approach is fully automatic but is not loss function based—it is rather specifically tailored to mimic the underlying variability of the data while employing binned sample means. Naturally, as we discuss in more detail below, given a fixed sample size, the resulting choices from the second approach (or any other approach) can always be rationalized as emerging from the first approach under a particular weighting scheme, and hence could be regarded as "optimal." In this sense, the first approach is useful at the very least insofar as it gives a formal, intuitive justification for any specific choice of number of bins in terms of trading off variance and bias of the underlying local means estimators entering the construction of the ES RD plot.

3.3.1 Weighted IMSE. This approach is based on a family of loss functions constructed by trading off variance and bias of the partitioning estimators. Specifically, to capture the variability of the underlying raw data, a natural approach is to undersmooth the binned sample means estimators (i.e., select a larger number of bins $J_{-,n}$ and $J_{+,n}$). This can be accomplished by trading off variance and bias differently: let $\omega_{\mathscr{V},-}, \omega_{\mathscr{B},-}, \omega_{\mathscr{V},+}, \omega_{\mathscr{B},+} > 0$ be fixed weights satisfying $\omega_{\mathscr{V},-} + \omega_{\mathscr{B},-} = 1$ and $\omega_{\mathscr{V},+} + \omega_{\mathscr{B},+} = 1$, then consider the family of weighted IMSE loss functions given by

$$\begin{aligned}&\mathrm{WIMSE}_{\mathrm{ES},-}(J_{-,n})\\&= \omega_{\mathscr{V},-}\mathrm{var}_{\mathrm{ES},-}(J_{-,n}) + \omega_{\mathscr{B},-}\mathrm{Bias}_{\mathrm{ES},-}(J_{-,n})\\&= \omega_{\mathscr{V},-}\frac{J_{-,n}}{n}\mathscr{V}_{\mathrm{ES},-}\{1+o_{\mathbb{P}}(1)\}+\omega_{\mathscr{B},-}\frac{1}{J_{-,n}^2}\mathscr{B}_{\mathrm{ES},-}\{1+o_{\mathbb{P}}(1)\}\end{aligned}$$

and

$$\begin{aligned}&\mathrm{WIMSE}_{\mathrm{ES},+}(J_{+,n})\\&= \omega_{\mathscr{V},+}\mathrm{var}_{\mathrm{ES},+}(J_{+,n}) + \omega_{\mathscr{B},+}\mathrm{Bias}_{\mathrm{ES},+}(J_{+,n})\\&= \omega_{\mathscr{V},+}\frac{J_{+,n}}{n}\mathscr{V}_{\mathrm{ES},+}\{1+o_{\mathbb{P}}(1)\} + \omega_{\mathscr{B},+}\frac{1}{J_{+,n}^2}\mathscr{B}_{\mathrm{ES},+}\{1+o_{\mathbb{P}}(1)\},\end{aligned}$$

where these expansions are formally justified under the conditions given in Theorem 1. It follows that the optimal choices based on the above loss functions are

$$J_{\mathrm{ES}\text{-}\omega,-,n} = \left\lceil \omega_-\left(\frac{2\mathscr{B}_{\mathrm{ES},-}}{\mathscr{V}_{\mathrm{ES},-}}\right)^{1/3} n^{1/3}\right\rceil$$

and

$$J_{\mathrm{ES}\text{-}\omega,+,n} = \left\lceil \omega_+\left(\frac{2\mathscr{B}_{\mathrm{ES},+}}{\mathscr{V}_{\mathrm{ES},+}}\right)^{1/3} n^{1/3}\right\rceil \qquad (2)$$

with $\omega_- = (\omega_{\mathscr{B},-}/\omega_{\mathscr{V},-})^{1/3}$ and $\omega_+ = (\omega_{\mathscr{B},+}/\omega_{\mathscr{V},+})^{1/3}$. The WIMSE objective functions are meant to offer more flexibility on the relative importance of variance and bias when searching for an optimal number bins, and hence the associated weights could be interpreted as capturing researchers' prior beliefs on the relative importance of variance and bias.

The result in (2) is a generalization of the choices given in (1) because $J_{\mathrm{ES}\text{-}\omega,-,n} = \lceil \omega_- J_{\mathrm{ES}\text{-}\mu,-,n}\rceil$ and $J_{\mathrm{ES}\text{-}\omega,+,n} = \lceil \omega_+ J_{\mathrm{ES}\text{-}\mu,+,n}\rceil$, and when variance and bias are weighted equally (i.e., $\omega_{\mathscr{V},-} = \omega_{\mathscr{B},-}$ and $\omega_{\mathscr{V},+} = \omega_{\mathscr{B},+}$), then $J_{\mathrm{ES}\text{-}\mu,-,n} = J_{\mathrm{ES}\text{-}\omega,-,n}$ and $J_{\mathrm{ES}\text{-}\mu,+,n} = J_{\mathrm{ES}\text{-}\omega,+,n}$. More generally, the larger the $\omega_{\mathscr{B},-}, \omega_{\mathscr{B},+} \in (0, 1)$, the larger the choice of number of bins $J_{\mathrm{ES}\text{-}\omega,-,n}$ and $J_{\mathrm{ES}\text{-}\omega,+,n}$ because the loss function puts more weight on bias and less on variance, allowing for more variability in the underlying local sample mean estimates. While it is not obvious how to choose a particular weighting scheme in empirical practice, this approach is very useful in justifying "manual" undersmoothing after selecting the number of bins using the IMSE-optimal choices. Specifically, for each choice of rescaling constants $\omega_-, \omega_+ > 0$, there exists a unique compatible weighting scheme:

$$(\omega_{\mathscr{V},-}, \omega_{\mathscr{B},-}) = \left(\frac{1}{1+\omega_-^3}, \frac{\omega_-^3}{1+\omega_-^3}\right)$$

and

$$(\omega_{\mathscr{V},+}, \omega_{\mathscr{B},+}) = \left(\frac{1}{1+\omega_+^3}, \frac{\omega_+^3}{1+\omega_+^3}\right),$$

which rationalizes the resulting choices of number of bins as optimal in the sense of minimizing the WIMSE loss function. This result may be of interest for practitioners because it helps explain how variability and bias are traded off when choosing a scaling factor to modify the IMSE-optimal choices for the number of bins, which can always be used as a starting point in the empirical investigation. In the supplemental appendix, we provide a table with the weights implied by different scaling factors ω_- and ω_+. Furthermore, for any initial ad hoc choice of number of bins used to construct the ES RD plot, the above logic can be used to find an IMSE-optimal choice and a scaling factor that is consistent with the ad hoc choice, thereby offering an objective interpretation of the ad hoc choice in terms of variance and bias trade-off.

3.3.2 Mimicking Variability. The weighted IMSE approach is useful to give a natural interpretation to "manual" undersmoothing and, more generally, to other ad hoc choices of number of bins used to construct ES RD plots approximating the underlying variability of the data. However, this approach is not fully automatic in general: while clearly objective and interpretable, its main drawback is that it requires the choice of a weighting scheme, and it is difficult to justify a scheme that works generally for all applications. For this reason, we also propose a second approach specifically targeted to capture the variability of the data while employing local sample means, which is fully automatic and can be easily implemented.

In this second approach, we choose the number of bins so that the binned sample means have an asymptotic (integrated) variability approximately equal to the amount of variability of the raw data. To describe the approach formally, let $\mathcal{V}_-$ and $\mathcal{V}_+$ denote, respectively, the sample variance of the outcome variables for control and treatment units, that is, the sample variance of the subsamples $\{Y_i : X_i < \bar{x}\}$ and $\{Y_i : X_i \geq \bar{x}\}$. Then, we select $J_{-,n}$ and $J_{+,n}$ so that

$$\mathrm{var}_{\mathrm{ES},-}(J_{-,n}) = \mathcal{V}_-$$

and

$$\mathrm{var}_{\mathrm{ES},+}(J_{+,n}) = \mathcal{V}_+$$

leading, respectively, to the "optimal" choices $\frac{\mathcal{V}_-}{\mathscr{V}_{\mathrm{ES},-}}n$ and $\frac{\mathcal{V}_+}{\mathscr{V}_{\mathrm{ES},+}}n$. The main intuition behind these choices is that we set the number of bins used so that the overall variability of the sample means, as measured by the asymptotic approximation obtained in Theorem 1, mimics the overall variability of the unrestricted scatterplot of the data.

This idea, while very intuitive, has a minor technical drawback: it leads to tuning parameter choices that do not satisfy the rate conditions of the results in Theorem 1. Thus, to make the end result theoretically coherent, we modify it slightly as follows:

$$J_{\mathrm{ES}\text{-}\vartheta,-,n} = \left\lceil \frac{\mathcal{V}_-}{\mathscr{V}_{\mathrm{ES},-}} \frac{n}{\log(n)^2} \right\rceil$$

and

$$J_{\mathrm{ES}\text{-}\vartheta,+,n} = \left\lceil \frac{\mathcal{V}_+}{\mathscr{V}_{\mathrm{ES},+}} \frac{n}{\log(n)^2} \right\rceil. \tag{3}$$

To summarize, the choice emerging for the number of bins in (3) mimics the overall variability of the data, up to a $\log(n)$ factor, and is fully consistent with the theoretical results given in Theorem 1. Importantly, the resulting number of bins will be in general larger than the one obtained in (1), which is consistent with the underlying distinct goals justifying these rules: $(J_{\mathrm{ES}\text{-}\mu,-,n}, J_{\mathrm{ES}\text{-}\mu,+,n})$ are developed explicitly to approximate the underlying regression function and hence they optimally trade-off variance and bias, while $(J_{\mathrm{ES}\text{-}\vartheta,-,n}, J_{\mathrm{ES}\text{-}\vartheta,+,n})$ are developed explicitly to approximate the variability of the data and hence the resulting underlying estimators lead to undersmoothing relative to the IMSE-optimal choices.

4. QUANTILE SPACED RD PLOTS

In addition to the popular ES RD plot, we also introduce and study an alternative plotting approach based on QS bins. This approach takes into account the sparsity of the data, forcing each bin to have approximately the same number of observations. This feature may be appealing because with QS bins the variability of the local sample means will change across bins only due to nonconstant conditional variances (i.e., due to the presence of heteroscedasticity), but not due to different sample sizes in each bin (as it occurs with an ES partition).

This section parallels the previous discussion for ES RD plots in Section 3, but now focusing on QS RD plots. In this case, we construct the partitioning scheme as follows:

$$p_{-,j} = \hat{F}_-^{-1}\left(\frac{j}{J_{-,n}}\right) \qquad \text{and} \qquad p_{+,j} = \hat{F}_+^{-1}\left(\frac{j}{J_{+,n}}\right),$$

with

$$\begin{aligned}
\hat{F}_-^{-1}(y) &= \inf\{x : \hat{F}_-(x) \geq y\}, \\
\hat{F}_-(x) &= \frac{1}{N_-}\sum_{i=1}^{n} \mathbb{1}(X_i < \bar{x})\mathbb{1}(X_i \leq x), \\
\hat{F}_+^{-1}(y) &= \inf\{x : \hat{F}_+(x) \geq y\}, \\
\hat{F}_+(x) &= \frac{1}{N_+}\sum_{i=1}^{n} \mathbb{1}(X_i \geq \bar{x})\mathbb{1}(X_i \leq x).
\end{aligned}$$

In words, the QS RD plot sets $p_{-,j}$ and $p_{+,j}$ to be the approximately $100(j/J_{-,n})$th quantiles of the subsample $\{X_i : X_i < \bar{x}\}$ and the approximately $100(j/J_{+,n})$th quantile of the subsample $\{X_i : X_i \geq \bar{x}\}$, respectively. This construction leads to the QS partitioning estimators denoted by $\hat{\mu}_{\mathtt{QS},-}(x; J_{-,n})$ and $\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n})$, which are estimators now employing the random partitioning schemes denoted by $\mathcal{P}_{\mathtt{QS},-,n}$ and $\mathcal{P}_{\mathtt{QS},+,n}$, respectively.

4.1 Variance and Bias Properties

We study first the integrated variance and squared bias of the estimators $\hat{\mu}_{\mathtt{QS},-}(x; J_{-,n})$ and $\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n})$, which are given by

$$\begin{aligned}
\mathrm{var}_{\mathtt{QS},-}(J_{-,n}) &= \int_{x_l}^{\bar{x}} \mathbb{V}\left[\hat{\mu}_{\mathtt{QS},-}(x; J_{-,n}) \middle| \mathbf{X}_n\right] w(x)dx, \\
\mathrm{var}_{\mathtt{QS},+}(J_{+,n}) &= \int_{\bar{x}}^{x_u} \mathbb{V}\left[\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n}) \middle| \mathbf{X}_n\right] w(x)dx,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Bias}_{\mathtt{QS},-}(J_{-,n}) &= \int_{x_l}^{\bar{x}} (\mathbb{E}\left[\hat{\mu}_{\mathtt{QS},-}(x; J_{-,n}) \middle| \mathbf{X}_n\right] - \mu_-(x))^2 w(x)dx, \\
\mathrm{Bias}_{\mathtt{QS},+}(J_{+,n}) &= \int_{\bar{x}}^{x_u} (\mathbb{E}\left[\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n}) \middle| \mathbf{X}_n\right] - \mu_+(x))^2 w(x)dx.
\end{aligned}$$

As in the case of ES RD plots, we propose several (optimal, data-driven) choices of the number of bins $J_{-,n}$ and $J_{+,n}$ by either trading off variance and bias of the underlying estimators, or by mimicking the overall variability of the raw data. The following result gives the formal expansions for the variance and bias of the underlying partitioning estimators.

Theorem 2. Suppose Assumption 1 holds with $S \geq 2$, and $w : [x_l, x_u] \mapsto \mathbb{R}_+$ is continuous.

$(-)$ If $J_{-,n}\log(J_{-,n})/n \to 0$ and $J_{-,n}/\log(n) \to \infty$, then

$$\begin{aligned}
\mathrm{var}_{\mathtt{QS},-}(J_{-,n}) &= \frac{J_{-,n}}{n}\mathscr{V}_{\mathtt{QS},-}\{1 + o_{\mathbb{P}}(1)\}, \\
\mathscr{V}_{\mathtt{QS},-} &= \frac{1}{P_-}\int_{x_l}^{\bar{x}} \sigma_-^2(x)w(x)dx, \\
\mathrm{Bias}_{\mathtt{QS},-}(J_{-,n}) &= \frac{1}{J_{-,n}^2}\mathscr{B}_{\mathtt{QS},-}\{1 + o_{\mathbb{P}}(1)\}, \\
\mathscr{B}_{\mathtt{QS},-} &= \frac{P_-^2}{12}\int_{x_l}^{\bar{x}} \left(\frac{\mu_-^{(1)}(x)}{f(x)}\right)^2 w(x)dx,
\end{aligned}$$

where $P_- = \mathbb{P}[X_i < \bar{x}]$.

(+) If $J_{+,n} \log(J_{+,n})/n \to 0$ and $J_{+,n}/\log(n) \to \infty$, then

$$\begin{aligned}
\operatorname{var}_{\mathtt{QS},+}(J_{+,n}) &= \frac{J_{+,n}}{n}\mathscr{V}_{\mathtt{QS},+}\{1 + o_{\mathbb{P}}(1)\},\\
\mathscr{V}_{\mathtt{QS},+} &= \frac{1}{P_+}\int_{\bar{x}}^{x_u} \sigma_+^2(x)w(x)dx,\\
\operatorname{Bias}_{\mathtt{QS},+}(J_{+,n}) &= \frac{1}{J_{+,n}^2}\mathscr{B}_{\mathtt{QS},+}\{1 + o_{\mathbb{P}}(1)\},\\
\mathscr{B}_{\mathtt{QS},+} &= \frac{P_+^2}{12}\int_{\bar{x}}^{x_u}\left(\frac{\mu_+^{(1)}(x)}{f(x)}\right)^2 w(x)dx,
\end{aligned}$$

where $P_+ = \mathbb{P}[X_i \geq \bar{x}]$.

The conclusion in this theorem is similar to that of Theorem 1, but its proof is different because the estimators are constructed using a random partitioning scheme. The partitioning scheme used in the ES RD plots ($\mathcal{P}_{\mathrm{ES},-,n}$ and $\mathcal{P}_{\mathrm{ES},+,n}$) requires $J_{-,n} \to \infty$ and $J_{+,n} \to \infty$ but could lead to empty bins in finite samples (this possibility disappears asymptotically; see Lemma SA1 in the supplemental appendix). In contrast, the partitioning scheme underlying the QS RD plots ($\mathcal{P}_{\mathtt{QS},-,n}$ and $\mathcal{P}_{\mathtt{QS},+,n}$) guarantees roughly the same number of observations ($\approx N_-/J_{-,n}$ and $\approx N_+/J_{+,n}$) in each bin. The slightly stronger rate conditions $J_{-,n}/\log(n) \to \infty$ and $J_{+,n}/\log(n) \to \infty$ are imposed to ensure consistency of the sample quantiles functions at the appropriate rate; see Lemma SA2 in the supplemental appendix.

The main difference between the conclusions in Theorems 1 and 2 is that the fixed, leading constants in the variance and bias approximations are different. Importantly, the rates derived are the same in both theorems. The fixed constants are different because the partitioning schemes used are different in each case, but nonetheless all the ideas previously discussed for ES RD plots also apply directly to QS RD plots. Thus, in the remainder of this section, we only briefly summarize the main results for completeness.

4.2 Approximating the Underlying Regression Functions

Using the results above, and under the assumptions of Theorem 2, we obtain an asymptotic expansion of the IMSE for QS RD plots given by

$$\begin{aligned}
&\mathrm{IMSE}_{\mathtt{QS},-}(J_{-,n})\\
&= \int_{x_l}^{\bar{x}} \mathbb{E}\left[(\hat{\mu}_{\mathtt{QS},-}(x; J_{-,n}) - \mu_-(x))^2 \middle| \mathbf{X}_n\right] w(x)dx\\
&= \frac{J_{-,n}}{n}\mathscr{V}_{\mathtt{QS},-}\{1 + o_{\mathbb{P}}(1)\} + \frac{1}{J_{-,n}^2}\mathscr{B}_{\mathtt{QS},-}\{1 + o_{\mathbb{P}}(1)\}
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{IMSE}_{\mathtt{QS},+}(J_{+,n})\\
&= \int_{\bar{x}}^{x_u} \mathbb{E}\left[(\hat{\mu}_{\mathtt{QS},+}(x; J_{+,n}) - \mu_+(x))^2 \middle| \mathbf{X}_n\right] w(x)dx\\
&= \frac{J_{+,n}}{n}\mathscr{V}_{\mathtt{QS},+}\{1 + o_{\mathbb{P}}(1)\} + \frac{1}{J_{+,n}^2}\mathscr{B}_{\mathtt{QS},+}\{1 + o_{\mathbb{P}}(1)\},
\end{aligned}$$

which imply the following IMSE-optimal choices of QS partition sizes (i.e., number of bins constructed using ES quantiles of the running variable):

$$J_{\mathtt{QS}\text{-}\mu,-,n} = \left\lceil \left(\frac{2\mathscr{B}_{\mathtt{QS},-}}{\mathscr{V}_{\mathtt{QS},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$J_{\mathtt{QS}\text{-}\mu,+,n} = \left\lceil \left(\frac{2\mathscr{B}_{\mathtt{QS},+}}{\mathscr{V}_{\mathtt{QS},+}}\right)^{1/3} n^{1/3} \right\rceil. \tag{4}$$

4.3 Approximating the Underlying Variability of the Data

4.3.1 Weighted IMSE. Employing the same notation for weights introduced above for ES RD plots, we can construct analogous weighted IMSE objective functions for QS RD plots: $\mathrm{WIMSE}_{\mathtt{QS},-}(J_{-,n}) = \omega_{\mathscr{V},-}\operatorname{var}_{\mathtt{QS},-}(J_{-,n}) + \omega_{\mathscr{B},-}\operatorname{Bias}_{\mathtt{QS},-}(J_{-,n})$ and $\mathrm{WIMSE}_{\mathtt{QS},+}(J_{+,n}) = \omega_{\mathscr{V},+}\operatorname{var}_{\mathtt{QS},+}(J_{+,n}) + \omega_{\mathscr{B},+}\operatorname{Bias}_{\mathtt{QS},+}(J_{+,n})$. Employing the approximations derived in Theorem 2 and optimizing we obtain

$$J_{\mathtt{QS}\text{-}\omega,-,n} = \left\lceil \omega_-\left(\frac{2\mathscr{B}_{\mathtt{QS},-}}{\mathscr{V}_{\mathtt{QS},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$J_{\mathtt{QS}\text{-}\omega,+,n} = \left\lceil \omega_+\left(\frac{2\mathscr{B}_{\mathtt{QS},+}}{\mathscr{V}_{\mathtt{QS},+}}\right)^{1/3} n^{1/3} \right\rceil \tag{5}$$

with $\omega_- = (\omega_{\mathscr{B},-}/\omega_{\mathscr{V},-})^{1/3}$ and $\omega_+ = (\omega_{\mathscr{B},+}/\omega_{\mathscr{V},+})^{1/3}$. The discussion given above for ES RD plots applies here as well, replacing the subindex ES with QS as appropriate in the different expressions given previously.

4.3.2 Mimicking Variability. We employ the same logic outlined for the case of ES RD plots, but now using QS binning. That is, letting $\mathcal{V}_-$ and $\mathcal{V}_+$ denote a population measure of variability of the outcome variables for control and treatment units, respectively, we select the number of bins for each group so that the asymptotic variability of the QS-based local sample means is approximately equal to the overall variability of the data. Thus, we propose the following "optimal" choice of number of bins:

$$J_{\mathtt{QS}\text{-}\vartheta,-,n} = \left\lceil \frac{\mathcal{V}_-}{\mathscr{V}_{\mathtt{QS},-}}\frac{n}{\log(n)^2} \right\rceil$$

and

$$J_{\mathtt{QS}\text{-}\vartheta,+,n} = \left\lceil \frac{\mathcal{V}_+}{\mathscr{V}_{\mathtt{QS},+}}\frac{n}{\log(n)^2} \right\rceil, \tag{6}$$

which has the same structure as given in (3) but with the subindex ES replaced by QS.

5. DATA-DRIVEN IMPLEMENTATIONS

Employing some reference model, we could easily construct rule-of-thumb estimates of the unknown constants ($\mathscr{V}_{\mathrm{ES},-}$, $\mathscr{B}_{\mathrm{ES},-}$, $\mathscr{V}_{\mathrm{ES},+}$, $\mathscr{B}_{\mathrm{ES},+}$) and ($\mathscr{V}_{\mathtt{QS},-}$, $\mathscr{B}_{\mathtt{QS},-}$, $\mathscr{V}_{\mathtt{QS},+}$, $\mathscr{B}_{\mathtt{QS},+}$) featuring in the different optimal choices of number of bins for ES and QS RD plots. Such implementations would require a given choice of weighting function $w(x)$ in practice, but would otherwise be straightforward to derive and easy to implement in practice; for further discussion see, for example, Wand and Jones (1995) for kernel estimation and Ruppert, Wand, and Carroll (2009) for series and penalized estimation.

A potential drawback of rule-of-thumb estimates is that they are inconsistent whenever the reference model used is incorrect. Thus, we propose instead easy-to-implement consistent nonparametric estimators for the unknown constants entering the optimal choices of number of bins in Equations (1) through (6). In the supplemental appendix, we outline a general approach allowing for any user-chosen known weighting function $w(x)$, which needs to be set in advance. Here, we discuss in detail our recommended choice, $w(x) = f(x)$, which removes a density from the denominators of the unknown constants and leads to particularly simple and intuitive data-driven rules. As we discuss further below, all of our approaches are not only theoretically justified, but also simple, easy-to-interpret and often more robust than the usual nonparametric alternatives.

We estimate the unknown constants using ideas related to spacings estimators (see, e.g., Ghosh and Jammalamadaka 2001; Lewbel and Schennach 2007; Baryshnikov, Penrose, and Yurich 2009, and references therein) and series estimators (see, e.g., Newey 1997; Chen 2007; Ruppert, Wand, and Carroll 2009; Belloni et al. 2015 for reviews). Spacings estimators are closely related to nearest neighbor estimators with fixed neighbors (e.g., Abadie and Imbens 2006, 2010), and may be more robust than other nonparametric estimators such as kernel-based estimators because they do not require additional tuning parameter choices in their implementation. To describe these estimators, we need to introduce notation for order statistics and concomitants; see David and Nagaraja (1998, 2003) for more details. For a collection of continuous random variables $\{(Z_i, W_i) : i = 1, 2, \ldots, n\}$, we let $W_{(i)}$ be the ith-order statistic of W_i and $Z_{[i]}$ its corresponding concomitant. That is, $W_{(1)} < W_{(2)} < \cdots < W_{(n)}$ and $Z_{[i]}$ denotes the Z-value associated with $W_{(i)}$ for all $i = 1, 2, \ldots, n$.

Spacings estimators are useful because they exploit properties of order statistics and concomitants to approximate the unknown density and moments of the random variables nonparametrically. To see this, heuristically, recall that $U_i = F_W(W_i)$ with $\{U_i : 1 \leq i \leq n\}$ uniform $[0, 1]$ random variables, $F_W(\cdot)$ the c.d.f. of W_i and $f_W(\cdot)$ the p.d.f. of W_i. Then, for some $\breve{u}_i \in [U_{(i-1)}, U_{(i)}]$ a Taylor series expansion gives

$$\begin{aligned} W_{(i)} - W_{(i-1)} &= F_W^{-1}(U_{(i)}) - F_W^{-1}(U_{(i-1)}) \\ &= \frac{U_{(i)} - U_{(i-1)}}{f_W(F_W^{-1}(\breve{u}_i))} \approx \frac{1}{n f_W(F_W^{-1}(\bar{U}_{(i)}))}, \end{aligned}$$

where $\bar{U}_{(i)} = (U_{(i-1)} + U_{(i)})/2$, and because $|U_{(i)} - U_{(i-1)} - 1/n| \approx 0$. Thus, heuristically, an average of a smooth function of the spacings statistic, $W_{(i)} - W_{(i-1)}$, will converge to the expectation of this function inversely weighted by the unknown density $f_W(\cdot)$, up to a scaling factor and some additional (technical) constants that may arise in the derivation. Similar arguments using concomitants and order statistics give $\mathbb{E}[(Z_{[i]} - Z_{[i-1]})^2 | W_{(1)}, \ldots, W_{(n)}] \approx \sigma^2(W_{(i)}) + \sigma^2(W_{(i-1)})$ with $\sigma^2(W_i) = \mathbb{V}[Z_i | W_i]$. Lemma SA3 in the supplemental appendix formalizes these results, which we use in the sequel to construct simple, nonparametric estimators of the unknown constants entering the number of bins selectors proposed in this article.

We also employ series (polynomial) nonparametric approximations to estimate $\mu_-^{(1)}(x)$ and $\mu_+^{(1)}(x)$, and $\sigma_-^2(x)$ and $\sigma_+^2(x)$ in some cases, trying to mimic as closely as possible current empirical practices—these polynomial approximations are already available as part of the RD plots.

5.1 ES RD Plots

Taking $w(x) = f(x)$, with $f(x)$ unknown, leads to the following simplified constants in Theorem 1:

$$\begin{aligned} \mathscr{V}_{\mathrm{ES},-} &= \frac{1}{\bar{x} - x_l} \int_{x_l}^{\bar{x}} \sigma_-^2(x) dx, \\ \mathscr{B}_{\mathrm{ES},-} &= \frac{(\bar{x} - x_l)^2}{12} \mathbb{E}[\mathbb{1}(X_i < \bar{x})(\mu_-^{(1)}(X_i))^2], \\ \mathscr{V}_{\mathrm{ES},+} &= \frac{1}{x_u - \bar{x}} \int_{\bar{x}}^{x_u} \sigma_+^2(x) dx \\ \mathscr{B}_{\mathrm{ES},+} &= \frac{(x_u - \bar{x})^2}{12} \mathbb{E}[\mathbb{1}(X_i \geq \bar{x})(\mu_+^{(1)}(X_i))^2], \end{aligned}$$

which feature in the number of bins selectors discussed above for the ES RD plots.

Letting $\{(Y_{-,i}, X_{-,i}) : i = 1, 2, \ldots, N_-\}$ and $\{(Y_{+,i}, X_{+,i}) : i = 1, 2, \ldots, N_+\}$ be the subsamples of control $(X_i < \bar{x})$ and treatment $(X_i \geq \bar{x})$ units, respectively, we propose the following estimators:

$$\hat{\mathscr{V}}_{\mathrm{ES},-} = \frac{1}{\bar{x} - x_l} \frac{1}{2} \sum_{i=2}^{N_-} (X_{-,(i)} - X_{-,(i-1)})(Y_{-,[i]} - Y_{-,[i-1]})^2 \tag{7}$$

$$\hat{\mathscr{B}}_{\mathrm{ES},-} = \frac{(\bar{x} - x_l)^2}{12n} \sum_{i=1}^{n} \mathbb{1}(X_i < \bar{x}) \left(\hat{\mu}_{-,k}^{(1)}(X_i)\right)^2, \tag{8}$$

and

$$\hat{\mathscr{V}}_{\mathrm{ES},+} = \frac{1}{x_u - \bar{x}} \frac{1}{2} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)})(Y_{+,[i]} - Y_{+,[i-1]})^2 \tag{9}$$

$$\hat{\mathscr{B}}_{\mathrm{ES},+} = \frac{(x_u - \bar{x})^2}{12n} \sum_{i=1}^{n} \mathbb{1}(X_i \geq \bar{x}) \left(\hat{\mu}_{+,k}^{(1)}(X_i)\right)^2, \tag{10}$$

with

$$\begin{aligned} \bar{X}_{-,(i)} &= \frac{X_{-,(i)} + X_{-,(i-1)}}{2}, \qquad i = 2, 3, \ldots, N_-, \\ \hat{\mu}_{-,k}^{(1)}(x) &= \mathbf{r}_k^{(1)}(x)' \hat{\boldsymbol{\beta}}_{-,k}, \\ \bar{X}_{+,(i)} &= \frac{X_{+,(i)} + X_{+,(i-1)}}{2}, \qquad i = 2, 3, \ldots, N_+, \\ \hat{\mu}_{+,k}^{(1)}(x) &= \mathbf{r}_k^{(1)}(x)' \hat{\boldsymbol{\beta}}_{+,k}, \end{aligned}$$

and $\mathbf{r}_k^{(1)}(x) = \partial \mathbf{r}_k(x)/\partial x = (0, 1, 2x, 3x^2, \ldots, kx^{k-1})'$. These estimators are particularly well suited for our purposes because they (i) avoid explicit estimation of the density $f(x)$ appearing in the denominators and (ii) do not require specific choices of tuning parameters (e.g., bandwidths in kernel-based estimation). For these reasons, and given their simple implementation, we recommend employing these spacings-based estimators whenever possible.

Thus, our proposed data-driven selectors for ES RD plots take the form

$$\hat{J}_{\text{ES-}\mu,-,n} = \left\lceil \left(\frac{2\hat{\mathscr{B}}_{\text{ES},-}}{\hat{\mathscr{V}}_{\text{ES},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$\hat{J}_{\text{ES-}\mu,+,n} = \left\lceil \left(\frac{2\hat{\mathscr{B}}_{\text{ES},+}}{\hat{\mathscr{V}}_{\text{ES},+}}\right)^{1/3} n^{1/3} \right\rceil, \tag{11}$$

$$\hat{J}_{\text{ES-}\omega,-,n} = \left\lceil \omega_- \left(\frac{2\hat{\mathscr{B}}_{\text{ES},-}}{\hat{\mathscr{V}}_{\text{ES},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and (12)

$$\hat{J}_{\text{ES-}\omega,+,n} = \left\lceil \omega_+ \left(\frac{2\hat{\mathscr{B}}_{\text{ES},+}}{\hat{\mathscr{V}}_{\text{ES},+}}\right)^{1/3} n^{1/3} \right\rceil, \tag{13}$$

$$\hat{J}_{\text{ES-}\vartheta,-,n} = \left\lceil \frac{\hat{\mathcal{V}}_-}{\hat{\mathscr{V}}_{\text{ES},-}} \frac{n}{\log(n)^2} \right\rceil$$

and

$$\hat{J}_{\text{ES-}\vartheta,+,n} = \left\lceil \frac{\hat{\mathcal{V}}_+}{\hat{\mathscr{V}}_{\text{ES},+}} \frac{n}{\log(n)^2} \right\rceil, \tag{14}$$

using the estimators in (7)–(10), and where $\hat{\mathcal{V}}_-$ and $\hat{\mathcal{V}}_+$ are consistent estimators of their population counterparts $\mathcal{V}_-$ and $\mathcal{V}_+$. The following theorem shows that, when the polynomial fits are viewed as nonparametric approximations with $k = k_n \to \infty$, the different number of bins selectors are nonparametric consistent.

Theorem 3. Suppose Assumption 1 holds with $S \geq 5$, and $Y_i(0)$ and $Y_i(1)$ are continuously distributed. If $k_n^7/n \to 0$ and $k_n \to \infty$, then

$$\frac{\hat{J}_{\text{ES-}\omega,-,n}}{J_{\text{ES-}\omega,-,n}} \to_{\mathbb{P}} 1, \quad \frac{\hat{J}_{\text{ES-}\vartheta,-,n}}{J_{\text{ES-}\vartheta,-,n}} \to_{\mathbb{P}} 1, \quad \frac{\hat{J}_{\text{ES-}\omega,+,n}}{J_{\text{ES-}\omega,+,n}} \to_{\mathbb{P}} 1, \quad \frac{\hat{J}_{\text{ES-}\vartheta,+,n}}{J_{\text{ES-}\vartheta,+,n}} \to_{\mathbb{P}} 1,$$

provided that $\hat{\mathcal{V}}_- \to_{\mathbb{P}} \mathcal{V}_-$ and $\hat{\mathcal{V}}_+ \to_{\mathbb{P}} \mathcal{V}_+$.

This theorem gives formal justification for employing any of the data-driven selectors for the number of bins introduced in this article. (Recall that $\hat{J}_{\text{ES-}\mu,-,n} = \hat{J}_{\text{ES-}\omega,-,n}$ and $\hat{J}_{\text{ES-}\mu,+,n} = \hat{J}_{\text{ES-}\omega,+,n}$ when equal weights are used in the WIMSE.) In the supplemental appendix, we also discuss the case where a given weighting scheme $w(x)$ is provided in advance, and show consistency of the associated number of bins selectors; those results cover the case $w(x) = 1$, for example.

Remark 1 (Discontinuous Outcomes). When $Y_i(0)$ and $Y_i(1)$ are not continuously distributed, the concomitant-based estimation method becomes invalid. In this case, we need to employ other more standard nonparametric techniques. For example, assuming that $\mathbb{E}[Y_i(t)^2|X_i = x]$, $t = 0, 1$, are twice continuously differentiable, we can use the following estimators: for $k \in \mathbb{Z}_+$ and $p \in \mathbb{Z}_{++}$,

$$\begin{aligned}
\check{\mathscr{V}}_{\text{ES},-} &= \frac{1}{\bar{x} - x_l} \sum_{i=2}^{N_-} (X_{-,(i)} - X_{-,(i-1)}) \hat{\sigma}^2_{-,k}(\bar{X}_{-,(i)}),\\
\hat{\sigma}^2_{-,k}(x) &= \hat{\mu}_{-,k,2}(x) - (\hat{\mu}_{-,k,1}(x))^2,\\
\check{\mathscr{V}}_{\text{ES},+} &= \frac{1}{x_u - \bar{x}} \sum_{i=2}^{N_+} (X_{+,(i)} - X_{+,(i-1)}) \hat{\sigma}^2_{+,k}(\bar{X}_{+,(i)}),\\
\hat{\sigma}^2_{+,k}(x) &= \hat{\mu}_{+,k,2}(x) - (\hat{\mu}_{+,k,1}(x))^2,
\end{aligned}$$

where

$$\begin{aligned}
\hat{\mu}_{-,k,p}(x) &= \mathbf{r}_k(x)'\hat{\boldsymbol{\beta}}_{-,k,p},\\
\hat{\boldsymbol{\beta}}_{-,k,p} &= \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} \mathbb{1}(X_i < \bar{x})(Y_i^p - \mathbf{r}_k(X_i)'\boldsymbol{\beta})^2,\\
\hat{\mu}_{+,k,p}(x) &= \mathbf{r}_k(x)'\hat{\boldsymbol{\beta}}_{+,k,p},\\
\hat{\boldsymbol{\beta}}_{+,k,p} &= \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} \mathbb{1}(X_i \geq \bar{x})(Y_i^p - \mathbf{r}_k(X_i)'\boldsymbol{\beta})^2,
\end{aligned}$$

$\hat{\mu}_{-,k}(x) = \hat{\mu}_{-,k,1}(x)$ and $\hat{\mu}_{+,k}(x) = \hat{\mu}_{+,k,1}(x)$ with our notation. We show in the appendix that the resulting partition-size selectors using the above estimators,

$$\check{J}_{\text{ES-}\mu,-,n} = \left\lceil \left(\frac{2\hat{\mathscr{B}}_{\text{ES},-}}{\check{\mathscr{V}}_{\text{ES},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$\check{J}_{\text{ES-}\mu,+,n} = \left\lceil \left(\frac{2\hat{\mathscr{B}}_{\text{ES},+}}{\check{\mathscr{V}}_{\text{ES},+}}\right)^{1/3} n^{1/3} \right\rceil, \tag{14}$$

$$\check{J}_{\text{ES-}\omega,-,n} = \left\lceil \omega_- \left(\frac{2\hat{\mathscr{B}}_{\text{ES},-}}{\check{\mathscr{V}}_{\text{ES},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$\check{J}_{\text{ES-}\omega,+,n} = \left\lceil \omega_+ \left(\frac{2\hat{\mathscr{B}}_{\text{ES},+}}{\check{\mathscr{V}}_{\text{ES},+}}\right)^{1/3} n^{1/3} \right\rceil, \tag{15}$$

$$\check{J}_{\text{ES-}\vartheta,-,n} = \left\lceil \frac{\hat{\mathcal{V}}_-}{\check{\mathscr{V}}_{\text{ES},-}} \frac{n}{\log(n)^2} \right\rceil$$

and

$$\check{J}_{\text{ES-}\vartheta,+,n} = \left\lceil \frac{\hat{\mathcal{V}}_+}{\check{\mathscr{V}}_{\text{ES},+}} \frac{n}{\log(n)^2} \right\rceil, \tag{16}$$

are also consistent in the sense of Theorem 3, under the conditions imposed in that theorem.

5.2 QS RD Plots

Paralleling the discussion above for ES RD plots, we propose consistent estimators for $J_{\texttt{QS-}\mu,-,n}$, $J_{\texttt{QS-}\mu,+,n}$, $J_{\texttt{QS-}\omega,-,n}$, $J_{\texttt{QS-}\omega,+,n}$, $J_{\texttt{QS-}\vartheta,-,n}$, and $J_{\texttt{QS-}\vartheta,+,n}$, when $w(x) = f(x)$ with $f(x)$ unknown. In this case, the target constants take the form

$$\begin{aligned}
\mathscr{V}_{\texttt{QS},-} &= \frac{1}{P_-} \mathbb{E}[\mathbb{1}(X_i < \bar{x})\sigma^2_-(X_i)],\\
\mathscr{B}_{\texttt{QS},-} &= \frac{P_-^2}{12} \int_{x_l}^{\bar{x}} \frac{1}{f(x)} \left(\mu_-^{(1)}(x)\right)^2 dx,
\end{aligned}$$

$$\mathscr{V}_{\texttt{QS},+} = \frac{1}{P_+}\mathbb{E}[\mathbb{1}(X_i \geq \bar{x})\sigma_+^2(X_i)],$$

$$\mathscr{B}_{\texttt{QS},+} = \frac{P_+^2}{12}\int_{\bar{x}}^{x_u} \frac{1}{f(x)}\left(\mu_+^{(1)}(x)\right)^2 dx.$$

Our preferred selectors employ spacings estimators, which are simple and easy-to-implement but require continuous outcomes. See Remark 2 for the case of noncontinuous outcomes. The estimators of the optimal selectors for QS partition size are based on the following estimators:

$$\hat{\mathscr{V}}_{\texttt{QS},-} = \frac{1}{2N_-}\sum_{i=2}^{N_-}(Y_{-,[i]} - Y_{-,[i-1]})^2 \tag{17}$$

$$\hat{\mathscr{B}}_{\texttt{QS},-} = \frac{N_-^2}{24n}\sum_{i=2}^{N_-}(X_{-,(i)} - X_{-,(i-1)})^2\left(\hat{\mu}_{-,k}^{(1)}(\bar{X}_{-,(i)})\right)^2, \tag{18}$$

and

$$\hat{\mathscr{V}}_{\texttt{QS},+} = \frac{1}{2N_+}\sum_{i=2}^{N_+}(Y_{+,[i]} - Y_{+,[i-1]})^2 \tag{19}$$

$$\hat{\mathscr{B}}_{\texttt{QS},+} = \frac{N_+^2}{24n}\sum_{i=2}^{N_+}(X_{+,(i)} - X_{+,(i-1)})^2\left(\hat{\mu}_{+,k}^{(1)}(\bar{X}_{+,(i)})\right)^2, \tag{20}$$

using the notation introduced above.

Therefore, in the QS partitions case, our data-driven selectors take the form

$$\hat{J}_{\texttt{QS-}\mu,-,n} = \left\lceil \left(\frac{2\hat{\mathscr{B}}_{\texttt{QS},-}}{\hat{\mathscr{V}}_{\texttt{QS},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$\hat{J}_{\texttt{QS-}\mu,+,n} = \left\lceil \left(\frac{2\hat{\mathscr{B}}_{\texttt{QS},+}}{\hat{\mathscr{V}}_{\texttt{QS},+}}\right)^{1/3} n^{1/3} \right\rceil, \tag{21}$$

$$\hat{J}_{\texttt{QS-}\omega,-,n} = \left\lceil \omega_- \left(\frac{2\hat{\mathscr{B}}_{\texttt{QS},-}}{\hat{\mathscr{V}}_{\texttt{QS},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$\hat{J}_{\texttt{QS-}\omega,+,n} = \left\lceil \omega_+ \left(\frac{2\hat{\mathscr{B}}_{\texttt{QS},+}}{\hat{\mathscr{V}}_{\texttt{QS},+}}\right)^{1/3} n^{1/3} \right\rceil, \tag{22}$$

$$\hat{J}_{\texttt{QS-}\vartheta,-,n} = \left\lceil \frac{\hat{\mathcal{V}}_-}{\hat{\mathscr{V}}_{\texttt{QS},-}} \frac{n}{\log(n)^2} \right\rceil$$

and

$$\hat{J}_{\texttt{QS-}\vartheta,+,n} = \left\lceil \frac{\hat{\mathcal{V}}_+}{\hat{\mathscr{V}}_{\texttt{QS},+}} \frac{n}{\log(n)^2} \right\rceil, \tag{23}$$

using the estimators in (17)–(20), and appropriate consistent estimators $\hat{\mathcal{V}}_-$ and $\hat{\mathcal{V}}_+$. As in the case of Theorem 3 for ES RD plots, the following theorem shows that these automatic partition-size selectors are nonparametric consistent if the polynomial fits are viewed as nonparametric approximations with $k = k_n \to \infty$.

Theorem 4. Suppose Assumption 1 holds with $S \geq 5$, and $Y_i(0)$ and $Y_i(1)$ are continuously distributed. If $k_n^7/n \to 0$ and $k_n \to \infty$, then

$$\frac{\hat{J}_{\texttt{QS-}\omega,-,n}}{J_{\texttt{QS-}\omega,-,n}} \to_{\mathbb{P}} 1, \quad \frac{\hat{J}_{\texttt{QS-}\vartheta,-,n}}{J_{\texttt{QS-}\vartheta,-,n}} \to_{\mathbb{P}} 1, \quad \frac{\hat{J}_{\texttt{QS-}\omega,+,n}}{J_{\texttt{QS-}\omega,+,n}} \to_{\mathbb{P}} 1,$$

$$\frac{\hat{J}_{\texttt{QS-}\vartheta,+,n}}{J_{\texttt{QS-}\vartheta,+,n}} \to_{\mathbb{P}} 1,$$

provided that $\hat{\mathcal{V}}_- \to_{\mathbb{P}} \mathcal{V}_-$ and $\hat{\mathcal{V}}_+ \to_{\mathbb{P}} \mathcal{V}_+$.

In the supplemental appendix, we also propose consistent estimators for a generic, known weighting scheme $w(x)$. These estimators are more flexible but also more complicated in general.

Remark 2 (Noncontinuous Outcomes). As mentioned in Remark 1, the concomitant-based estimation approach cannot be used when $Y_i(0)$ and $Y_i(1)$ are not continuously distributed. For the latter cases, alternatively, we can use the series polynomial estimation approach already introduced above. Assuming that $\mathbb{E}[Y_i(t)^2|X_i = x]$, $t = 0, 1$, are twice continuously differentiable, we may use the following estimators:

$$\check{\mathscr{V}}_{\texttt{QS},-} = \frac{1}{N_-}\sum_{i=1}^{n}\mathbb{1}(X_i < \bar{x})\hat{\sigma}_{-,k}^2(X_i)$$

and

$$\check{\mathscr{V}}_{\texttt{QS},+} = \frac{1}{N_+}\sum_{i=1}^{n}\mathbb{1}(X_i \geq \bar{x})\hat{\sigma}_{+,k}^2(X_i),$$

where $\hat{\sigma}_{-,k}^2(x)$ and $\hat{\sigma}_{+,k}^2(x)$ are the polynomial approximations discussed in Remark 1. The corresponding data-driven partition-size selectors in this case are

$$\check{J}_{\texttt{QS-}\mu,-,n} = \left\lceil \left(\frac{2\hat{\mathscr{B}}_{\texttt{QS},-}}{\check{\mathscr{V}}_{\texttt{QS},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$\check{J}_{\texttt{QS-}\mu,+,n} = \left\lceil \left(\frac{2\hat{\mathscr{B}}_{\texttt{QS},+}}{\check{\mathscr{V}}_{\texttt{QS},+}}\right)^{1/3} n^{1/3} \right\rceil, \tag{24}$$

$$\check{J}_{\texttt{QS-}\omega,-,n} = \left\lceil \omega_- \left(\frac{2\hat{\mathscr{B}}_{\texttt{QS},-}}{\check{\mathscr{V}}_{\texttt{QS},-}}\right)^{1/3} n^{1/3} \right\rceil$$

and

$$\check{J}_{\texttt{QS-}\omega,+,n} = \left\lceil \omega_+ \left(\frac{2\hat{\mathscr{B}}_{\texttt{QS},+}}{\check{\mathscr{V}}_{\texttt{QS},+}}\right)^{1/3} n^{1/3} \right\rceil, \tag{25}$$

$$\check{J}_{\texttt{QS-}\vartheta,-,n} = \left\lceil \frac{\hat{\mathcal{V}}_-}{\check{\mathscr{V}}_{\texttt{QS},-}} \frac{n}{\log(n)^2} \right\rceil$$

and

$$\check{J}_{\texttt{QS-}\vartheta,+,n} = \left\lceil \frac{\hat{\mathcal{V}}_+}{\check{\mathscr{V}}_{\texttt{QS},+}} \frac{n}{\log(n)^2} \right\rceil, \tag{26}$$

which, as we show in the appendix, are also consistent in the sense of Theorem 4, provided the conditions in that theorem hold.

6. NUMERICAL RESULTS

This section reports numerical evidence on the performance of our proposed methods employing real data from several empirical applications, and simulated data from a Monte Carlo experiment. We also compare numerically the two partitioning schemes studied in this article, ES and QS, in terms of their asymptotic IMSE. All the results in this section are obtained using the R and STATA software packages described in Calonico, Cattaneo, and Titiunik (2014a, 2015).

6.1 Empirical Applications

We illustrate our methods using data from several RD empirical applications. To conserve space, we report here only results using the data from Lee (2008), already mentioned in the Introduction. The supplemental appendix includes the other empirical applications, which employ data from (i) U.S. Senate elections (see Cattaneo, Frandsen, and Titiunik 2015 for details), (ii) Progresa/Oportunidades anti-poverty conditional cash transfer program (see Calonico, Cattaneo, and Titiunik 2014c, sec. S.4 for details), and (iii) Head Start funding program (see Ludwig and Miller 2007 for details). As mentioned above, Lee (2008) studied the incumbency advantage in U.S. House elections; the forcing variable is the margin of victory of the Democratic party in a given U.S. House election, the threshold is $\bar{x} = 0$, and the outcome variable is the Democratic vote share in the following U.S. House election, which occurs 2 years later. The unit of observation is the U.S. House district. All U.S. House elections between 1948 and 2008 are included, with the exception of years when district boundaries change; the dataset we employ has a total of $n = 6558$ complete district-year observations.

The main goal of this empirical application is to show how our selectors perform when applied to a realistic dataset. It is difficult to compare our results to alternatives or benchmarks because we are not aware of any other selectors available in the literature to construct RD plots. The standard practice appears to be that each researcher explores the data and selects the number of bins in an ad hoc, nonsystematic way. Thus, we focus on discussing the graphical properties of the resulting RD plots when our methods are employed.

Figures 2 and 3 collect six graphs in two rows. Each graph depicts the global fourth degree polynomial fits $\hat{\mu}_{-,4}(x)$ and $\hat{\mu}_{+,4}(x)$ as a solid blue line. Graphs (a) and (d) are the scatterplot of the raw data, which we include here for visual comparison. The remaining four graphs in each figure are RD plots constructed using ES bins (Figure 2) or QS bins (Figure 3) with data-driven choices employing either spacings estimators or series estimators. In each of these four graphs, the black dots correspond to the sample mean within each bin when the number of bins is selected to mimic the variability of the data (graphs (b) and (e)), while the black triangles correspond to the sample mean within each bin when the number of bins is selected to minimize the IMSE of the underlying regression function estimator (graphs (c) and (f)).

When analyzing each figure row-wise, we may see graphically how the variability of the data is summarized by each method. In particular, the scatterplots (graphs (a) and (d) in each figure) give a graphical representation of the raw data and are therefore extremely variable and arguably uninformative. Next, the mimicking variance RD plots (graphs (b) and (e) in each figure) reduce variability substantially when plotting the binned sample means of the raw data, but they are still able to provide a disciplined graphical representation of the overall variability of the RD design. Finally, the IMSE-optimal RD plots (graphs (c) and (f) in each figure) deliver "smoother" local (disjoint) sample means essentially trying to trace out the underlying unknown conditional expectation functions.

To summarize, the data-driven selectors introduced in this article seem to perform very well in all the empirical applications we considered. Specifically, the data-driven IMSE-optimal spacings-based selectors ($\hat{J}_{\text{ES-}\mu,-,n}$, $\hat{J}_{\text{ES-}\mu,+,n}$) and series-based selectors ($\check{J}_{\text{ES-}\mu,-,n}$, $\check{J}_{\text{ES-}\mu,+,n}$) generate a collection of binned sample means "tracing out" the estimated regression function (we use the polynomial fit as benchmark), which provides visual evidence in favor of continuity of the conditional expectations. Furthermore, the data-driven spacings-based selectors ($\hat{J}_{\text{ES-}\vartheta,-,n}$, $\hat{J}_{\text{ES-}\vartheta,+,n}$) and series-based selectors ($\check{J}_{\text{ES-}\vartheta,-,n}$, $\check{J}_{\text{ES-}\vartheta,+,n}$) mimicking the underlying variability of the data generate a disciplined scatterplot with substantial more variability than the IMSE-optimal binned means case, but yet less variable than the raw data, which in this case provides a nice visual representation of the RD design. As shown in the supplemental appendix, very similar findings emerge from the other empirical applications mentioned previously.

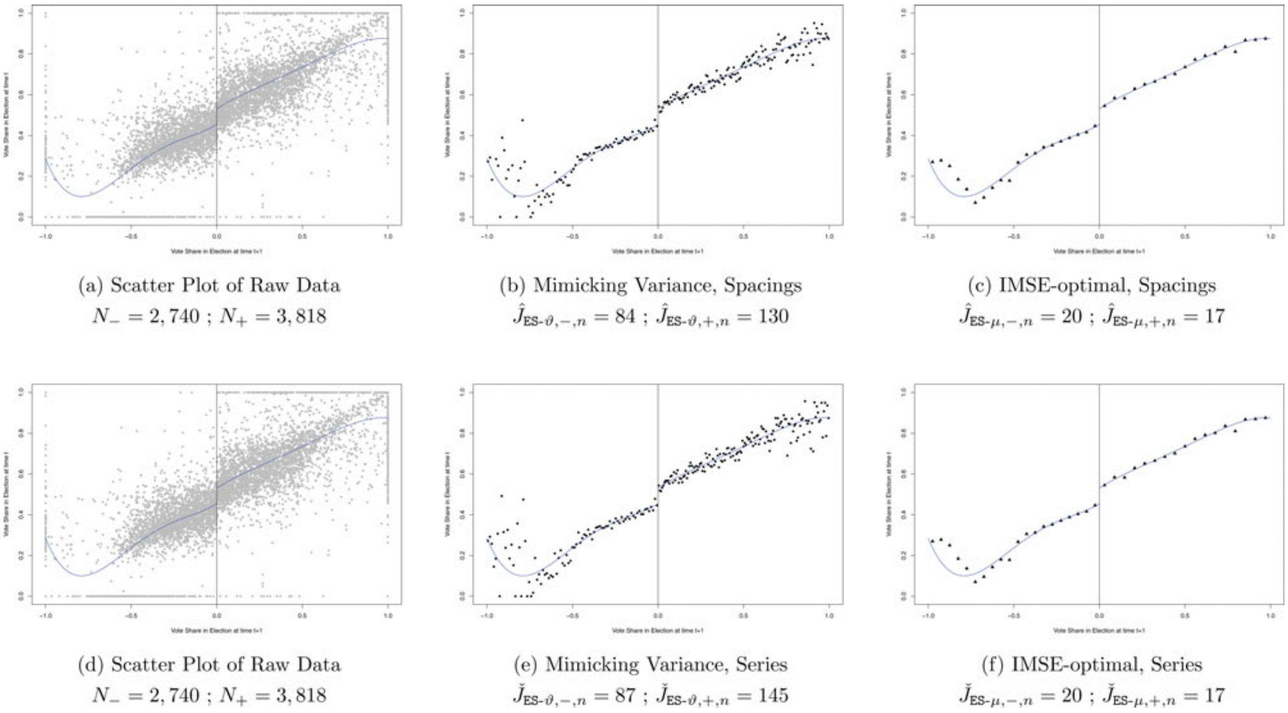
6.2 Simulated Data

We briefly report an example of the results from an extensive Monte Carlo experiment we conducted to study the finite-sample behavior of our proposed methods. The full simulation experiment considers 16 distinct data-generating processes, which vary in the distribution of the running variable, the form of the conditional variance, and the distribution of the unobserved error term in the regression function. To conserve space, here we only discuss the simplest case that assumes a uniform distribution for X_i and homoscedasticity, but the full set of results is reported in the supplemental appendix. We found the same qualitative results in all cases.

Specifically, we discuss the simulation model $\{(Y_i, X_i)' : i = 1, 2, \ldots, n\}$ iid with $Y_i = \mu(X_i) + \varepsilon_i$, $X_i \sim \text{Uniform}(-1, 1)$, $\varepsilon_i \sim \text{Normal}(0, 1)$, and where

$$\mu(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 \\ \quad \text{if} x < 0 \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 \\ \quad \text{if} x \geq 0 \end{cases}.$$

The functional form of $\mu(x)$ is obtained using the original data of Lee (2008). All details are given in the supplemental appendix. Table 1 reports the simulation results for this simple model. This table includes results for both ES and QS RD plots organized in two distinct panels. Panel A focuses attention on the IMSE of different partitioning schemes in finite samples, as well as the performance of the associated IMSE-optimal data-driven selectors. All IMSEs are normalized relative to the IMSE evaluated at the optimal partition-size choice to avoid any scaling issue. Panel B reports several features of the empirical (finite-sample) distribution of the different data-driven number of bins selectors



(a) Scatter Plot of Raw Data $N_- = 2,740$; $N_+ = 3,818$

(b) Mimicking Variance, Spacings $\hat{J}_{\texttt{ES-}\vartheta,-,n} = 84$; $\hat{J}_{\texttt{ES-}\vartheta,+,n} = 130$

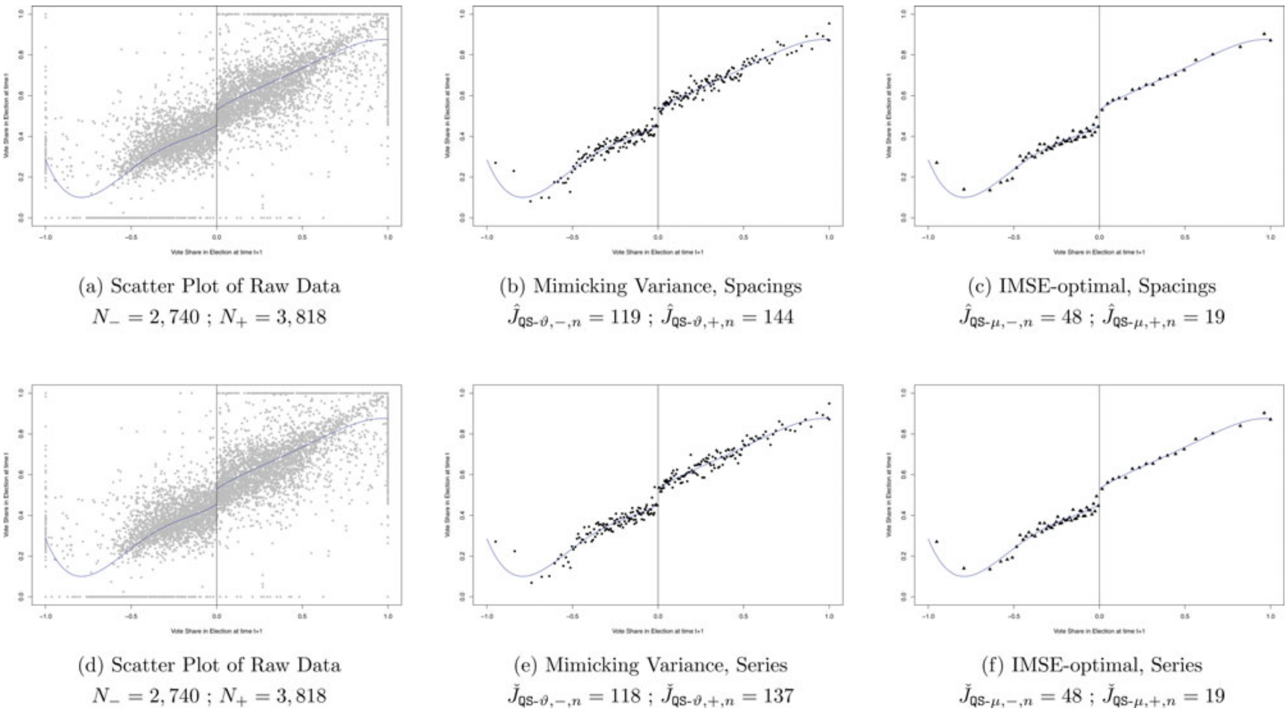
(c) IMSE-optimal, Spacings $\hat{J}_{\texttt{ES-}\mu,-,n} = 20$; $\hat{J}_{\texttt{ES-}\mu,+,n} = 17$

(d) Scatter Plot of Raw Data $N_- = 2,740$; $N_+ = 3,818$

(e) Mimicking Variance, Series $\check{J}_{\texttt{ES-}\vartheta,-,n} = 87$; $\check{J}_{\texttt{ES-}\vartheta,+,n} = 145$

(f) IMSE-optimal, Series $\check{J}_{\texttt{ES-}\mu,-,n} = 20$; $\check{J}_{\texttt{ES-}\mu,+,n} = 17$

Figure 2. Scatterplot and automatic data-driven ES RD plots for U.S. House elections data. (a) Scatterplot of raw data $N_- = 2740$; $N_+ = 3818$ (b) Mimicking variance, spacings $\hat{J}_{\text{ES-}\vartheta,-,n} = 84$; $\hat{J}_{\text{ES-}\vartheta,+,n} = 130$ (c) IMSE-optimal, spacings $\hat{J}_{\text{ES-}\mu,-,n} = 20$; $\hat{J}_{\text{ES-}\mu,+,n} = 17$ (d) Scatterplot of raw data $N_- = 2740$; $N_+ = 3818$ (e) Mimicking variance, series $\check{J}_{\text{ES-}\vartheta,-,n} = 87$; $\check{J}_{\text{ES-}\vartheta,+,n} = 145$ (f) IMSE-optimal, series $\check{J}_{\text{ES-}\mu,-,n} = 20$; $\check{J}_{\text{ES-}\mu,+,n} = 17$. Notes: (i) sample size is $n = 6558$; (ii) N_- and N_+ denote the sample sizes for control and treatment units, respectively; (iii) solid blue lines depict fourth-order polynomial fits using control and treated units separately.



(a) Scatter Plot of Raw Data $N_- = 2,740$; $N_+ = 3,818$

(b) Mimicking Variance, Spacings $\hat{J}_{\texttt{QS-}\vartheta,-,n} = 119$; $\hat{J}_{\texttt{QS-}\vartheta,+,n} = 144$

(c) IMSE-optimal, Spacings $\hat{J}_{\texttt{QS-}\mu,-,n} = 48$; $\hat{J}_{\texttt{QS-}\mu,+,n} = 19$

(d) Scatter Plot of Raw Data $N_- = 2,740$; $N_+ = 3,818$

(e) Mimicking Variance, Series $\check{J}_{\texttt{QS-}\vartheta,-,n} = 118$; $\check{J}_{\texttt{QS-}\vartheta,+,n} = 137$

(f) IMSE-optimal, Series $\check{J}_{\texttt{QS-}\mu,-,n} = 48$; $\check{J}_{\texttt{QS-}\mu,+,n} = 19$

Figure 3. Scatterplot and automatic data-driven QS RD plots for U.S. House elections data. (a) Scatterplot of raw data $N_- = 2740$; $N_+ = 3818$ (b) Mimicking variance, spacings $\hat{J}_{\texttt{QS-}\vartheta,-,n} = 119$; $\hat{J}_{\texttt{QS-}\vartheta,+,n} = 144$ (c) IMSE-optimal, spacings $\hat{J}_{\texttt{QS-}\mu,-,n} = 48$; $\hat{J}_{\texttt{QS-}\mu,+,n} = 19$ (d) Scatterplot of raw data $N_- = 2740$; $N_+ = 3818$ (e) Mimicking variance, series $\check{J}_{\texttt{QS-}\vartheta,-,n} = 118$; $\check{J}_{\texttt{QS-}\vartheta,+,n} = 137$ (f) IMSE-optimal, series $\check{J}_{\texttt{QS-}\mu,-,n} = 48$; $\check{J}_{\texttt{QS-}\mu,+,n} = 19$. Notes: (i) sample size is $n = 6558$; (ii) N_- and N_+ denote the sample sizes for control and treatment units, respectively; (iii) solid blue lines depict fourth-order polynomial fits using control and treated units separately.

Table 1. Simulations results

Panel A: IMSE for grid of number of bins and estimated choices

$J_{-,n}$	$\frac{\text{IMSE}_{\text{ES},-}(J_{-,n})}{\text{IMSE}^*_{\text{ES},-}}$	$J_{+,n}$	$\frac{\text{IMSE}_{\text{ES},+}(J_{+,n})}{\text{IMSE}^*_{\text{ES},+}}$	$J_{-,n}$	$\frac{\text{IMSE}_{\texttt{QS},-}(J_{-,n})}{\text{IMSE}^*_{\texttt{QS},-}}$	$J_{+,n}$	$\frac{\text{IMSE}_{\texttt{QS},+}(J_{+,n})}{\text{IMSE}^*_{\texttt{QS},+}}$
20	1.047	11	1.148	20	1.047	11	1.148
21	1.027	12	1.081	21	1.027	12	1.081
22	1.013	13	1.039	22	1.013	13	1.039
23	1.005	14	1.014	23	1.005	14	1.014
24	1.000	15	1.002	24	1.000	15	1.002
25	1.000	16	1.000	25	1.000	16	1.000
26	1.003	17	1.006	26	1.003	17	1.006
27	1.008	18	1.017	27	1.008	18	1.017
28	1.016	19	1.033	28	1.016	19	1.033
29	1.025	20	1.053	29	1.025	20	1.053
30	1.036	21	1.076	30	1.036	21	1.076
$\hat{J}_{\text{ES-}\mu,-,n}$	1.033	$\hat{J}_{\text{ES-}\mu,+,n}$	0.9435	$\hat{J}_{\texttt{QS-}\mu,-,n}$	1.072	$\hat{J}_{\texttt{QS-}\mu,+,n}$	0.9351
$\check{J}_{\text{ES-}\mu,-,n}$	1.034	$\check{J}_{\text{ES-}\mu,+,n}$	0.9428	$\check{J}_{\texttt{QS-}\mu,-,n}$	1.073	$\check{J}_{\texttt{QS-}\mu,+,n}$	0.9347

Panel B: Summary statistics for the estimated number of bins

Pop. par.		Min.	1st qu.	Median	Mean	3rd qu.	Max.	Std. dev.
$J_{\text{ES-}\mu,-,n} = 25$	$\hat{J}_{\text{ES-}\mu,-,n}$	22	25	26	25.95	27	29	0.93
	$\check{J}_{\text{ES-}\mu,-,n}$	23	25	26	25.93	26	29	0.87
$J_{\text{ES-}\vartheta,-,n} = 118$	$\hat{J}_{\text{ES-}\vartheta,-,n}$	105	116	120	119.6	123	139	5.05
	$\check{J}_{\text{ES-}\vartheta,-,n}$	110	117	119	119.3	121	131	2.72
$J_{\text{ES-}\mu,+,n} = 16$	$\hat{J}_{\text{ES-}\mu,+,n}$	14	15	15	15.34	16	17	0.57
	$\check{J}_{\text{ES-}\mu,+,n}$	14	15	15	15.34	16	17	0.55
$J_{\text{ES-}\vartheta,+,n} = 116$	$\hat{J}_{\text{ES-}\vartheta,+,n}$	103	113	117	116.7	120	139	4.71
	$\check{J}_{\text{ES-}\vartheta,+,n}$	107	115	117	116.7	118	128	2.65
$J_{\texttt{QS-}\mu,-,n} = 25$	$\hat{J}_{\texttt{QS-}\mu,-,n}$	23	26	27	26.91	27	30	0.92
	$\check{J}_{\texttt{QS-}\mu,-,n}$	23	26	27	26.89	27	30	0.90
$J_{\texttt{QS-}\vartheta,-,n} = 118$	$\hat{J}_{\texttt{QS-}\vartheta,-,n}$	108	117	120	119.6	122	134	3.66
	$\check{J}_{\texttt{QS-}\vartheta,-,n}$	110	117	119	119.3	121	131	2.71
$J_{\texttt{QS-}\mu,+,n} = 16$	$\hat{J}_{\texttt{QS-}\mu,+,n}$	14	15	15	15.21	15	17	0.51
	$\check{J}_{\texttt{QS-}\mu,+,n}$	14	15	15	15.21	15	17	0.50
$J_{\texttt{QS-}\vartheta,+,n} = 116$	$\hat{J}_{\texttt{QS-}\vartheta,+,n}$	106	114	117	116.6	119	130	3.50
	$\check{J}_{\texttt{QS-}\vartheta,+,n}$	107	115	117	116.7	118	128	2.65

NOTES: (i) Population quantities: $J_{\text{ES-}\mu,\cdot,n}$ = IMSE-optimal partition size for ES RD plot (Equation (1)). $J_{\text{ES-}\vartheta,\cdot,n}$ = Mimicking variance partition size for ES RD plot (Equation (3)). $J_{\texttt{QS-}\mu,\cdot,n}$ = IMSE-optimal partition size for QS RD plot (Equation (4)). $J_{\texttt{QS-}\vartheta,\cdot,n}$ = Mimicking variance partition size for QS RD plot (Equation (6)). $\text{IMSE}^*_{\text{ES},\cdot} = \text{IMSE}_{\text{ES},\cdot}(J_{\text{ES-}\mu,\cdot,n})$ = ES IMSE function evaluated at optimal choice. $\text{IMSE}^*_{\texttt{QS},\cdot} = \text{IMSE}_{\texttt{QS},\cdot}(J_{\texttt{QS-}\mu,\cdot,n})$ = QS IMSE function evaluated at optimal choice.(ii) Estimators: $\hat{J}_{\text{ES-}\mu,\cdot,n}$ = spacings estimator of $J_{\text{ES-}\mu,\cdot,n}$ (Equation (11)); $\check{J}_{\text{ES-}\mu,\cdot,n}$ = polynomial estimator of $J_{\text{ES-}\mu,\cdot,n}$ (Equation (14)). $\hat{J}_{\text{ES-}\vartheta,\cdot,n}$ = spacings estimator of $J_{\text{ES-}\vartheta,\cdot,n}$ (Equation (13)); $\check{J}_{\text{ES-}\vartheta,\cdot,n}$ = polynomial estimator of $J_{\text{ES-}\vartheta,\cdot,n}$ (Equation (16)). $\hat{J}_{\texttt{QS-}\mu,\cdot,n}$ = spacings estimator of $J_{\texttt{QS-}\mu,\cdot,n}$ (Equation (21)); $\check{J}_{\texttt{QS-}\mu,\cdot,n}$ = polynomial estimator of $J_{\texttt{QS-}\mu,\cdot,n}$ (Equation (24)). $\hat{J}_{\texttt{QS-}\vartheta,\cdot,n}$ = spacings estimator of $J_{\texttt{QS-}\vartheta,\cdot,n}$ (Equation (23)); $\check{J}_{\texttt{QS-}\vartheta,\cdot,n}$ = polynomial estimator of $J_{\texttt{QS-}\vartheta,\cdot,n}$ (Equation (26)).

introduced in this article: (i) spacings-based selectors for ES RD plots (Equations (11) and (13)), (ii) polynomial-based selectors for ES RD plots (Equations (14) and (16)), (iii) spacings-based selectors for QS RD plots (Equations (21) and (23)), and (iv) polynomial-based selectors for QS RD plots (Equations (24) and (26)). Our Monte Carlo experiment is designed to (i) capture the finite-sample performance of Theorems 1 and 2 that give an approximation to the IMSE (Panel A), and (ii) capture the finite-sample performance of Theorems 3 and 4 as well as the other consistency results discussed in the remarks above (Panel B).

The numerical results are very encouraging. First, in all cases the IMSE is minimized at the corresponding IMSE-optimal number of bins choice derived in this article, suggesting that the Theorems 1 and 2 do provide a good finite-sample approximation. For example, the first two columns in Panel A of Table 1 present the normalized IMSE of the binned sample means underlying the ES RD plot for the control group, which is minimized at $J_{-,n} = 25$ because for other numbers of bins in the grid the ratio of actual IMSE to the IMSE evaluated at the optimal number of bins proposed in this article is larger than one. Therefore, the theoretical IMSE-optimal number of bins coincide with the simulated IMSE-optimal number of bins.

Second, in all cases our proposed data-driven implementations of the number of bins selectors perform quite well, exhibiting a concentrated finite-sample distribution centered at the target population choice introduced in this article. For example, the first two rows in Panel B of Table 1 give summary statistics for the data-driven implementations of the population IMSE-optimal choice $J_{\text{ES-}\mu,-,n}$ (ES RD plot for the control group), when using either spacings estimators ($\hat{J}_{\text{ES-}\mu,-,n}$) or polynomial estimators ($\check{J}_{\text{ES-}\mu,-,n}$). In this case, the target quantity is $J_{\text{ES-}\mu,-,n} = 25$, as mentioned above, and both data-driven implementations are well centered (e.g., sample means are 25.95 and 25.93) and concentrated (e.g., standard deviation are 0.93 and 0.87). Similarly, the third and fourth rows of Panel B present the sampling behavior of the mimicking variance estimators $\hat{J}_{\text{ES-}\vartheta,-,n}$ and $\check{J}_{\text{ES-}\vartheta,-,n}$, which perform equally well.

In sum, our simulation results briefly discussed here (and available in the supplemental appendix in full) suggest that our proposed optimal data-driven tuning parameter choices for constructing RD plots perform well in finite samples.

6.3 Comparison of Partitioning Schemes

We proposed two alternative ways of constructing RD plots, one employing ES partitioning and the other employing QS partitioning. Our proposed selection rules for ES and QS partitioning coincide when the underlying distribution of X_i is uniform. In general, however, neither partitioning approach dominates the other in terms of asymptotic variability or IMSE. For example, the IMSE of ES and QS sample means will depend on the unknown density of the running variable, as well as the unknown regression and conditional variance functional forms. In the supplemental appendix, we derive the exact formulas for the optimal IMSE for both ES and QS partitioning, and show formally that neither dominates the other in general. We also employ the 16 simulation models mentioned before to compare the performance of the partition-size selectors for ES and QS RD plots: according to our numerical evidence, QS RD plots seem to perform better than ES RD plots in terms of IMSE when the density $f(x)$ is low in some regions of the support, although this conclusion depends in part on the other unknown features of the data-generating process.

7. EXTENSIONS

We discuss briefly two extensions that are practically relevant. (We thank a reviewer for suggesting that we address these issues.) The first extension discusses how our results apply to fuzzy RD designs, while the second focuses on how other covariates could be incorporated in the analysis.

7.1 Fuzzy RD Designs

In the so-called fuzzy RD design, treatment assignment and treatment status may be different for each unit (i.e., imperfect treatment compliance). The basic RD model introduced in Section 2 can be expanded to account for this possibility. Specifically, similarly to the potential outcomes $Y_i(0)$ and $Y_i(1)$ already introduced, define

$$\begin{aligned} T_i &= T_i(0)\cdot\mathbb{1}(X_i < \bar{x}) + T_i(1)\cdot\mathbb{1}(X_i \geq \bar{x}) \\ &= \begin{cases} T_i(0) & \text{if } X_i < \bar{x} \\ T_i(1) & \text{if } X_i \geq \bar{x} \end{cases}, \end{aligned}$$

where $T_i(0)$ and $T_i(1)$ denote the potential actual treatment status for each unit. In this more general setting, the treatment effect of interest is defined as the ratio of the reduced form effect (i.e., the effect of treatment assignment on the outcome) and the first-stage effect (i.e., the effect of treatment assignment on actual treatment status).

Our previous results for RD plots continue to apply without change to the reduced form model for Y_i and X_i: that is, for the conditional expectations $\mathbb{E}[Y_i(0)|X_i = x]$, $x < \bar{x}$, and $\mathbb{E}[Y_i(1)|X_i = x]$, $x \geq \bar{x}$. Furthermore, by imposing conditions analogous to Assumption 1 but now for the conditional expectations $\mathbb{E}[T_i(0)|X_i = x]$ and $\mathbb{E}[T_i(1)|X_i = x]$, and the conditional variances $\mathbb{V}[T_i(0)|X_i = x]$ and $\mathbb{V}[T_i(1)|X_i = x]$, our results will also apply to the first-stage model. The only distinct aspect of the first-stage model is that $T_i \in \{0, 1\}$, and thus one needs to employ the data-driven selectors discussed in Remarks 1 and 2, for ES and QS RD plots, respectively.

From a practical perspective, in the fuzzy RD design, RD plots can be used to depict both the reduced form effect as well as the take-up effect (i.e., the first-stage effect). Our results apply to both by simply changing the outcome variable used (either Y_i or T_i). For a related approach, see also Bertanha and Imbens (2014).

7.2 Incorporating Covariates

In some applications, researchers may want to incorporate covariates in the empirical analysis employing RD plots. We briefly discuss two such approaches for sharp RD designs but, as mentioned above, the upcoming discussion also applies to fuzzy RD designs. Suppose $\mathbf{Z}_i \in \mathbb{R}^d$, $i = 1, 2, \ldots, n$, is an observed covariate for each unit, and consider constructing an RD plot for the conditional expectations $\mathbb{E}[Y_i(0)|X_i = x, \mathbf{Z}_i = z]$, $x < \bar{x}$, and $\mathbb{E}[Y_i(1)|X_i = x, \mathbf{Z}_i = z]$, $x \geq \bar{x}$. Two simple approaches are (i) conditioning and (ii) employing generalized additive models (GAM).

A first conceptually straightforward approach to incorporate covariates in RD plots is to condition on them. Handling continuous covariates in this case is hard, while incorporating discrete covariates such as gender or age is easy. Specifically, the results in this article can be applied to the subsamples generated by the conditioning set of interest, or an approximation thereof (when the number of conditioning variables is large or Z_i is continuously distributed), provided the assumptions given above are extended to hold conditional on the appropriate conditioning set (and other appropriate regularity conditions hold).

A second way of incorporating covariates in the RD plots is to employ ideas from the GAM literature (Hastie and Tibshirani 1990), together with some method to remove the covariates such as backfitting. Specifically, in this approach it is assumed that the underlying conditional expectations take the form $\mathbb{E}[Y_i(0)|X_i = x, \mathbf{Z}_i = z] = \mathbb{E}[Y_i(0)|X_i = x] + g_0(z)$ and $\mathbb{E}[Y_i(1)|X_i = x, \mathbf{Z}_i = z] = \mathbb{E}[Y_i(1)|X_i = x] + g_1(z)$, for some unknown smooth functions $g_0(\cdot)$ and $g_1(\cdot)$, and then a flexible (nonparametric) method is used to remove the effect of the covariates. Once the covariates are removed, the RD plot can be constructed as discussed in this article but employing the appropriately adjusted outcome Y_i and running variable X_i.

8. CONCLUSION

This article introduced several optimal data-driven partition-size selectors for RD plots, focusing on the commonly used ES RD plot and also on an alternative QS RD plot. The resulting selectors lead to practical RD plots that are constructed in an automatic and objective way using the available data.

We developed two kinds of selectors, one tailored to approximate the underlying regression function and another to represent the underlying variability of the raw data. These selectors provide a benchmark for graphical analysis in the context of RD designs: the optimal choices of number of bins introduced can be interpreted as balancing variance and bias of a partitioning estimator of the underlying conditional expectations, and hence an empirical researcher may use these selectors to construct undersmoothed (more bins) or oversmoothed (fewer bins) RD plots or to give a formal interpretation to an ad hoc choice.

SUPPLEMENTARY MATERIALS

The supplemental appendix contains the proofs of the main theorems, additional methodological and technical results, more detailed simulation evidence, and other empirical illustrations.

[Received October 2014. Revised January 2015.]

REFERENCES

Abadie, A., and Imbens, G. W. (2006), “Large Sample Properties of Matching Estimators for Average Treatment Effects,” *Econometrica*, 74, 235–267. [1762]

——— (2010), “Estimation of the Conditional Variance in Paired Experiments,” *Annales d’Économie et de Statistique*, 91, 175–187. [1762]

Baryshnikov, Y., Penrose, M. D., and Yurich, J. E. (2009), “Gaussian Limits For Generalized Spacings,” *Annals of Applied Probability*, 19, 158–185. [1762]

Belloni, A., Chernozhukov, V., Chetverikov, D., and Kato, K. (2015), “Some New Asymptotic Theory for Least Squares Series: Pointwise and Uniform Results,” *Journal of Econometrics*, 186, 345–366. [1756,1762]

Bertanha, M., and Imbens, G. W. (2014), “External Validity in Fuzzy Regression Discontinuity Designs,” NBER Working Paper 20773, New York: National Bureau of Economic Research. [1768]

Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014a), “Robust Data-Driven Inference in the Regression-Discontinuity Design,” *Stata Journal*, 14, 909–946. [1756,1765]

——— (2014b), “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs,” *Econometrica*, 82, 2295–2326. [1753,1756,1758]

——— (2014c), “Supplement to ‘Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs’,” *Econometrica Supplemental Material*, 82, available at *http://dx.doi.org/10.3982/ECTA11757*. [1765]

——— (2015), “`rdrobust`: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs,” *R Journal*, 7, 38–51. [1756,1765]

Cattaneo, M. D., and Farrell, M. H. (2013), “Optimal Convergence Rates, Bahadur Representation, and Asymptotic Normality of Partitioning Estimators,” *Journal of Econometrics*, 174, 127–143. [1757]

Cattaneo, M. D., Frandsen, B., and Titiunik, R. (2015), “Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate,” *Journal of Causal Inference*, 3, 1–24. [1753,1765]

Chen, X. (2007), “Large Sample Sieve Estimation of Semi-Nonparametric Models,” in *Handbook of Econometrics* (Vol. VI), eds. J. J. Heckman and E. Leamer, New York: Elsevier Science B.V., pp. 5549–5632. [1756,1762]

Cook, T. D. (2008), “Waiting for Life to Arrive: A History of the Regression-discontinuity Design in Psychology, Statistics and Economics,” *Journal of Econometrics*, 142, 636–654. [1753]

David, H. A., and Nagaraja, H. N. (1998), “Concomitants of Order Statistics,” in *Handbook of Statistics* (Vol. 16), eds. N. Balakrishnan and C. R. Rao, New York: Elsevier Science B.V., pp. 487–513. [1762]

——— (2003), *Order Statistics*, Hoboken, NJ: Wiley. [1762]

Gelman, A., and Imbens, G. W. (2014), “Why High-Order Polynomials Should Not be Used in Regression Discontinuity Designs,” NBER Working Paper 20405, New York: National Bureau of Economic Research. [1757]

Ghosh, K., and Jammalamadaka, R. (2001), “A General Estimation Method Using Spacing,” *Journal of Statistical Planning and Inference*, 93, 71–82. [1762]

Hahn, J., Todd, P., and van der Klaauw, W. (2001), “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design,” *Econometrica*, 69, 201–209. [1753,1756]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, New York: Chapman & Hall/CRC. [1768]

Imbens, G., and Lemieux, T. (2008), “Regression Discontinuity Designs: A Guide to Practice,” *Journal of Econometrics*, 142, 615–635. [1753]

Imbens, G. W., and Kalyanaraman, K. (2012), “Optimal Bandwidth Choice for the Regression Discontinuity Estimator,” *Review of Economic Studies*, 79, 933–959. [1753,1756]

Imbens, G. W., and Wooldridge, J. M. (2009), “Recent Developments in the Econometrics of Program Evaluation,” *Journal of Economic Literature*, 47, 5–86. [1756]

Lee, D. S. (2008), “Randomized Experiments from Non-random Selection in U.S. House Elections,” *Journal of Econometrics*, 142, 675–697. [1753,1757,1765]

Lee, D. S., and Lemieux, T. (2010), “Regression Discontinuity Designs in Economics,” *Journal of Economic Literature*, 48, 281–355. [1753]

Lewbel, A., and Schennach, S. (2007), “A Simple Ordered Data Estimator for Inverse Density Weighted Expectations,” *Journal of Econometrics*, 136, 189–211. [1762]

Ludwig, J., and Miller, D. L. (2007), “Does Head Start Improve Children’s Life Chances? Evidence from a Regression Discontinuity Design,” *Quarterly Journal of Economics*, 122, 159–208. [1765]

Newey, W. K. (1997), “Convergence Rates and Asymptotic Normality for Series Estimators,” *Journal of Econometrics*, 79, 147–168. [1756,1762]

Porter, J. (2003), “Estimation in the Regression Discontinuity Model,” Working Paper, University of Wisconsin. [1753,1756]

Ruppert, D., Wand, M., and Carroll, R. (2009), *Semiparametric Regression*, New York: Cambridge University Press. [1755,1756,1761,1762]

Thistlethwaite, D. L., and Campbell, D. T. (1960), “Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment,” *Journal of Educational Psychology*, 51, 309–317. [1753]

Wand, M., and Jones, M. (1995), *Kernel Smoothing*, Boca Raton, FL: Chapman & Hall/CRC. [1761]