A Practical Guide to Regression Discontinuity Designs in Political Science[*]

Christopher Skovron[†] Rocío Titiunik[‡]

October 12, 2015

Abstract

We provide a practical guide to the analysis and interpretation of the regression discontinuity (RD) design, an empirical strategy that political scientists are increasingly employing to estimate causal effects with observational data. The defining feature of the RD design is that a treatment is assigned based on whether the value of a score exceeds a known cutoff. We review core conceptual issues, discussing the differences and similarities between RD designs and randomized controlled experiments, the basic conditions for identification of RD effects, and parametric and nonparametric alternatives to inference. We distinguish between a continuity-based RD approach, which relies on the continuity of the relevant regression functions and justifies polynomial estimation and inference methods, and a randomization-based RD approach, which relies on the assumption of random assignment of the treatment near the cutoff. We discuss best practices for estimation and inference in both cases. We illustrate all practical recommendations with a political science application, and provide companion `R` and `Stata` code to reproduce all empirical results. We conclude by offering concrete guidelines for the empirical analysis of RD designs in political science research.

[*] We have benefited from discussions with many people over the years, including Sebastián Calonico, Matias Cattaneo, Bob Erikson, Brigham Frandsen, Sebastián Galiani, Luke Keele, Marko Klašnja, Brendan Nyhan, Jas Sekhon, Gonzalo Vazquez-Bare, and José Zubizarreta. We thank Alexander Fouirnaies and Andy Hall for making their replication data available. Skovron gratefully acknowledges funding from the National Science Foundation Graduate Research Fellowship Program. Titiunik gratefully acknowledges funding from the National Science Foundation (SES 1357561).

[†] Graduate Student, Department of Political Science, University of Michigan, cskovron@umich.edu.

[‡] Corresponding author. Assistant Professor, Department of Political Science, University of Michigan, titiunik@umich.edu.

1 Introduction

The regression discontinuity (RD) design has become widely used by political scientists in recent years. In the simplest version of this design, units are assigned a score, and a treatment is given to those units whose value of the score exceeds a known cutoff and withheld from units whose value of the score is below the cutoff. For example, in single-member district elections with exactly two parties, a party wins the seat if the vote share it obtains is 50% or higher, but loses the election if its vote share falls short of the 50% cutoff. The key feature of the design is that the probability of receiving the treatment (winning) changes abruptly at the known threshold (50%). This discontinuous change in the probability of receiving treatment can be used to infer the effect of the treatment on the outcome of interest because, under certain assumptions, it makes units whose scores are barely below the cutoff comparable to units whose scores are barely above it.

The RD design was originally proposed by Thistlethwaite and Campbell (1960), and has become increasingly common in the social sciences in the last two decades. In political science, the number of articles that use RD designs has increased steadily in the last five years. In a review of the literature in which we included more than thirty general interest and field-specific Political Science journals, we found dozens of recently published articles that use this design (see Supplemental Appendix). By far, studies that use vote share as the score—and thus winning the election as the treatment—are the most common. These studies focus on the effects of electoral victory on various outcomes, including incumbents' future electoral performance, distributive decisions, and economic performance. RD designs based on elections have been found to work well in many contexts (Eggers et al. 2015), although Caughey and Sekhon (2011) show that the application to postwar U.S. House elections is problematic.

Our review of the literature revealed that, despite their substantive similarities, RD studies in political science differ significantly in how the authors choose to estimate the effects of interest, make statistical inferences, present their results, evaluate the plausibility of the RD assumptions, and interpret the estimated effects. This lack of consensus about the best way to perform validation, estimation, inference, and interpretation of RD results makes it hard for scholars and policy-makers to judge the plausibility of the results presented and to compare results across different RD studies. This is the motivation behind our article. We review the different approaches that are possible in the analysis and interpretation of RD designs, and provide clear practical recommendations for making defensible inferences from RD designs. Our ultimate goal is to provide an accessible practical guide that increases the transparency in RD empirical analysis in all subfields of political science.

Our review focuses on the most basic and defining features of the RD design, clarifying important conceptual differences between RD designs and experiments, and offering simple guidelines for researchers who wish to employ this design in their empirical research. We try as much as possible to present this information in a way that is accessible to non-experts, prioritizing conceptual over technical distinctions wherever possible. Our decision to illustrate and discuss core concepts in great detail and at length means that we were forced to exclude from our review some of the most advanced and recent RD topics. The list of topics that we do not discuss includes RD designs with multiple scores (Papay, Willett, and Murnane 2011) and multiple cutoffs (Cattaneo et al. 2015), methods for extrapolation away from the cutoff (Angrist and Rokkanen 2016; Wing and Cook 2013), geographic RD designs (Keele and Titiunik 2015; Keele, Titiunik, and Zubizarreta 2015), quantile RD treatment effects (Frandsen, Frölich, and Melly 2012) and regression kink designs (Card et al. 2015), among others. Prior articles that review methods for RD analysis include Imbens and Lemieux (2008) and Lee and Lemieux (2010)—but many of the topics we discuss are more recent and therefore not discussed in these pieces.

The rest of the article is organized as follows. In the following section we introduce the basic RD model, explain the notation, and discuss the differences between RD designs and experiments. In Section 3, we briefly discuss how to present RD effects in a graphical way. In Section 4 we discuss the continuity-based approach to estimation and inference in RD designs, highlighting its advantages and limitations. In Section 5, we present an alternative approach based on a local-randomization assumption, and discuss the type of applications where this approach is most useful. In Section 6, we discuss how to evaluate the plausibility of the RD assumptions in graphical and formal ways. In Section 7, we apply all the methods discussed to a re-analysis of the study by Fouirnaies and Hall (2014) on the effect of incumbency on campaign contributions in the U.S. House and state legislatures. We conclude in Section 8 with a list of recommendations for practice.

2 How to recognize and interpret an RD design

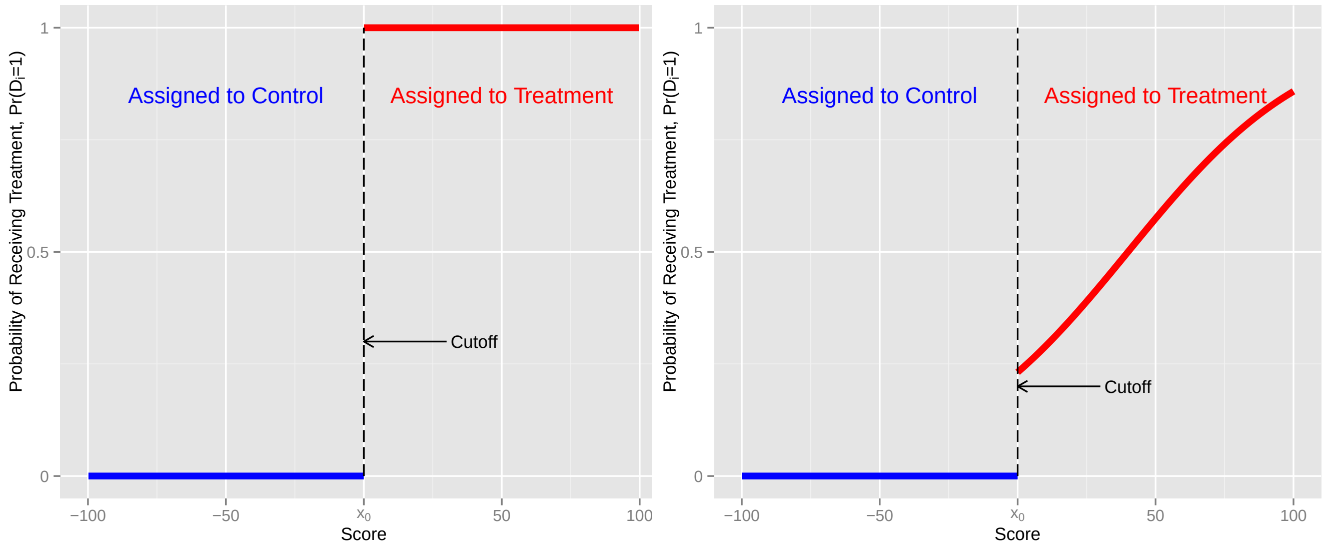
In the RD design, all the units in the study receive a score, and a treatment is assigned to those units whose score is above a known cutoff and withheld from those units whose score is below the cutoff. For example, as mandated by Section 203 of the Voting Rights Act, all U.S. counties that have a language minority that includes more than 10,000 citizens must provide Spanish-language ballots. Hopkins (2011) uses this discontinuity in Spanish-ballot access to study the impact of language assistance on voter turnout among Latino citizens. In this RD design, the score (also known as the *running variable*) is the population size of the language minority in each county, the cutoff is 10,000, and the treatment is the presence of Spanish-language ballots—a treatment that is received by all counties whose population exceeds 10,000.

These three components—score, cutoff, and treatment—define RD designs in generality. When all units in the study comply with the treatment condition they have been assigned, we say that the RD is *sharp*. In contrast, when some of the units fail to receive the treatment despite having a score above the cutoff and/or some units receive the treatment despite having been assigned to the control condition, we say that the RD design is *fuzzy*. This occurs, for example, when units with score above the cutoff are eligible to participate in a program, but participation is not mandatory.

We now introduce some notation to formalize these definitions. We assume that we have n units, indexed by $i = 1, 2, \ldots, n$, each unit has a score or running variable X_i, and x_0 is a known cutoff. Units with $X_i \geq x_0$ are assigned to the treatment condition, and units with $X_i < x_0$ are assigned to the control condition. This assignment, denoted with the variable Z_i, is defined as $Z_i = \mathbb{1}(X_i \geq x_0)$, where $\mathbb{1}(\cdot)$ is the indicator function. In addition, the binary variable D_i denotes whether the treatment was actually received. In a sharp RD design, units comply perfectly with their assignment, so $Z_i = D_i$ for all i, which implies that the treatment received is a deterministic function of the score. In contrast, in a fuzzy RD design we have $Z_i \neq D_i$ for some units.

The difference between the sharp and fuzzy RD designs is illustrated in Figure 1, where we plot the probability of receiving treatment, $\Pr(D_i = 1)$, as a function of the score. As shown in Figure 1(a), in a sharp RD design the probability of receiving treatment changes exactly from zero to one at the cutoff. In contrast, in a fuzzy RD design, the change in the probability of receiving treatment at the cutoff is always less than one. Figure 1(b) illustrates a fuzzy RD design where units with score below the cutoff comply perfectly with the treatment, but compliance with the treatment is imperfect for units with score above the cutoff—for these units, the probability of receiving treatment is always less than one, and increases with the value of the score.

Figure 1: Probability of Receiving Treatment in Sharp vs. Fuzzy RD Designs



(a) Sharp RD

(b) Fuzzy RD (One-Sided)

Because of space considerations, our discussion in the remainder of the article will focus on the sharp RD design where treatment assignment and treatment received are identical—but we include a discussion of the fuzzy RD design in Section S4 of the Supplemental Appendix. We note, however, that any fuzzy RD design can be seen as a sharp RD design where the treatment of interest is redefined to be the treatment assignment, so most of the concepts and recommendations we discuss below are directly relevant to the fuzzy case.

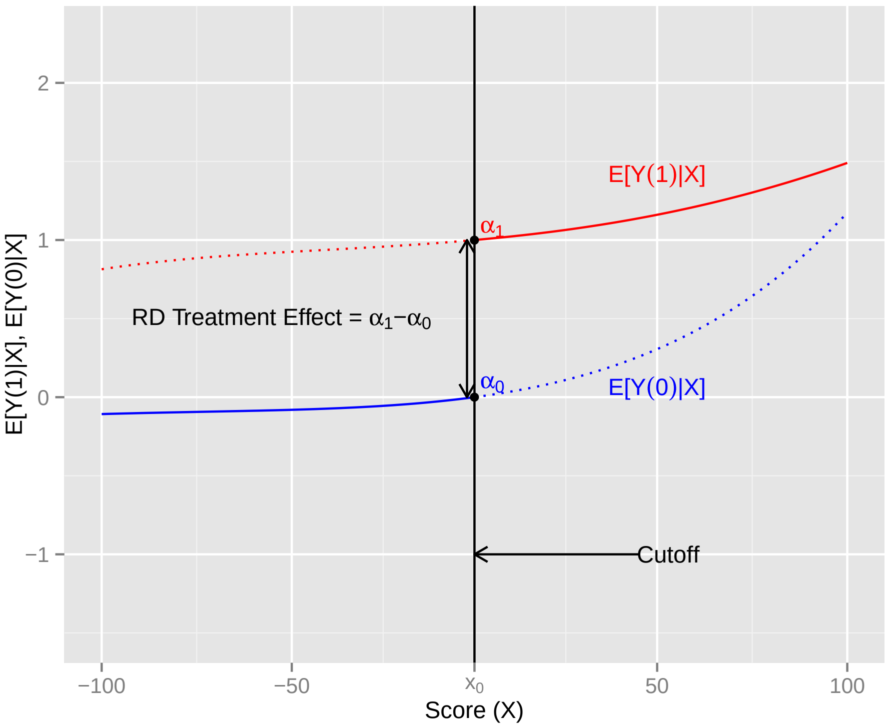
Following the RD literature, we assume that each unit has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, which correspond, respectively, to the outcomes that would be observed if the unit received treatment or control. We adopt the usual econometric perspective that sees the data $(Y_i, X_i)_{i=1}^{n}$ as a sample from a larger population and thus the potential outcomes $(Y_i(1), Y_i(0))_{i=1}^{n}$ as stochastic variables—but we consider an alternative perspective in Section 5.

It follows that the observed outcome is

$$Y_i = \begin{cases} Y_i(0) & \text{if } X_i < x_0, \\ Y_i(1) & \text{if } X_i \geq x_0. \end{cases}$$

Applying this notation to the study by Hopkins (2011), X_i refers to the size of the language minority for county i, $x_0 = 10,000$, $Z_i = \mathbb{1}(X_i \geq 10,000)$, and Y_i is the observed Latino turnout at the county level. The fundamental problem of causal inference is that, for those units whose score is below the cutoff, we only observe the outcome under the control condition and, for those units whose score is above the cutoff, we only observe the outcome under treatment. For example, we cannot observe the level of Latino turnout that would have occurred in counties with language minorities over 10,000 citizens if those counties had not had access to Spanish-language ballots.

Figure 2: RD Treatment Effect in Sharp RD Design



We illustrate this missing data problem in Figure 2, which plots the average potential outcomes given the score, $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$, against the different values x that the score may take. Note that $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ are functions of X_i. In statistics, conditional expectation functions such as $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$ are called *regression functions*.

As shown in Figure 2, the regression function $\mathbb{E}[Y_i(1)|X_i]$ is observed for values of the score to the right of the cutoff—because when $X \geq x_0$, the observed outcome Y_i is equal to the potential outcome under treatment, $Y_i(1)$, for every i. This is represented with the solid red line. However, to the left of the cutoff, all units are untreated, and therefore $\mathbb{E}[Y_i(1)|X_i]$ is not observed (represented by a dashed red line). A similar phenomenon occurs for $\mathbb{E}[Y_i(0)|X_i]$, which is observed for values of the score to the left of the cutoff (solid blue line), $X < x_0$, but unobserved for $X \geq x_0$ (dashed blue line). In other words, the observed average outcome given the score is as follows:

$$\mathbb{E}[Y_i|X_i] = \begin{cases} \mathbb{E}[Y_i(0)|X_i] & \text{if } X_i < x_0, \\ \mathbb{E}[Y_i(1)|X_i] & \text{if } X_i \geq x_0. \end{cases}$$

As seen in Figure 2, the average treatment effect at any value of the score is the vertical distance between the two curves $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$. This distance cannot be directly estimated because we never observe both curves simultaneously, that is, for the same value of X. However, a special situation occurs at the cutoff $X = x_0$, because this is the only point at which we "almost" observe both curves. To see this, we imagine having units with score exactly equal to x_0, and units with score barely below x_0—that is, with score $X = x_0 - \varepsilon$ for a small and positive ε. The former units would receive treatment, and the latter would receive control. Yet if the values of the average potential outcomes at x_0 are not abruptly different from their values at points near x_0, the units with $X_i = x_0$ and $X_i = x_0 - \varepsilon$ would be identical except for their treatment status. The result is that we can calculate the vertical distance at x_0 and learn about the average treatment effect at this point. This notion of comparability between units with very similar values of the score but on opposite sides of the cutoff is the fundamental concept on which all RD designs are based.

This intuitive explanation of how RD designs allow us to learn about the average treatment effect was derived formally by Hahn, Todd, and van der Klaauw (2001), who showed that if, among other conditions, the regression functions $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$ are continuous at x_0, then in a sharp RD design we have

$$\lim_{x \downarrow x_0} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow x_0} \mathbb{E}[Y_i|X_i = x] = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x_0] \tag{1}$$

Equation 1 says that if the average potential outcomes are continuous functions of the score at x_0, the difference between the limits of the treated and control average *observed* outcomes as the score converges to x_0 is equal to the average treatment effect at x_0. We call this effect the *RD treatment effect*, defined as the right-hand-side of Equation 1:

$$\tau^{\texttt{RD}} \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x_0]$$

2.1 Analogies and differences between RD designs and experiments

A very influential interpretation of the result in Equation 1 was offered by Lee (2008), who argued that an RD design can be as credible as a randomized experiment for units near the cutoff. Lee showed that, as long as there is a random chance element to the value of the score that each unit ultimately receives *and* the probability of this "error" does not change abruptly at the cutoff, the RD design can be interpreted as an experiment that randomly assigns units to treatment and control in an neighborhood of the cutoff. This holds even when the units' unobservable characteristics and choices affect the score.

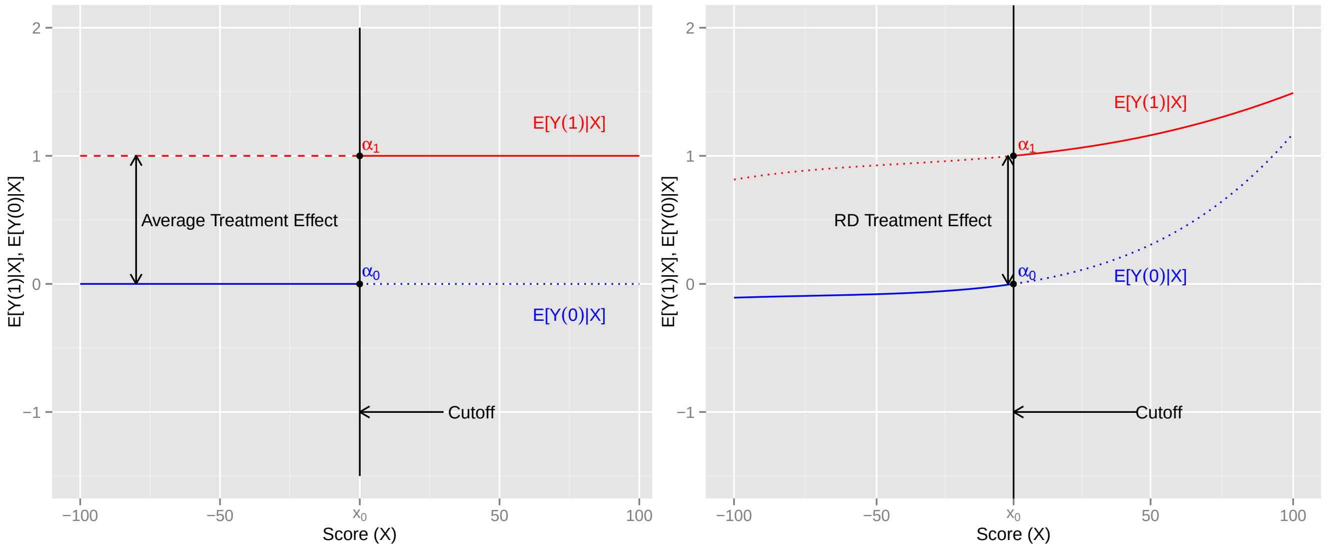
This interpretation is very useful, since it is more intuitive to conceive of RD designs as local experiments than it is to think about the continuity of regression functions. Thinking about continuity, however, cannot be completely avoided because the as-if random interpretation of RD designs proposed by Lee (2008) still requires that probability of the score's random error component be continuous for every individual. This condition will fail, for example, if some units have the ability to exactly control their score value.

Clarifying the differences and similarities between RD designs and experiments is important for two reasons. First, the interpretation of the treatment effects estimated by RD designs changes depending on whether we strictly interpret an RD design as a local experiment or we use the as-good-as-randomized interpretation as an heuristic notion. Second, the procedures for RD estimation and inference under a continuity condition may not be the most appropriate under a local randomization assumption, and vice versa.

To illustrate, consider a valid RD design where the continuity of $\mathbb{E}[Y_i(1)|X]$ and $\mathbb{E}[Y_i(0)|X]$ holds and units do not have the ability to interfere with the assignment mechanism or "manipulate" the value of their score. There is a crucial distinction between this ideal RD design and a randomized experiment. In an experiment, there is no need to make assumptions about the shape of the average potential outcomes, since these functions are by construction equal in treatment and control groups. In contrast, in an RD design, even when units do not manipulate their score, inferences depend crucially on the functional form of the regression functions.

To see this, it is helpful to note that any experiment can be recast as an RD design where the score is a random number and the cutoff is chosen to ensure a certain treatment probability. For example, consider an experiment in a student population that randomly assigns a scholarship with probability $1/2$. This experiment can be seen as an RD design where each student is assigned a uniform random number between 0 and 100, and the scholarship is given to students whose score is above 50, a scenario we illustrate in Figure 3(a). The crucial feature of an experiment recast as an RD design is that the value of the score, by virtue of being a randomly generated number, is unrelated to the average potential outcomes. This is the reason why, in the figure, the average potential outcomes $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$ take the same constant value for all values of X. Since the regression functions are flat, the vertical distance between them can be recovered by the difference between the average observed outcomes among all units in the treatment and control groups, i.e. $\mathbb{E}[Y_i|X_i \geq 50] - \mathbb{E}[Y_i|X_i < 50] = \mathbb{E}[Y_i(1)|X_i \geq 50] - \mathbb{E}[Y_i(0)|X_i < 50] = \mathbb{E}[Y_i(1)] - \mathbb{E}[Y_i(0)]$.

Figure 3: Experiment versus RD Design



(a) Randomized Experiment

(b) RD Design

In contrast, Figure 3(b) illustrates an RD design where the average treatment effect at the cutoff is the same as in the experimental setting in Figure 3(a), $\alpha_1 - \alpha_0$, but where the average potential outcomes are non-constant functions of the score. This relationship between score and potential outcomes is characteristic of most RD designs: since the score is often related to the the units' ability, resources or performance (poverty index, vote shares, test scores), units with higher score values are often systematically different from units whose scores are lower. For example, counties where Hispanics are a large proportion of the total population may be more likely to have Hispanic representatives at different levels of government, which may increase voter turnout among Hispanics due to an empowerment effect. In this scenario, the percentage of Hispanics (the score) will be positively related to the county's average Hispanic turnout with and without Spanish-language ballots (the treatment), leading to positive slopes as in Figure 3(b).

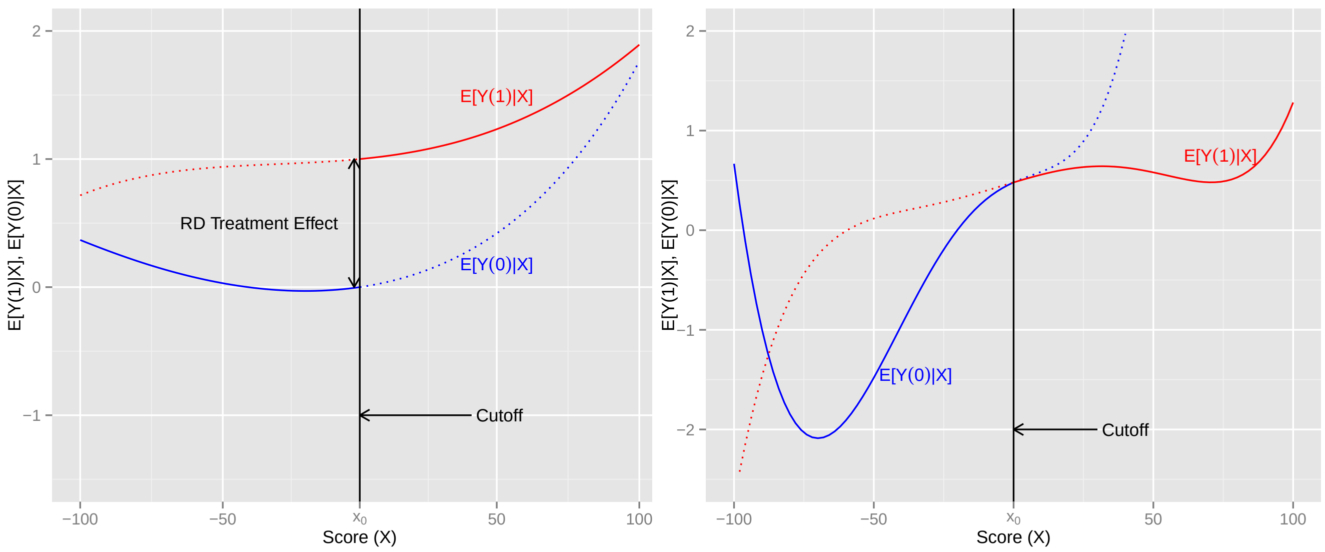
The crucial difference between the scenarios in Figures 3(a) and 3(b) is our knowledge about the functional forms. As shown in Equation 1, the average treatment effect in 3(b) can be estimated by calculating the limit of the average observed outcomes as the score approaches the cutoff for the treatment and control groups, $\lim_{x \downarrow x_0} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow x_0} \mathbb{E}[Y_i|X_i = x]$. The estimation of these limits requires that the researcher know or at least approximate the regression functions, and using the incorrect functional form will lead to invalid estimates of the RD treatment effect. This is in stark contrast to the experiment depicted in Figure 3(a), where the random assignment of the score implies that the average potential outcomes are unrelated to the score and estimation does not require functional form assumptions—since the regression functions are constant in the entire region where the score was randomly assigned.

2.2 The local nature of RD effects

Another important aspect of the interpretation of RD effects has to do with their external validity. As we explained, the RD effect can be interpreted graphically as the vertical difference between $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$ at the point where the score equals the cutoff, $X_i = x_0$. In the general case where the average effect of treatment varies as a function of X, the RD effect may not be informative of the average effect of treatment at values of X different from x_0. For this reason, in the absence of specific assumptions about the global shape of the regression functions, the effect recovered by the RD design is the *local* average effect of treatment at x_0.

How much can be learn from such a local effect will depend on each particular application. For example, in the scenario illustrated in Figure 4(a), the vertical distance between $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$ at x_0 is considerably higher than at other points, such as $X = -100$ and $X = 100$, but the effect is positive everywhere. A more heterogeneous scenario is shown in Figure 4(b), where the effect is zero at the cutoff but ranges from positive to negative at other points. Since in real examples the counterfactual (dotted) regression functions are not observed, it is not possible to know with certainty the degree of external validity of any given RD application. Increasing the external validity of the RD estimates can be achieved by making additional assumptions that allow for extrapolation (Angrist and Rokkanen 2016; Wing and Cook 2013) or replicating similar RD designs in different settings. On this regard, RD designs are no different from experiments.

Figure 4: Local Nature of RD Effect



(a) Mild Heterogeneity

(b) Severe Heterogeneity

3 Graphical illustration of RD treatment effects

An appealing feature of the RD design is that it can be illustrated graphically. This graphical representation, in combination with the formal approaches to estimation and inference discussed below, adds transparency to the analysis by plotting all the observations used for estimation and inference, and allows researchers to readily summarize the RD effect.

One alternative would be to simply create a scatter plot of the observed outcomes against the score values, separately identifying the points above and below the cutoff. However, this strategy is rarely useful, since it is hard to see "jumps" by simply looking at the raw data. Figure 5(a) shows raw simulated data corresponding to a sharp RD design with treatment assignment defined by $Z_i = \mathbb{1}(X_i \geq 50)$ and a true average treatment effect at the cutoff equal to one unit. Despite the true RD treatment effect being large and positive, a jump in the values of the outcome at the cutoff is hard to see to the naked eye.

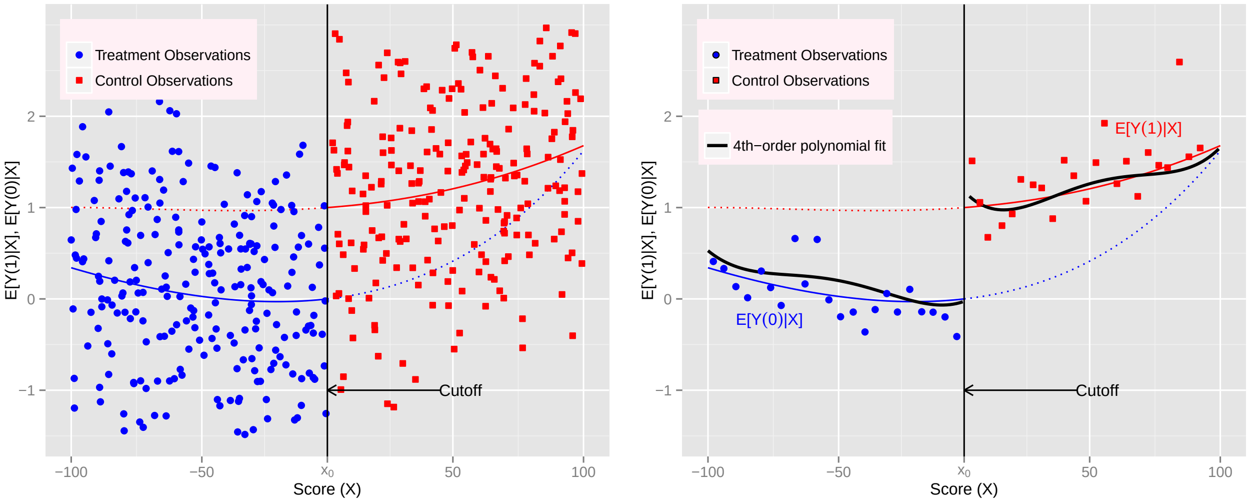
A more useful approach is to aggregate the data before plotting. The typical RD plot presents two summaries: (i) a global polynomial fit and (ii) local sample means. The global polynomial fit is simply the predicted values from two fourth- or fifth-order polynomials of the outcome on the score, fitted separately above and below the cutoff. The local means are created by choosing disjoint intervals or bins of the score, calculating the mean of the outcome within each bin, and then plotting the binned outcomes against the mid point of the bin. The number and length of the bins can be chosen manually, although data-driven methods are preferable. See Calonico, Cattaneo, and Titiunik (2016) for a detailed discussion of the features of RD plots and several methods to choose the number of bins that are data-driven and optimal according to various criteria.

Figure 5(b) plots the same data used in Figure 5(a), but using local means instead of raw data points and adding a fourth-order polynomial fit. The RD effect, which was hidden in the raw scatter plot, is now clearly visible.

4 How to Estimate Effects and Test Hypothesis in an RD Design: The Continuity-Based Approach

We now discuss the most common framework for estimation and inference in RD designs. This framework is based on the continuity assumption discussed above, and uses nonparametric regression methods to approximate the unknown regression functions. All the methods we describe have been implemented in the statistical programs `R` and `Stata`—see Calonico, Cattaneo, and

Figure 5: Visualization of RD Effects



(a) Raw data

(b) Binned means

Titiunik (2014a,c), and Cattaneo, Titiunik, and Vazquez-Bare (2015b). We illustrate the use of these methods in practice in Section 7.

Local Polynomial Point estimation

A fundamental feature of the sharp RD design is that there are no control observations with score equal to the cutoff x_0: by construction, every unit whose score is x_0 (or higher) is treated. This means that we cannot use observations at the cutoff to estimate $\mathbb{E}[Y_i(0)|X_i = x_0]$. Instead, we must use control observations that are below the cutoff, preferably observations that are very near it. In contrast, the treatment assignment rule does allow for treated observations to have score equal to x_0. However, because the score is continuous, the probability that unit i has a score exactly equal to the cutoff value is zero. Therefore, in practice, there are no treated observations exactly at the cutoff with which to estimate $\mathbb{E}[Y_i(1)|X_i = x_0]$ and, analogously to the case of $\mathbb{E}[Y_i(0)|X_i = x_0]$, the estimation of $\mathbb{E}[Y_i(1)|X_i = x_0]$ must be based on observations above, and near, x_0. Thus, estimation of the RD effect must necessarily rely on *extrapolation*.

Since $\tau^{\texttt{RD}}$ is the vertical distance between the $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ at x_0, estimation of this effect involves estimating the points α_0 and α_1 in Figure 2 and then calculating their difference. As explained above, the exact functional form of $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$

is unknown; estimation therefore proceeds by *approximating* these functions, that is, by choosing a functional form that, according to some objective criteria, can be considered close enough to the true functions.

The problem of approximating an unknown function is well understood in calculus. According to the Taylor Theorem, any sufficiently smooth function can be approximated by a polynomial function in a neighborhood of a point. Applied to the RD point estimation problem, this result says that the unknown regression function $\mathbb{E}[Y_i|X_i = x]$ can be approximated in a neighborhood of x_0 by a polynomial on the normalized score—i.e., on $x - x_0$. (For additional details, see Fan and Gijbels 1996, .)

The preferred method is to estimate this polynomial *locally*, that is, using only observations near the cutoff point. This approach uses only observations that are between $x_0 - h$ and $x_0 + h$, where h is some chosen bandwidth. Moreover, within this bandwidth, observations closer to x_0 often receive more weight than observations further away, where the weights are determined by a *kernel* function $K(\cdot)$. This estimation approach, usually called local polynomial modeling, is *nonparametric* because it does not assume a particular parametric form of the unknown underlying regression functions. A local approach is preferable to a global approximation because, in the latter, observations far from the cutoff can distort the approximation near the cutoff and lead to a misleading effect—see Gelman and Imbens (2014), and Section S2 of the Supplemental Appendix for a graphical illustration.

Local-polynomial estimation consists of the following steps:

1. Choose bandwidth h.

2. For each observation i, calculate weight $w_i = K(\frac{x_i - x_0}{h})$.

3. For observations above the cutoff—i.e. with $X_i \geq x_0$, fit a weighted least squares regression of the outcome Y_i on a constant, $(X_i - x_0)$, $(X_i - x_0)^2, \ldots, (X_i - x_0)^p$ where p is the chosen polynomial order, with weight w_i for each observation. The estimated intercept from this local weighted regression, $\hat{\alpha}_1$ is an estimate of the point α_1.

4. For observations below the cutoff—i.e. with $X_i < x_0$, fit a weighted least squares regression of the outcome Y_i on a constant, $(X_i - x_0)$, $(X_i - x_0)^2, \ldots, (X_i - x_0)^p$ where p is the

same polynomial order chosen above, with weights w_i for each observation. The estimated intercept from this local weighted regression, $\hat{\alpha}_0$ is an estimate of the point α_0.

5. Calculate the RD point estimate as $\hat{\tau}^{\texttt{RD}} = \hat{\alpha}_1 - \hat{\alpha}_0$.

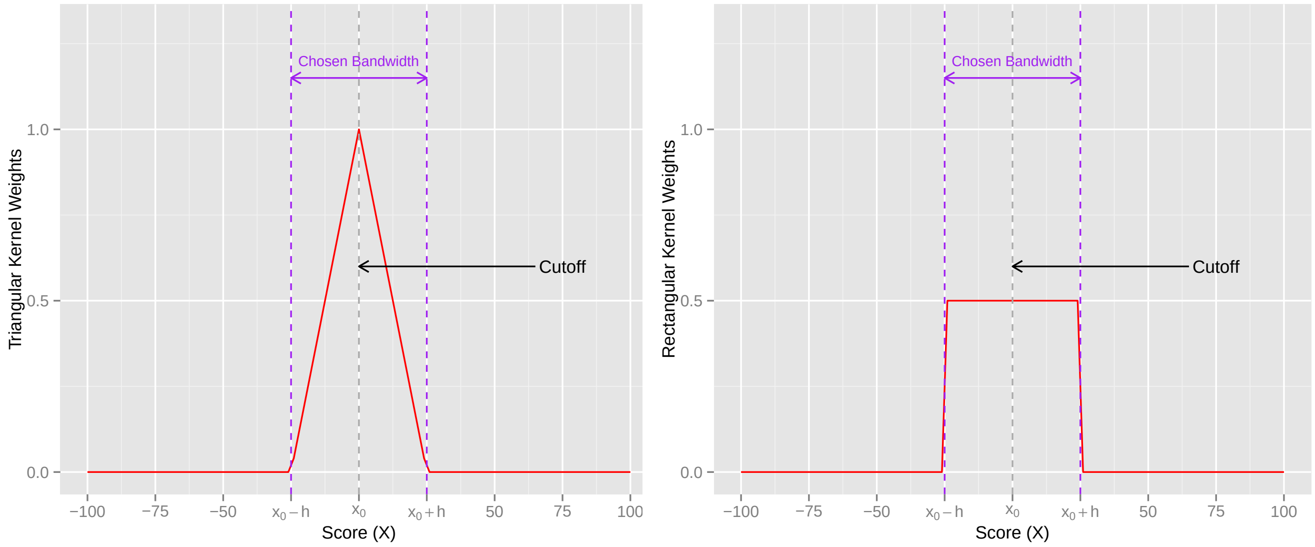
The implementation of a local polynomial approach thus requires the choice of three ingredients: the bandwidth h, the kernel function $K(\cdot)$, and the order of the polynomial p. We discuss each choice below.

Choice of kernel function. The kernel function $K(z)$ assigns non-negative weights and satisfies $\int K(z)dz = 1$. The recommended choice is the triangular kernel function, $K(\frac{X_i - x_0}{h}) = \left(1 - \left|\frac{X_i - x_0}{h}\right|\right) \mathbb{1}\left(\left|\frac{X_i - x_0}{h}\right| \leq 1\right)$ because, when using an optimal bandwidth, it leads to a point estimator with optimal variance and bias properties. As illustrated in Figure 6(a), this kernel function assigns zero weight to all observations with score outside the interval $[x_0 - h, x_0 + h]$, and positive weights to all observations within this interval. The weight is maximized at $X_i = x_0$ and declines symmetrically as the value of the score gets farther from the cutoff.

Despite the desirable asymptotic properties of the triangular kernel, researchers sometimes prefer to use the more simple uniform or rectangular kernel. $K(\frac{X_i - x_0}{h}) = 0.5 \cdot \mathbb{1}\left(\left|\frac{X_i - x_0}{h}\right| \leq 1\right)$, which gives zero weight all observations with score outside $[x_0 - h, x_0 + h]$, and equal weight to all the observations whose scores are in this interval—see Figure 6(b). Employing a local-linear estimation with bandwidth h and a rectangular kernel is therefore equivalent to estimating a simple linear regression without weights using observations whose distance from the cutoff is at most h, i.e. observations with $X_i \in [x_0 - h, x_0 + h]$. In practice, estimation results are typically insensitive to the choice of kernel.

Choice of bandwidth. The choice of bandwidth h is the most important. This choice controls the width of the neighborhood around the cutoff that is used to fit the local polynomial that approximates the unknown functions. In general, choosing a very small h will reduce the error or bias of the local polynomial approximation, but will increase the variance of the estimated coefficients because very few observations will be available for estimation. On the other hand, a very large h may result in a large bias if the unknown function differs considerably from the polynomial model used for approximation, but will result in estimates with low variance because the number of observations in the interval $[x_0 - h, x_0 + h]$ will be large when h is large. For this reason, the

Figure 6: Kernel Weights for RD Estimation



(a) Triangular Kernel

(b) Rectangular Kernel

choice of bandwidth is said to involve a bias-variance trade-off.

Since the results will often depend on the choice of bandwidth, we recommend selecting h in a data-driven, automatic way to avoid specification searching and ad-hoc decisions. The standard approach is to use data-driven methods to choose the bandwidth that minimizes an approximation to the asymptotic mean squared error (MSE) of the RD point estimator, $\hat{\tau}^{\mathtt{RD}}$. Since the MSE of an estimator is the sum of its bias squared plus its variance, this approach effectively chooses h to optimize the bias-variance trade-off. The procedure involves deriving the asymptotic MSE approximation, optimizing it with respect to h, and estimating the unknown quantities in the resulting formula (Calonico, Cattaneo, and Titiunik 2014b; Imbens and Kalyanaraman 2012). The technical details are outside the scope of our review, but it is important to note that the MSE-optimal bandwidth is designed to be optimal for point estimation, *not* for inference. Thus, although this bandwidth leads to an RD point estimator that has minimum MSE, it leads to standard confidence intervals that are invalid in most cases, which is why alternative confidence intervals are needed (Calonico, Cattaneo, and Titiunik 2014b). We discuss this in more detail in the inference section below.

Choice of polynomial order. The final choice involves the order of the local polynomial used.

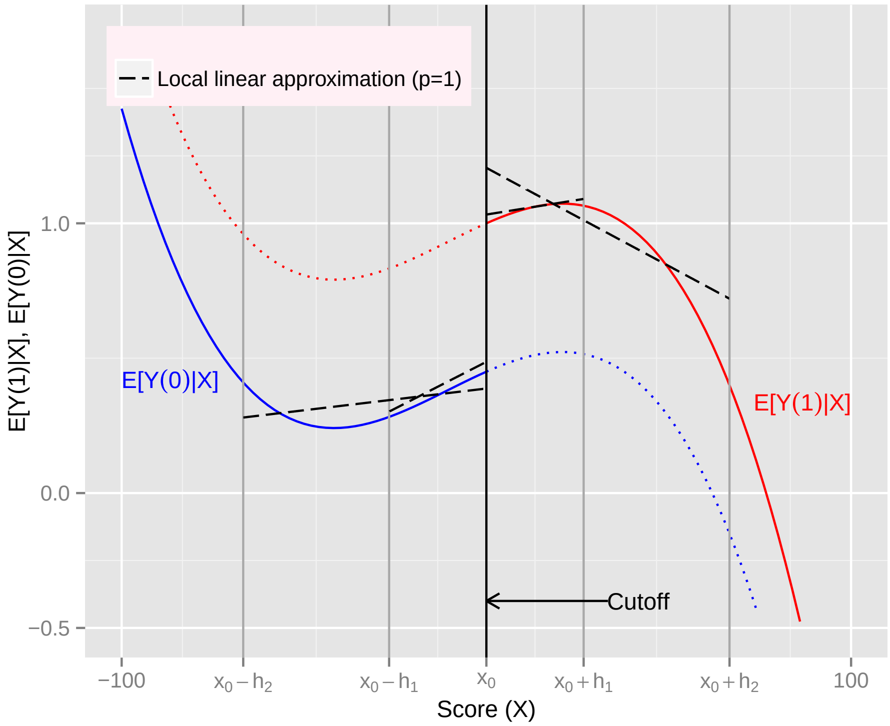
Since the accuracy of the approximation is essentially controlled by the bandwidth, the order of the polynomial should be kept low. Bandwidth selectors optimize the bias-variance trade-off *given* the chosen polynomial order. For example, if a linear fit results in an inaccurate approximation within a given bandwidth, the approximation can always be improved by reducing the size of the bandwidth as needed. Moreover, high order polynomials can lead to severe approximation errors due to over-fitting (Gelman and Imbens 2014).

Several issues should be considered when choosing the specific order of the local polynomial. First, a polynomial of order zero—a constant fit—should be avoided, as its has undesirable properties in boundary points, which is precisely where RD estimation must occur. Second, for a given bandwidth, increasing the order of the polynomial generally improves the accuracy of the approximation, but it increases the variability. More precisely, it can be shown that the asymptotic variance of the local polynomial fit increases when going from an odd to an even order, but stays constant when going from an even to an odd order. Thus, the recommendation is to use the smallest odd order possible. In the case of the RD point estimate, since the object of interest is a conditional expectation (i.e., a derivative of order zero), the recommended choice is a polynomial of order one, that is, a local—i.e., inside the bandwidth—linear regression.

Figure 7 illustrates how the error in the approximation is directly related to the bandwidth choice. The unknown regression functions in the figure, $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$, have considerable curvature. At first, it would seem inappropriate to approximate these functions with a linear function. Indeed, inside the interval $[x_0 - h_2, x_0 + h_2]$, a linear approximation yields an RD effect equal to $a_2 - b_2$, which is considerably different from the true effect $\alpha_1 - \alpha_0$. Thus, a linear regression within bandwidth h_2 results in a large approximation error. However, reducing the bandwidth from h_2 to h_1 improves the linear approximation considerably, as now the approximated effect $a_1 - b_1$ is much closer to the true effect. The reason is that the regression functions are nearly linear in the interval $[x_0 - h_1, x_0 + h_1]$, and therefore the linear approximation results in a very small error. This illustrates the general principle that, given a polynomial order, the accuracy of the approximation can always be improved by reducing the bandwidth.

The procedure of local polynomial RD point estimation is illustrated in Figure 8, where a polynomial of order one is fit locally within a the MSE-optimal bandwidth h_1—observations outside

Figure 7: Bias in Local Approximations



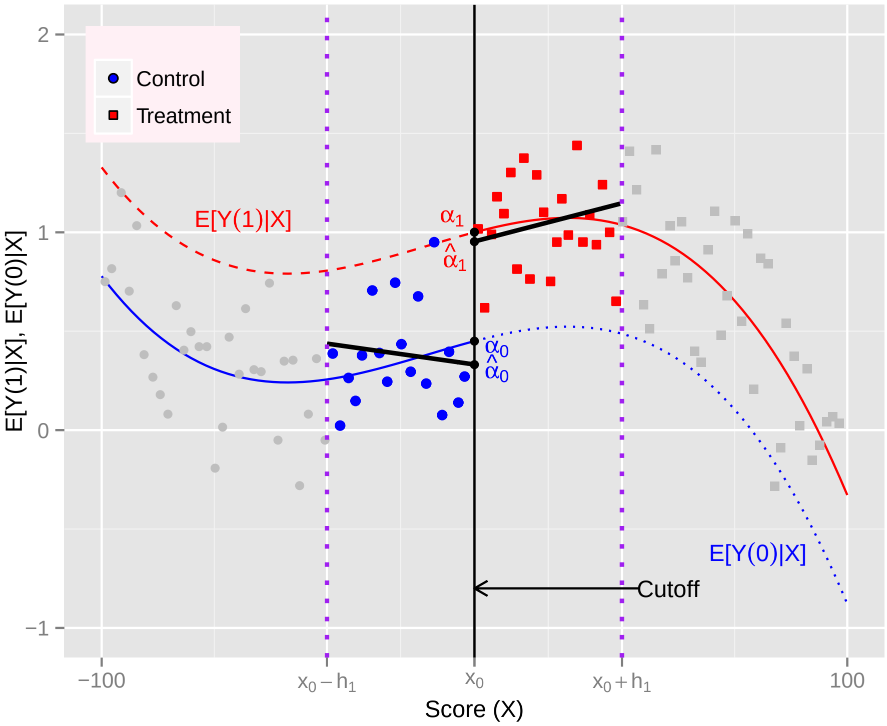
this bandwidth are not used in the estimation. The RD effect is $\alpha_1 - \alpha_0$ and the local polynomial estimator of this effect is $\hat{\alpha}_1 - \hat{\alpha}_0$. The dots represent binned outcome means.

Properties of the local polynomial RD point estimator. Assuming that the bandwidth sequence shrinks appropriately as the sample size increases, the local polynomial point estimator is consistent for $\tau^{\texttt{RD}}$. This follows directly from the general properties of nonparametric local polynomial modeling (see Fan and Gijbels 1996). Moreover, it can be shown that MSE-optimal bandwidth satisfies this rate restriction. Thus, using a local polynomial estimator within a MSE-optimal bandwidth leads to a consistent and optimal (in an asymptotic MSE sense) RD point estimator.

Local Polynomial Inference

Once a point estimator has been obtained with local polynomial methods, we will be interested in testing hypotheses and constructing confidence intervals. At first glance, it seems that ordinary least-squares inference methods should be appropriate, since local polynomial estimation, in practice, involves nothing more than fitting two weighted least-squares regressions within a region

Figure 8: RD Estimation with local polynomial



near x_0 (controlled by the bandwidth). However, using OLS results would treat local polynomial estimation as parametric, while in fact this estimation method is nonparametric.

The MSE-optimal bandwidth discussed above results in an optimal and consistent RD point estimator. Point estimation and inference, however, are different concepts, and a well-known result in local polynomial modeling is that the rate of convergence of the MSE-optimal bandwidth leads to a bias in the distributional approximation of the estimator that is used to create confidence intervals.

More precisely, the MSE-optimal bandwidth h leads to a local polynomial RD point estimator $\hat{\tau}^{\texttt{RD}}$ that has an approximate distribution $\hat{\tau}_n^{\texttt{RD}} \sim \mathcal{N}(\tau^{\texttt{RD}} + \texttt{B}_n, \texttt{V}_n)$, where $\texttt{B}_n$ is the asymptotic bias or approximation error of the local polynomial estimator, and $\texttt{V}_n$ is its asymptotic variance. We use sub-indices to indicate that these quantities depend on the sample size n. Given this distributional approximation, an asymptotic 95-percent confidence interval for $\tau^{\texttt{RD}}$ is given by $I = \left[(\hat{\tau}_n^{\texttt{RD}} - \texttt{B}_n) \pm 1.96 \cdot \sqrt{\texttt{V}_n}\right]$. These confidence intervals (CI) depend on the unknown bias $\texttt{B}_n$, and any practical procedure that ignores this bias term will lead to incorrect inferences unless the misspecification error is negligible.

This bias term arises because the local polynomial approach is nonparametric: instead of *assuming* that the underlying function is a polynomial (as would occur in ordinary least-squares estimation), this approach uses the polynomial to *approximate* this function. Thus, unless the unknown function happens to be a polynomial of exactly the same order used in the nonparametric approximation (in which case the bias term will be exactly zero), there will always be some approximation or misspecification error. This is why the term $\mathtt{B}_n$ appears in the distributional approximation when nonparametric methods are employed but not when the method of estimation is parametric.

A common mistake in practice is to treat the local polynomial approach as parametric and ignore the bias term, a procedure that leads to invalid inferences in all cases except when the approximation error is so small that can be ignored. When the bias term is zero, the approximate distribution of the RD estimator is $\hat{\tau}_n \overset{as}{\sim} \mathcal{N}(\tau^{\mathtt{RD}}, \mathtt{V}_n)$ and confidence intervals are simply $I^{\mathtt{conv}} = \left[\ \hat{\tau}_n \pm 1.96 \cdot \sqrt{\mathtt{V}_n} \ \right]$, the same as in parametric least-squares estimation. Thus, using conventional confidence intervals is equivalent to assuming that the chosen polynomial gives an exact approximation of the true functions $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$. Since these functions are fundamentally unobservable, this assumption is not verifiable and will rarely be credible. Thus we strongly discourage researchers from using conventional inference when using local polynomial methods.

A theoretically sound but ad-hoc procedure is to use undersmooothing (p. 629 Imbens and Lemieux 2008). This procedure involves choosing a bandwidth smaller than the MSE-optimal choice, and using conventional CI with this smaller bandwidth. The theoretical justification is that, for bandwidths smaller than the MSE-optimal choice, the bias term will become negligible in large samples. The main drawback of this procedure is that there are no clear and transparent criteria for shrinking the bandwidth below the MSE-optimal value. Some researchers might estimate the MSE-optimal choice and divide by two, others may chose to divide by three, and yet others may decide to subtract a small number ε from it. Although these procedures can be justified in a strictly theoretical sense, they are all *ad-hoc* and can result in lack of transparency and specification searching. Moreover, this general strategy leads to a loss of statistical power because a smaller bandwidth results in less observations used for estimation.

In a recent contribution, Calonico, Cattaneo, and Titiunik (2014b) proposed *robust* confidence intervals that lead to faster rates of decay in the error $\mathtt{B}_n$ of the distributional approximation, and thus lead to valid inferences even when the MSE-optimal bandwidth is used (no undersmoothing is necessary). These confidence intervals are based on bias-correction, a procedure that first estimates the bias term with $\hat{\mathtt{B}}_n$ and then removes this term from the RD point estimator. The derivation of these robust confidence intervals allows the estimated bias term to converge in distribution to a random variable and thus contribute to the distributional approximation of the RD point estimator. This results in an asymptotic variance $\mathtt{V}_n^{\mathtt{bc}}$ that, unlike the variance $\mathtt{V}_n$ used by the conventional approach, incorporates the contribution of the bias-correction step to the variability of the bias-corrected point estimator.

This approach leads to the robust bias-corrected confidence intervals

$$I^{\mathtt{rbc}} = \left[\ \left(\hat{\tau}_n^{\mathtt{RD}} - \hat{\mathtt{B}}_n\right) \pm 1.96 \cdot \sqrt{\mathtt{V}_n^{\mathtt{bc}}} \ \right]$$

which are constructed by subtracting the bias estimate from the local polynomial estimator and using the new variance formula. Note that these confidence intervals are centered around the bias-corrected point estimate,$\hat{\tau}_n^{\mathtt{RD}} - \hat{\mathtt{B}}_n$, not around the uncorrected estimate $\hat{\tau}_n^{\mathtt{RD}}$.

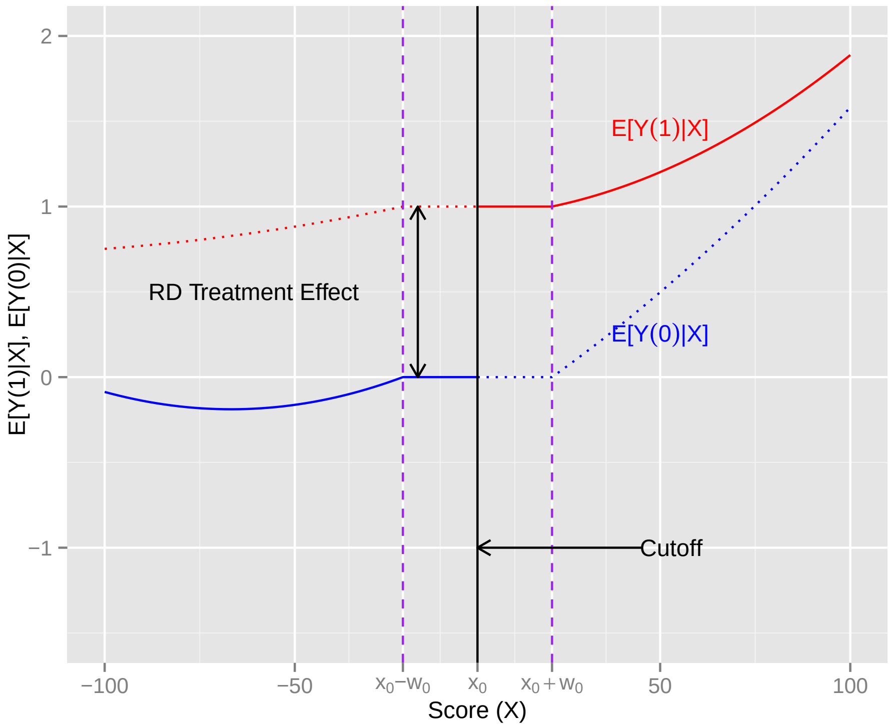
These robust confidence intervals result in valid inferences when the MSE-optimal is used, because they have smaller coverage errors and are therefore less sensitive to tuning parameter choices (see Calonico, Cattaneo, and Farrell 2015).

5 An Alternative Framework for Estimation and Inference: The RD Design as a Local Experiment

An alternative "local randomization" approach to the analysis of RD designs is based directly on the assumption that the RD approximates an experiment near the cutoff (Lee 2008). Our discussion in this section follows Cattaneo, Frandsen, and Titiunik (2015), who proposed to use finite-sample randomization-based inference to analyze RD designs based on a local randomization assumption (see also Cattaneo, Titiunik, and Vazquez-Bare 2015a, who compare RD analysis in continuity-based and randomization-based approaches). For details on the `Stata` implementation of the methods discussed in this section, see Cattaneo, Titiunik, and Vazquez-Bare (2015b). We also provide details in our empirical example in Section 7, and the accompanying `R` and `Stata` code available in the supplementary materials.

When the RD is based on a local randomization assumption, instead of assuming that the unknown regression functions $\mathbb{E}[Y_1|X]$ and $\mathbb{E}[Y_0|X]$ are continuous at the cutoff, the researcher assumes that there is a small window around the cutoff, $W_0 = [x_0 - w_0, x_0 + w_0]$, such that for all units whose scores fall in that window, their placement above or below the cutoff is assigned as in a randomized experiment. As illustrated in Figure 9, under this assumption, the shape of the regression functions $\mathbb{E}[Y_1|X]$ and $\mathbb{E}[Y_0|X]$ is known inside W_0: since placement above or below the cutoff is unrelated to the potential outcomes, the regression functions must be flat.

Figure 9: RD as a Local Experiment



The local randomization assumption is strictly stronger than the continuity assumption, in the sense that if there is a window around x_0 in which the regression functions are flat, then these regression functions will also be continuous at x_0—but the converse is not true. Thus, we must consider the type of applications in which it will be appropriate to use a local randomization approach instead of the standard continuity-based framework. Why would researchers want to impose stronger assumptions to make their inferences? This question is particularly important because, unlike in an experiment, the assignment mechanism in an RD design (a rule that gives treatment based on whether a score exceeds a cutoff) does not logically imply that the treatment is randomly assigned within some window. Like the continuity assumption, the local randomization assumption must be made *in addition* to the assignment mechanism, and is not directly testable.

In order to see in what type of situations the stronger assumption of local randomization is appropriate, it is useful to remember that the local-polynomial approach, although based on the weaker condition of continuity, necessarily relies on extrapolation. In contrast to the local randomization condition which imposes flatness, the assumption of continuity at the cutoff does not imply a specific functional form of the regression functions near the cutoff. This makes the continuity assumption more appealing if there are enough observations near the cutoff to approximate the shape of the regression functions with reasonable accuracy—but possibly inadequate when the number of observations is small. In cases where the sample size around the cutoff is sparse, a continuity-based approach may result in untrustworthy results. In these cases, the local randomization approach will have the advantage that, by relying on the observations very close to the cutoff, it will require minimal extrapolation.

Another situation in which a local randomization approach may be preferable is when the running variable is discrete—i.e., when the set of values that the score can take is countable. Examples include age measured in years or days, population counts or vote totals. By definition, a discrete running variable will have mass points: multiple units will have a score of the same value. In contrast, when the score is continuous, the probability that two units have the same score value is zero. Examples of continuous running variable include vote shares, income, poverty rates, etc. The continuity-based approach described above requires a continuous running variable and is not applicable when the running variable is discrete. The local randomization approach is a natural approach to RD analysis when this happens. Assuming that X_i takes values $x_1, x_2, \ldots, x_k, x_0, x_{k+1}, x_{k+2}, \ldots$, and all units with score greater than or equal to x_0 are treated, the smallest possible window around x_0 that can be selected is $[x_k, x_0]$. Assuming that the parameter $\mathbb{E}[Y_1|X = x_0] - \mathbb{E}[Y_0|X = x_k]$ is of interest, a local-randomization approach can be used to base inferences on a comparison of all control units with $X_i = x_k$ to all treated units with $X_i = x_0$.

Estimation and Inference Within the Window

Adopting a local-randomization approach to RD analysis means that, inside the window W_0 where the treatment is assumed to be randomly assigned, we can analyze the data as we would analyze an experiment. We discuss two possible approaches to estimation and inference, one that is appropriate when the number of observations inside W_0 is large enough, and another that is valid even if the number of observations in W_0 is so small as to render conventional large-sample approximations invalid. Unlike the local-polynomial approach discussed in Section 4, these approaches view the potential outcomes as non-stochastic, and rely on the random assignment of treatment to construct confidence intervals and hypothesis tests. Our discussion assumes that W_0 is known, but we provide guidance on window selection methods at the end of this section.

Large-sample approximation approach.

We first discuss how to make inferences based on large-sample approximations. This approach, which is the most frequently chosen in the analysis of experiments, is appropriate to analyze RD designs under a local randomization assumption when the number of observations inside W_0 is large enough to ensure that these approximations are similar to the finite-sample distributions of the test-statistics of interest. In this case, a natural parameter of interest is the (finite-sample) average treatment effect inside the window,

$$\tau_{\texttt{LRRD}} = \overline{Y}(1) - \overline{Y}(0),$$

where $\overline{Y}(1) = \frac{1}{N}\sum_{i:X_i \in W_0} Y_i(1)$ and $\overline{Y}(0) = \frac{1}{N}\sum_{i:X_i \in W_0} Y_i(0)$ are the average potential outcomes and N is the total number of units inside the window. In this definition, we have assumed that the potential outcomes are non-stochastic. Note that the parameter $\tau_{\texttt{LRRD}}$ is different from the conventional RD parameter $\tau_{\texttt{RD}}$ defined in Section 4: while the former is an average effect inside an interval (W_0), the latter is an average at a single point where, by construction, the number of observations is zero. Thus, the decision to adopt a continuity-based approach versus a randomization-based approach directly affects the definition of the parameter of interest. Naturally, if the window W_0 is extremely small, $\tau_{\texttt{LRRD}}$ and $\tau_{\texttt{RD}}$ become more conceptually similar.

The effect $\tau_{\texttt{LRRD}}$ can be easily estimated by the difference between the average observed outcomes in the treatment and control groups, $\hat{\tau}^{\texttt{LLRD}} = \overline{Y}_t - \overline{Y}_c$, where $\overline{Y}_t = \frac{1}{N_{t,W_0}} \sum_{i:X_i \in W_0} Y_i \mathbb{1}\left(X_i \geq x_0\right)$ and $\overline{Y}_c = \frac{1}{N_{c,W_0}} Y_i \mathbb{1}\left(X_i < x_0\right)$ are the average treated and control observed outcomes and N_{t,W_0} and $N_{c,W_0} = N_{W_0} - N_{t,W_0}$ are, respectively, the number of treated and control units inside W_0. Under the assumption of complete randomization inside W_0 with N_{t,W_0} assigned to treated and $N_{W_0} - N_{t,W_0}$ to control, this observed difference-in-means is an unbiased estimator of $\tau_{\texttt{LRRD}}$. A conservative estimator of the variance of $\tau_{\texttt{LRRD}}$ is given by the sum of the sample variance in each group, $\widehat{\mathbb{V}} = \frac{s_t^2}{N_{t,W_0}} + \frac{s_c^2}{N_{c,W_0}}$, where $s_j^2 = \frac{1}{N_{j,W_0}-1} \sum_{i:X_i \in W_0} \left(Y_i \mathbb{1}\left(X_i \geq x_0\right) - \overline{Y}_j\right)^2$ for $j = \{t, c\}$.

A confidence $1 - \alpha$ interval can be constructed in the usual way relying on a normal large-sample approximations, $I^{\texttt{LR}} = \left[\hat{\tau}^{\texttt{LLRD}} \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathbb{V}}}\right]$. Testing of the null hypothesis that the average treatment effect is zero can also be based on normal approximations. Letting $H_0^{\texttt{N}} : \overline{Y}(1) - \overline{Y}(0) = 0$, we can construct a usual t-statistic using the point and variance estimators just introduced, $t = \frac{\overline{Y}_t - \overline{Y}_c}{\sqrt{\widehat{\mathbb{V}}}}$. Using, for example, a two-sided test, the p-value associated with a test of $H_0^{\texttt{N}}$ is $2\left(1 - \Phi(t)\right)$, where $\Phi(\cdot)$ is the Normal CDF.

If, instead, we see the units inside W_0 as a random sample from a (large) super-population, the potential outcomes within W_0 become stochastic by virtue of the random sampling and the parameter of interest is the super-population average treatment effect, $\mathbb{E}[Y_i(1) - Y_i(0) | X_i \in W_0]$. Adopting this super-population perspective, however, does not change the estimation or inference procedures discussed above (see Imbens and Rubin 2015, Chapter 6 for details).

Finite-sample approach

The inference procedures described above rely on large-sample approximations, and are most appropriate when the sample size inside the window is sufficiently large. In many cases, however, the number of observations within W_0 will be very small. The reason is that, in most RD applications, there is a tension between the plausibility of the local randomization assumption and the length of the window around the cutoff where this assumption is invoked: the smaller the window, the more similar the values of the score for units inside the window, and the more credible the local randomization assumption tends to be. Since a small window will tend to have a small number of observations, the large-sample methods described above will often be unreliable.

An alternative proposed by Cattaneo, Frandsen, and Titiunik (2015) is to use a finite-sample Fisherian framework, which leads to correct inferences for any sample size because it is finite-sample exact. The Fisherian approach is similar to the Neyman approach in that the potential outcomes are seen as fixed, but it differs in several important ways. In Fisherian inference, the total number of units in the study is seen as fixed, and inferences do not rely on assuming that this number is large. Moreover, the null hypothesis of interest in the Fisherian approach is not that the average treatment effect is zero, as in the Neyman approach. Instead, the Fisherian null hypothesis is that the treatment has no effect for any unit, $H_0^{\mathtt{F}} : Y_i(0) = Y_i(1) \text{ for all } i$, also known as the *sharp null hypothesis*. For detailed discussions on the randomization-based Fisherian framework, see Bowers, Fredrickson, and Panagopoulos (2013), Imbens and Rosenbaum (2005), Keele, McConnaughy, and White (2012), and Rosenbaum (2002, 2010).

This framework allows us to make inferences that are correct for any sample size because, under $H_0^{\mathtt{F}}$, both potential outcomes can be imputed for every unit and there is no missing data. Under the sharp null, $Y_i(1) = Y_i(0) = Y_i$, and the observed outcome of each unit is equal to both unit's potential outcomes. When the treatment assignment is known, the fact that all potential outcomes are observed under the null hypothesis allows us to derive the null distribution of any test statistic of interest from the randomization distribution of the treatment assignment. Since the latter distribution is finite-sample exact, the Fisherian framework allows researchers to make inferences without relying on large-sample approximations.

Applying this framework to RD analysis thus requires, in addition to knowledge of W_0, knowledge of the specific way in which the treatment was randomized—that is, knowledge of the distribution of the treatment assignment. In practice, the latter will not be known, but can be accurately approximated by assuming a complete randomization in W_0 with with N_{t,W_0} assigned to treated and $N - N_{t,W_0}$ to control. We let $\mathbf{Z}$ be the treatment assignment for the N units in W_0, and collect in the set Ω all the possible treatment assignments that can occur given the assumed randomization mechanism. In a complete randomization, Ω includes all vectors of length N_{W_0} such that each vector has N_{t,W_0} ones and $N_{W_0} - N_{t,W_0}$ zeros. We also need to choose a test statistic, which we denote

$t(\mathbf{Z}, \mathbf{Y})$, that is a function of the treatment assignment $\mathbf{Z}$ and the vector $\mathbf{Y}$ of observed outcomes for the N units in the experiment. Of all the possible values of the treatment vector $\mathbf{Z}$ that can occur, only one will have occurred in W_0; we call this value the observed treatment assignment, $\mathbf{z}^{\mathsf{obs}}$, and we denote T^{obs} the observed value of the test-statistic associated with $\mathbf{z}^{\mathsf{obs}}$.

Then, the one-sided finite-sample exact p-value associated with a test of the sharp null hypothesis $H_0^{\mathtt{F}}$ is the probability that the test-static exceeds its observed value:

$$p^{\mathtt{F}} = \Pr(t(\mathbf{Z}, \mathbf{Y}) \geq T^{\mathsf{obs}}) = \sum_{\mathbf{z} \in \Omega} \mathbf{1}\left\{t(\mathbf{z}, \mathbf{Y}) \geq T^{\mathsf{obs}}\right\} \cdot \Pr(\mathbf{Z} = \mathbf{z}).$$

When each of the treatment assignments in Ω is equally likely, this expression simplifies to the number of times the test-statistic exceeds the observed value divided by the total number of test-statistics that can possibly occur, $p^{\mathtt{F}} = \Pr(t(\mathbf{Z}, \mathbf{Y}) \geq T^{\mathsf{obs}}) = \frac{\aleph\left\{t(\mathbf{z}, \mathbf{Y}) \geq T^{\mathsf{obs}}\right\}}{\aleph\{\Omega\}}$, where $\aleph\{\cdot\}$ denotes the number of elements in a set. Note also that, under $H_0^{\mathtt{F}}$ we have $\mathbf{Y} = \mathbf{Y}(1) = \mathbf{Y}(0)$, so that $t(\mathbf{Z}, \mathbf{Y}) = t(\mathbf{Z}, \mathbf{Y}(0))$. Thus, the only randomness in $t(\mathbf{Z}, \mathbf{Y})$ comes through the random assignment of the treatment. Confidence intervals can be obtained by inverting these hypothesis tests. For example, under the constant treatment effect model $Y_i(1) = Y_i(0) + \tau$, a $1 - \alpha$ confidence interval for τ can be obtained by collecting the set of all the values τ_0 that fail to be rejected with an α-level test of the hypothesis $H_0 : \tau = \tau_0$. See Cattaneo, Frandsen, and Titiunik (2015) and Rosenbaum (2002) for details.

Window selection when W_0 ***is unknown.*** In most RD applications, the window where a local randomization assumption is plausible will be unknown. Cattaneo, Frandsen, and Titiunik (2015) proposed a window selection procedure based on a series of nested balance tests on important predetermined covariates, where the chosen window is the window such that covariate balance holds in that window and all windows contained in it. Due to space limitations, we discuss this procedure in more detail in Section S1 of the Supplemental Appendix. Additional details can also be found in Cattaneo, Frandsen, and Titiunik (2015), and Cattaneo, Titiunik, and Vazquez-Bare (2015a,b).

6 How to validate an RD design

A main advantage of the RD design is that the mechanism by which treatment is assigned is known and based on observable quantities, giving researchers an objective basis to distinguish pre-treatment from post-treatment variables and to identify qualitative information regarding the treatment assignment process that can be helpful to justify assumptions. However, a known rule that assigns treatment based on whether a score exceeds a cutoff is not by itself enough to guarantee that the assumptions needed to recover the causal effect of interest are met. For example, a scholarship may be assigned based on whether the grade students receive on a test is above a cutoff, but if the cutoff is known to the students' parents and there are mechanisms to appeal the grade, this may invalidate the assumption that the average potential outcomes are continuous at the cutoff. If, for example, some parents appeal the grade when their child is barely below the cutoff and are successful in changing their child's score so that it reaches the cutoff, it is likely that those parents who appeal are more involved than average in their children's education. If parents' involvement affects students' future academic achievement then, on average, the potential outcomes of those students above the cutoff may be discontinuously different from the potential outcome of those students below the cutoff.

In general, if the cutoff that determines treatment is known to the units that will be the beneficiaries of the treatment, researchers must worry about the possibility of units actively changing or manipulating the value of their score when they miss the treatment barely. The first type of information that can be provided is whether an institutionalized mechanism to appeal the score exists, and if so, how often it is used to successfully change the score and which units use it. Qualitative data about the administrative process by which scores are assigned, cutoffs determined and publicized, and treatment decisions appealed, is extremely useful to validate the design.

In many cases, however, qualitative information will be limited and the possibility of units manipulating their score cannot be ruled out. Crucially, the fact that there are no institutionalized mechanisms to appeal and change scores does not mean that there are no informal mechanisms by which this may happen. Thus, an essential step in evaluating the plausibility of the RD assumptions is to provide empirical evidence. Naturally, the continuity and local randomization assumptions that guarantee the validity of the RD design are about unobservable quantities and as such are inherently untestable. However, there are empirical implications of these unobservable assumptions that can be expected to hold in most cases and can provide indirect evidence about the validity of the design. We consider three such empirical tests: (i) the density of the running variable around the cutoff, (ii) the treatment effect on pre-treatment covariates or placebo outcomes, and (iii) the treatment effect at alternative non-cutoff values of the score. As we discuss below, the implementation of each of the tests differs according to whether a continuity or a local randomization assumption is invoked.

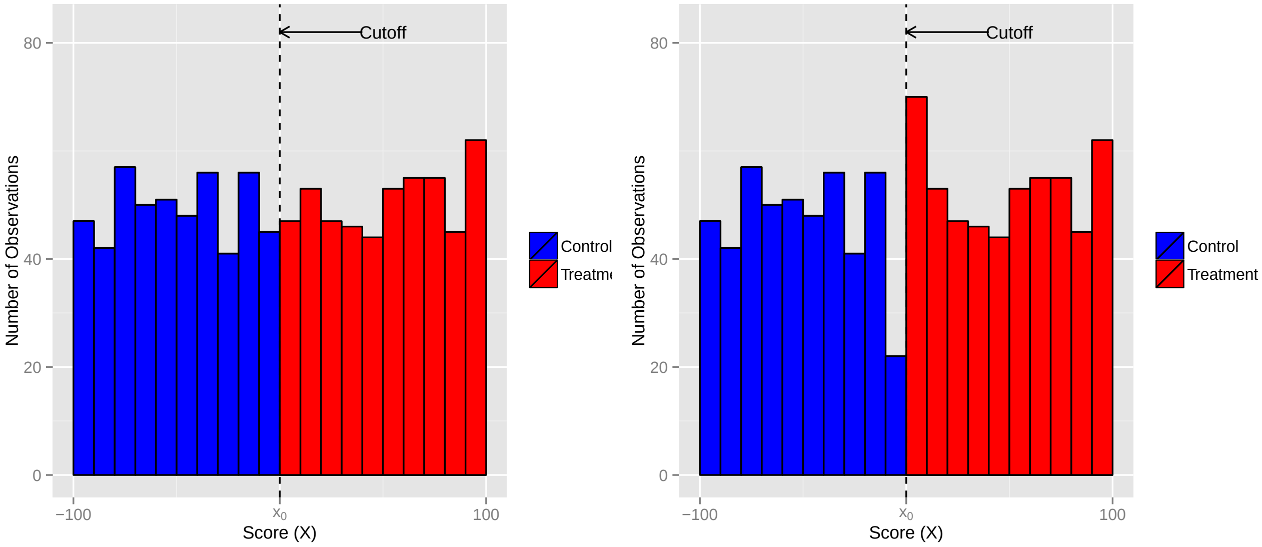
Density of the Running Variable

The first type of falsification test examines whether, in a local neighborhood near the cutoff, the number of observations below the cutoff is considerably different from the number of observations above it. The underlying assumption is that if individuals do not have the ability to precisely manipulate the value of the score that they receive, the number of treated observations just above the cutoff should be approximately similar to the number of control observations below it. Although this assumption is neither necessary nor sufficient for the validity of an RD design, RD applications where there is an unexplained abrupt change in the number of observations right at the cutoff will tend to be less credible. This kind of test, first introduced by McCrary (2008), is often called a density test.

Figure 10 shows a histogram of the running variable in two hypothetical RD examples. In the scenario illustrated in Figure 10(a), the number of observations above and below the cutoff is very similar. In contrast, Figure 10(b) illustrates a case in which the density of the score right below the cutoff is considerably lower than just above it—a finding that is consistent with units systematically increasing the value of their original score so that they are assigned to the treatment group instead of the control.

In addition to a graphical illustration of the density of the running variable, researchers should test the assumption more formally. The implementation of the formal test depends on whether one adopts a continuity-based or a local-randomization-based approach to RD. In the former approach, the null hypothesis is that the density of the running variable is continuous at the cutoff, and its implementation requires the estimation of the density of observations near the cutoff, separately

Figure 10: Histogram of Score



(a) No sorting

(b) Sorting

for observations above and below the cutoff. Cattaneo, Jansson, and Ma (2015a) propose a local polynomial density estimator that does not require pre-binning of the data and leads to size and power improvements relative to other implementation approaches. (implementation in `Stata` is discussed in Cattaneo, Jansson, and Ma 2015b, .)

The implementation is different under the local-randomization approach. In this case, the null hypothesis is that, within the window W_0 where the treatment is assumed to be randomly assigned, the number of the number of observations in the control and treatment groups should be approximately similar to what would be expected in sample of N Bernoulli trials with a pre-specified treatment probability $p \in (0, 1)$. This implies that the number of treated units in W_0 and the number of control units in W_0 should follow a binomial distribution. The null hypothesis of the test is that the probability of success in a series of N Bernoulli experiments is p. As we have discussed, the true probability of treatment is unknown, but in practice $p = 1/2$ is the most natural choice in the absence of additional information. The binomial test is implemented in all common statistical software, and is also part of the `rdlocrand` `Stata` commands (Cattaneo, Titiunik, and Vazquez-Bare 2015b).

Treatment Effect on Predetermined Covariates and Placebo Outcomes

Another important falsification test involves examining whether, near the cutoff, treated units are similar to control units. The idea behind this approach is simply that, if units lack the ability to precisely manipulate the value score they receive, there should be no systematic differences between units with similar values of the score. Thus, except for their treatment status, units just above and just below the cutoff should be similar in all those characteristics that could not have been affected by the treatment. These variables can be divided into two groups: variables that are determined before the treatment is assigned (which we call *predetermined covariates*) and variables that are determined after the treatment is assigned but, according to substantive knowledge about the treatment's causal mechanism, could not possibly have been affected by the treatment (which we call *placebo outcomes*).[1]

Although, once again, the particular implementation of this type of falsification test depends on whether researchers adopt a continuity-based or a local-randomization-based approach, a fundamental principle applies to both: *all predetermined covariates and placebo outcomes should be analyzed in the same way as the outcome of interest*. In the continuity-based approach, this principle means that for each predetermined covariate (and placebo outcome), researchers should first choose an optimal bandwidth, and then use local-polynomial techniques within that bandwidth to estimate the "treatment effect". Since the variable could not have been affected by the treatment, the expectation is that the null hypothesis of no treatment effect will fail to be rejected.

In the local-randomization RD approach, this principle means that the null hypothesis of no (average) treatment effect should be tested within W_0 for all predetermined covariates and placebo outcomes, using the same inference procedures and the same assumptions regarding the treatment assignment mechanism and the same test statistic used for the analysis of the outcome of interest. Note that, since in this approach W_0 is the window where the treatment is assumed to have been randomly assigned, all covariates and placebo outcomes should be analyzed within the window W_0. This illustrates a fundamental difference between the continuity-based and the randomization-based approach: in the former, in order to estimate and test hypotheses about treatment effects

[1] For example, if the treatment is access to clean water and the outcome of interest is child mortality, a treatment effect is expected on mortality due to water-bone illnesses but not on mortality due to other causes. Thus, mortality from road accidents is a placebo outcome in this example.

we need to approximate the unknown functional form of outcomes, predetermined covariates and placebo outcomes, which requires estimating separate bandwidths for each variable analyzed; in the latter, since the treatment is randomly assigned in W_0, all analyses occur within the same window, W_0.

Treatment Effect For Alternative Cutoffs

A final falsification test we mention briefly is to replace the true cutoff value with another value (a value at which the treatment status does not really change) and perform estimation and inference using this "fake" cutoff. The expectation is that a significant treatment effect should occur only at the true cutoff value and not at other values of the score where the treatment status is constant. A graphical implementation of this falsification can be done by creating the RD plots described in Section 3 and observing that there are no jumps in the observed regression functions at points other than the true cutoff.

A formal implementation of this idea is to repeat the entire analysis using the new, fake value of the cutoff. Once again, the implementation depends the approach adopted. In the continuity-based approach, we would use local-polynomial methods within an optimally-chosen bandwidth around the fake cutoff to estimate treatment effects on the true outcome. In the local-randomization approach, we would choose a window around the fake cutoff where randomization is plausible, and make inferences for the true outcome within that window.

7 Empirical Illustration

In this section, we use an example with real political science data to illustrate our previous discussion. Our example comes from Fouirnaies and Hall (2014), who study the effect of partisan incumbency in a district on future campaign contributions, which they call "the financial incumbency effect." Using data from the U.S. House and state legislatures, they find that parties that narrowly win in a district at time t command a much greater share of campaign contributions in that district at time $t + 1$ than parties that narrowly lost at time t.

Fouirnaies and Hall conduct several analyses in the paper. For simplicity, we focus on replicating only their analysis for state legislative districts. The paper is clear about its assumptions and already follows several of our recommendations, so we make only minor changes to the analysis that they present. As part of the supplemental materials accompanying this manuscript, we provide code to replicate these analyses in both `R` and Stata.

The state legislative data is notable because of its large sample size (32,670 observations), which translates into a high density of observations even close to the cutoff. Not all analysts will be so lucky and, as we discussed, having sparse data—especially close to the cutoff—can make some inference methods less reliable.

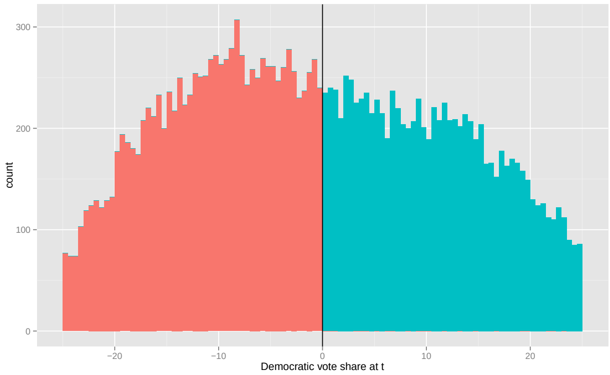
The first step in RD analysis is to *identify a triplet of running variable, cutoff, and outcome*. Fouirnaies and Hall, like many others who study the incumbency advantage, choose Democratic margin of victory in an election at time t as the running variable or score. Margin of victory is defined as the difference between vote percentage obtained by the Democratic party minus the vote percentage obtained by its strongest opponent. Specifying this variable is specified as a percentage ensures that it is continuous. The Democratic party wins the election when its margin of victory is positive and loses when it is negative. Thus, the cutoff is zero. The outcome of interest we consider is the Democratic share of total contributions in the following election cycle ($t + 1$)—the authors test other outcomes in the paper.

The second step is to *validate the design*, followed by estimation and inference of the effects of interest. As discussed, the implementation of the analysis varies according to whether a continuity-based or local-randomization-based approach is adopted. We illustrate each method in turn.

Continuity-based Analysis

We start by conducting a continuity-based analysis. We validate the design analyzing the density of the running variable near the cutoff and estimating treatment effects for predetermined covariates and placebo outcomes. We use the local polynomial density estimators proposed developed by Cattaneo, Jansson, and Ma (2015a) to test the null hypothesis that the density of the Democratic margin of Victory at t is continuous at the cutoff—see (Cattaneo, Jansson, and Ma 2015b) for the `Stata` implementation. The resulting p-value is 0.7648, meaning we have no evidence that the density is discontinuous at the cutoff. We also plot a fine histogram of the data

Figure 11: Histogram of Democratic Margin of Victory at t in 0.5 percentage-point bins



(binned by 0.5 percentage points) near the cutoff in Figure 11. Visual inspection of this figure supports the results of the density test. Thus, based on our density analysis, there is no evidence that the Democratic party is able to barely win more often than it barely loses.

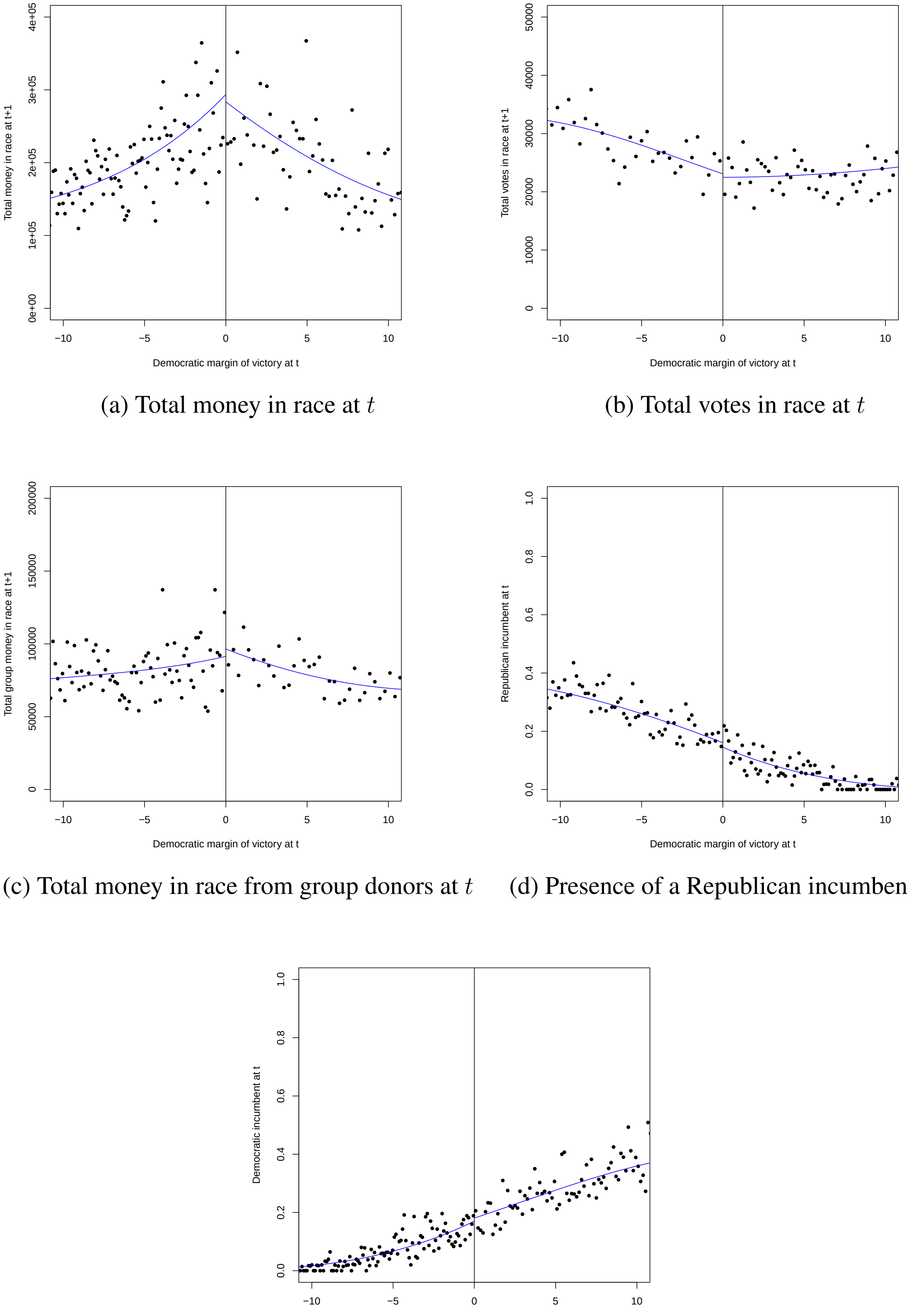
The next step to validate the design is to *analyze predetermined covariates as we do outcomes*. The choice of covariates may be limited by data availability, but it is desirable to examine covariates that are known to be related to both the treatment and the outcome. Most of the covariates in the Fouirnaies and Hall replication data have to do with electoral returns in the district or campaign contributions. Some of the best covariates in the dataset for investigating the validity of the design are the total money and the total number of votes in the race, which will vary both with district size and with turnout. We also analyze whether there was a Democratic or Republican incumbent at time t, as well as the total amount of money spent in the race by organized groups. We analyze each of these predetermined covariates in the same way we will analyze the outcome variable: using local linear estimation within a MSE-optimal bandwidth, with robust confidence intervals for inference. We use the `rdrobust` command, available both in `R` and `Stata`, for both bandwidth selection and local polynomial estimation and inference. We use the CCT implementation of the MSE-optimal bandwidth (Calonico, Cattaneo, and Titiunik 2014b) as this has better properties than the IK (Imbens and Kalyanaraman 2012) implementation—see Calonico, Cattaneo, and Titiunik (2014b,a) for details.

Table 1 shows the results from testing these covariates. There is no evidence of a treatment effect for any covariate—none of the estimated effects are significantly different from zero when robust p-values are considered. We also show RD plots for each variable in Figure 12, where we can see that the covariates do not change abruptly around the cutoff. These results suggest that the RD continuity assumptions are likely met, and give credibility to the design.

Table 1: Testing covariates as placebo outcomes. Results presented are from robust estimation using CCT bandwidths.

				Robust Inference
Covariate	Bandwidth	Point Estimator ($\hat{\tau}^{\mathtt{RD}}$)	p-val	95% CI
Total money	10.84	-8,531.3	0.57	[-70,969.49, 39,131.71]
Total money from groups	11.28	2,455.6	0.77	[-12568.33, 16,883.05]
Total votes	9.08	-1,008	0.52	[-4,158.57, 2,094.49]
Democratic incumbent	7.861	0.01188	0.83	[-0.036911, 0.045788]
Republican incumbent	6.96	-0.01902	0.61	[-.052013 , .030615]

Figure 12: Predetermined Covariates



(a) Total money in race at t

(b) Total votes in race at t

(c) Total money in race from group donors at t

(d) Presence of a Republican incumbent at t.

(e) Presence of a Democratic incumbent at t.

Having found no (continuity-based) evidence that the RD design is invalid, we proceed to perform estimation and inference for the outcome of interest—the share of the district's total campaign contributions at $t+1$ received by the Democratic party. As we did for the predetermined covariates, we use local-polynomial methods with CCT MSE-optimal bandwidth. The results of the estimation are reported in Table 2. In this first row, we report results from local-linear estimation. The CCT bandwidth selector chooses a MSE-optimal bandwidth of 8.97 percentage points for point-estimation of $\tau^{\texttt{RD}}$ with a polynomial of order one.

As Fouirnaies and Hall (2014) do, we find a large and strong treatment effect. Using only observations with score within $[-8.97, 8.97]$ weighted by triangular kernel, the point estimator is 20.53. Substantively, this means that the Democratic party's share of the total campaign contributions in the district increases by 20 percentage points after a narrow win compared to after a narrow loss.

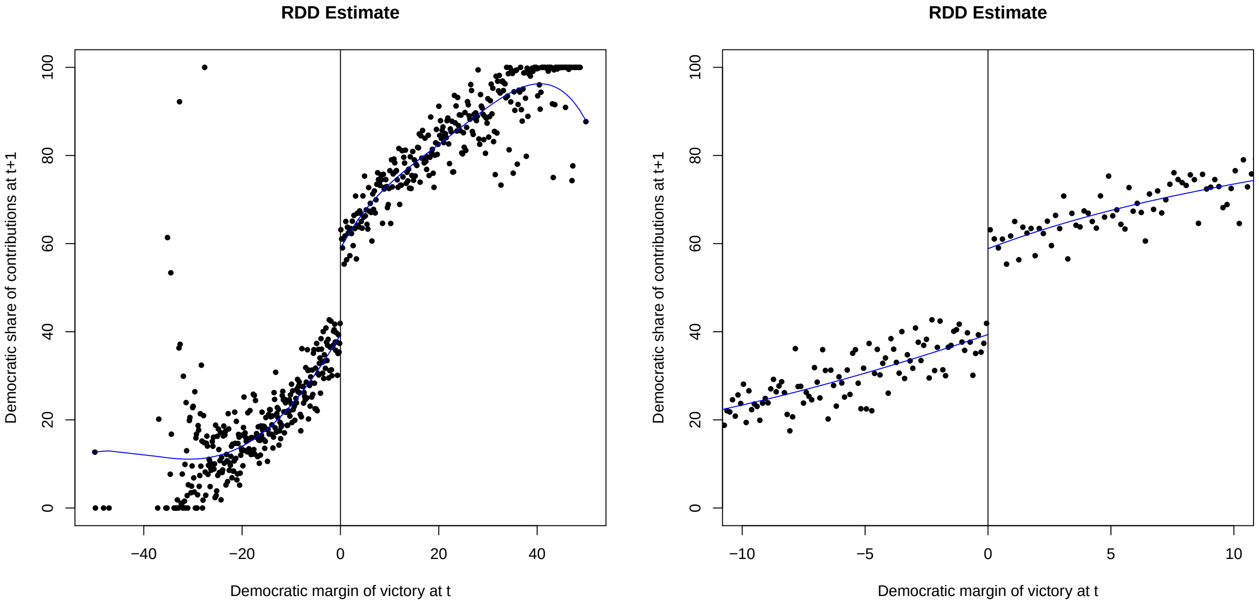
For illustration purposes, in the second panel of Table 2 we estimate the effect using a local quadratic polynomial. The CCT MSE-optimal bandwidth is 16.26, almost twice as large as the bandwidth for local-linear estimation. This difference is expected. The 8.97 bandwidth associated with the local-linear specification optimally balances bias and variance for a linear approximation. A polynomial of higher order improves the approximation and hence reduces bias; thus, keeping the bandwidth at 8.97 when the polynomial order is increased would be suboptimal: when the bandwidth increases to 16.26, the bias increases and the variance decreases, restoring optimality. As we can see in Table 2, the results are extraordinarily robust to the polynomial specification: in both the local-linear and local-quadratic specifications, the point estimator is about 20 percentage-points and the robust 95% confidence intervals range approximately from 17 to 23 percentage-points.

Table 2: Local polynomial estimation with CCT MSE-optimal bandwidth

			Conventional Inference		Robust Inference	
Polynomial Order	Bandwidth	Point Estimator ($\hat{\tau}^{\texttt{RD}}$)	p-val	95% CI	p-val	95% CI
1	8.97	20.53	0.00	[17.97, 23.09]	0.00	[17.12, 22.88]
2	16.26	19.93	0.00	[17.13, 22.72]	0.00	[16.74, 23.11]

We plot these effects in Figure 13. For visual clarity, we plot the entire data range as well as only the data close to the cutoff, within 10 percentage points above and below the cutoff. The plots are easily generated with the `rdplot` function from the `rdrobust` package in `R` and `Stata`. It is clear from the plots that the jump at the cutoff is large and sharp—this is a large and robust effect.

Figure 13: The Financial Incumbency Advantage



(a) All Races

(b) Races decided by 10 points or less

The final continuity-based robustness check we do is to use *placebo cutoffs*, where we vary the cutoff slightly and re-estimate the main RD effect. This test should result in an RD effect that is not significantly different from zero, as we are using false cutoffs in the running variable that do not reflect the actual cutoff that assigns treatment. In Table 3, we show that for two placebo cutoffs, the point estimator is more than ten times smaller in absolute value than the true effect and is statistically indistinguishable from zero, another important validation for the design.

Table 3: Local-linear RD estimation and inference at placebo cutoffs

			Conventional Inference		Robust Inference	
Placebo Cutoff	Bandwidth	Point Estimator ($\hat{\tau}^{\mathtt{RD}}$)	p-val	95% CI	p-val	95% CI
1	3.536	-0.759	0.72	[-4.912, 3.393]	0.42	[-6.200 , 2.595]
-3	4.566	-1.932	0.29	[-5.505, 1.642]	0.17	[-6.650, 1.206]

7.1 Local-randomization Approach

We now illustrate how to analyze an RD design based on the assumption of local randomization. This can be implemented using the `rdlocrand` package in `Stata` and with `R` code provided in the supplemental materials.

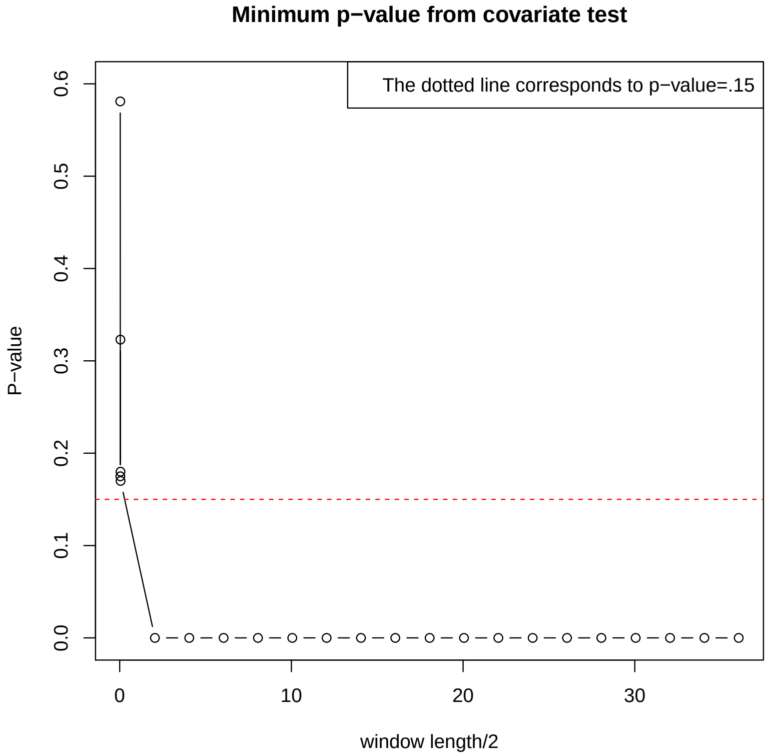
The first step in the local-randomization-based approach is to *select a window* based on several covariates. As above, we recommend using crucial covariates that are believed to be strongly related to both the running variable and the outcome of interest. If a continuity-based approach was also used for analysis, we recommend using the same covariates that were used in the continuity-based approach to validate the design.

Our window selection is based on the five covariates that we tested above: total money in the race at $t+1$, total votes cast in the race at $t+1$, whether there was a Democratic incumbent at t, whether there was a Republican incumbent at t, and the total money in the race that came from groups. The window selection algorithm, implemented in the `rdwinselect` command in Stata, searches for the largest window around the cutoff in which chosen covariates are balanced. We recommend using $p < 0.15$ as the threshold for window selection, which is the default opinion in `rdwinselect`. This implies that the window selected will be the largest window such the p-values associate with covariate balance tests are no smaller than 0.15 for the chosen window and all smaller windows contained in it.

Figure 14 helps show the intuition for the window selector. As the window selector evaluates wider windows, moving right across the figure, the p-values get smaller, indicating less balance in the covariates. See Cattaneo, Frandsen, and Titiunik (2015) and section S1 of the Supplemental Appendix for more details on window selection. In the state legislative level data we re-analyze, the window selector chooses an optimal window of $W_0 = [-0.052, 0.052]$. This leaves us with 45 observations, 29 below the cutoff and 16 above the cutoff.

We then use the `rdrandinf` function to use randomization inference to perform estimation and inference within the chosen window. Even in this very small window, we estimate a similar effect size, 24.33 percentage points. Using the Fisherian framework, we reject the sharp null hypothesis that the treatment has no effect for any unit (p-value 0.004), and we obtain the 95% confidence interval [7.65, 40.95] by hypothesis test inversion under a constant treatment effect model.

Figure 14: Minimum p-values for covariates at various windows evaluated by the window selection algorithm.



We arrive at a similar conclusion using a Neyman framework: we reject the null hypothesis that the average share of contributions received by the Democratic party at $t+1$ is the same in districts where the Democratic party won at t as in districts where it lost at t. The confidence interval for this average treatment effect, [7.76 ,40.90], is very similar to the Fisherian confidence interval.

Table 4: Local randomization estimates for main outcome variable

Chosen Window	Sample Size		Point Estimator ($\hat{\tau}^{\mathtt{LLRD}}$)	Fisherian Inference		Neyman Inference	
	Tr	Co		p-val	95% CI	p-val	95% CI
[-0.052, 0.052]	16	29	24.332	0.004	[7.65, 40.95]	0.004	[7.76 ,40.90]

The analysis plan that we have demonstrated with the Fouirnaies and Hall (2014) data is appropriate for nearly all RD applications. The availability of covariates may affect some analysts' ability to follow all of our recommended steps, but we encourage researchers to follow this recommended set of practices closely.

8 A Summary of Recommendations for Practice

We conclude with some general recommendations for practice. We hope these broad guidelines for the empirical analysis of RD designs will contribute to the transparency, replicability and comparability of RD applications in political science.

1. The first step is to make sure that the empirical application being considered fits the definition of an RD design: the design must have a score, a cutoff, a treatment that is assigned completely or partially based on whether the score exceeds the cutoff, and an outcome of interest that is measured at the same level at which the treatment is assigned (or lower).

2. Once a proper RD application design is identified, researchers should focus on validation and falsification of the RD assumptions. We suggest the following steps:

    - Collect qualitative evidence about the treatment assignment mechanism. How was the score decided? Did the units know the cutoff value and that this cutoff would be used as the basis for treatment assignment? Were there mechanisms for units to appeal the value of the score that was initially assigned to them?
    - Create an RD plot for the outcome(s) of interest and show that there are no jumps in the outcome except at the cutoff. If the outcome does jump abruptly at other values, researchers should investigate whether another intervention was assigned based on the score—in cases where other treatments co-exist with the treatment of interest, the interpretation of the RD effect will be considerably more complicated.
    - Perform density tests, both graphically and with formal statistical tests.
    - Look at the treatment effect on important predetermined covariates and, if available, placebo outcomes. For each of the variables considered, the analysis should be analogous to the way in which the main outcome of interest will be analyzed. For example, if a continuity-based approach is adopted and the effect on the outcome of interest will be estimated with local polynomials, the effects on the covariates should also use local polynomials, treating each covariate as an outcome. Unlike in the local randomization approach to RD, the bandwidth is not the region around the cutoff where the treatment

is as-if randomly assigned; thus, it is not appropriate to use the same bandwidth for the outcome of interest as for the covariates. The bandwidth is outcome-specific, and as such one different bandwidth should be estimated for each main outcome, predetermined covariate and placebo outcome considered.

In contrast, if a local randomization assumption is adopted, all covariate tests should be carried out in the same window W_0 where the local randomization is invoked. In fact, as we saw, a recommended method to choose this window is by detecting the region where no relationship can be detected between the treatment and the covariates and/or placebo outcomes.

If these falsification steps "fail", in the sense that they show abrupt changes in the number of observations and the characteristics of units at the cutoff, the application will lack credibility (Caughey and Sekhon 2011, for an example, see). The most credible applications are those in which there are no concerns raised during the falsification stage.

3. After successful falsification, analyze the outcome of interest, both graphically via an RD plot and formally with appropriate estimation and inference methods. If the number of observations near the cutoff is large enough, polynomial estimation and inference based on continuity assumptions will be appropriate. We advise against using global polynomials; instead, we recommend using nonparametric, local polynomial methods. The bandwidth required in local polynomial estimation and inference should not be chosen in an ad-hoc way, as this reduces the transparency of the analysis study. Instead, we recommend choosing a MSE-optimal bandwidth using a data-driven and automated method, and then reporting a local-polynomial point estimator and robust confidence intervals using that bandwidth.

   If the number of observations near the cutoff is very small or the running variable is not continuous, we recommend using a local randomization approach. Even when the number of observations near the cutoff is large enough, a local randomization approach can be employed in addition to local polynomial methods as a robustness check *if the local randomization is plausible*.

4. The features of the RD design implicitly define the unit of analysis. If the treatment is as-

signed at a higher level than the outcome is observed (e.g., counties are assigned a treatment and researcher has survey data on individuals within each county) , the first step in the analysis should be to aggregate the outcome up to the level at which the treatment is assigned, and carry out the analysis at that level.

References

Angrist, Joshua, and Miikka Rokkanen. 2016. “Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff.” *Journal of American Statistical Association*, forthcoming.

Bowers, Jake, Mark M Fredrickson, and Costas Panagopoulos. 2013. “Reasoning About Interference Between Units: A General Framework.” *Political Analysis* 21 (1): 97–124.

Calonico, Sebastian, Matias D. Cattaneo, and Max H. Farrell. 2015. “On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference.” Working Paper, University of Michigan.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014a. “Robust Data-Driven Inference in the Regression-Discontinuity Design.” *Stata Journal* 14 (4): 909–946.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014b. “Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs.” *Econometrica* 82 (6): 2295–2326.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2014c. “`rdrobust`: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs.” *R Journal* 7 (1): 38–51.

Calonico, Sebastian, Matias D. Cattaneo, and Rocío Titiunik. 2016. “Optimal Data-Driven Regression Discontinuity Plots.” *Journal of American Statistical Association*, forthcoming.

Card, David, David S. Lee, Zhuan Pei, and Andrea Weber. 2015. “Inference on Causal Effects in a Generalized Regression Kink Design.” *Econometrica,* forthcoming.

Cattaneo, Matias D., Brigham Frandsen, and Rocío Titiunik. 2015. “Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate.” *Journal of Causal Inference* 3 (1): 1–24.

Cattaneo, Matias D., Luke Keele, Rocío Titiunik, and Gonzalo Vazquez-Bare. 2015. “Interpreting Regression Discontinuity Designs with Multiple Cutoffs.” Working Paper, University of Michigan.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2015a. "Simple Local Regression Distribution Estimators with an Application to Manipulation Testing." Working Paper, University of Michigan.

Cattaneo, Matias D., Michael Jansson, and Xinwei Ma. 2015b. "`rddensity`: Manipulation Testing Based on Density Discontinuity." Working Paper, University of Michigan.

Cattaneo, Matias D., Rocío Titiunik, and Gonzalo Vazquez-Bare. 2015a. "Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality." Working Paper, University of Michigan.

Cattaneo, Matias D., Rocío Titiunik, and Gonzalo Vazquez-Bare. 2015b. "`rdlocrand`: Inference in Regression Discontinuity Designs under Local Randomization." Working Paper, University of Michigan.

Caughey, Devin, and Jasjeet S Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close US House Races, 1942–2008." *Political Analysis* 19: 385–408.

Eggers, Andrew, Anthony Fowler, Jens Hainmueller, Andrew B Hall, and James M Snyder. 2015. "On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from over 40,000 Close Races." *American Journal of Political Science* 59 (1): 259–274.

Fan, J., and I. Gijbels. 1996. *Local Polynomial Modelling and Its Applications*. New York: Chapman & Hall/CRC.

Fouirnaies, Alexander, and Andrew B Hall. 2014. "The Financial Incumbency Advantage: Causes and Consequences." *Journal of Politics* 76 (03): 711–724.

Frandsen, Brigham, Markus Frölich, and Blaise Melly. 2012. "Quantile Treatments Effects in the Regression Discontinuity Design." *Journal of Econometrics* 168 (2): 382–395.

Gelman, Andrew, and Guido Imbens. 2014. "Why High-order Polynomials Should Not be Used in Regression Discontinuity Designs." NBER Working Paper No. 20405.

Hahn, Jinyong, Petra Todd, and Wilbert van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69 (1): 201–209.

Hopkins, Daniel J. 2011. "Translating into Votes: The Electoral Impacts of Spanish-Language Ballots." *American Journal of Political Science* 55 (4): 814–830.

Imbens, Guido, and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142 (2): 615–635.

Imbens, Guido W., and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press.

Imbens, Guido W., and Karthik Kalyanaraman. 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *Review of Economic Studies* 79 (3): 933–959.

Imbens, Guido W., and Paul Rosenbaum. 2005. "Robust, Accurate Confidence Intervals with a Weak Instrument: Quarter of Birth and Education." *Journal of the Royal Statistical Society, Series A* 168: 109–126.

Keele, Luke, Corrine McConnaughy, and Ismail White. 2012. "Strengthening the Experimenter's Toolbox: Statistical Estimation of Internal Validity." *American Journal of Political Science* 56 (2): 484–499.

Keele, Luke J., and Rocío Titiunik. 2015. "Geographic Boundaries as Regression Discontinuities." *Political Analysis* 23 (1): 127–155.

Keele, Luke, Rocío Titiunik, and José R Zubizarreta. 2015. "Enhancing a Geographic Regression Discontinuity Design through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178 (1): 223–239.

Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142 (2): 675–697.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48 (2): 281–355.

McCrary, Justin. 2008. “Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test.” *Journal of Econometrics* 142 (2): 698–714.

Papay, John P., John B. Willett, and Richard J. Murnane. 2011. “Extending the Regression-Discontinuity Approach to Multiple Assignment Variables.” *Journal of Econometrics* 161 (2): 203-207.

Rosenbaum, Paul R. 2002. *Observational Studies*. 2nd ed. New York: Springer.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York: Springer.

Thistlethwaite, Donald L., and Donald T. Campbell. 1960. “Regression-discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment.” *Journal of Educational Psychology* 51 (6): 309–317.

Wing, Coady, and Thomas D. Cook. 2013. “Strengthening The Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison.” *Journal of Policy Analysis and Management* 32 (4): 853–877.