A Practical Introduction to Regression Discontinuity Designs: Foundations

Matias D. Cattaneo* Nicolás Idrobo† Rocío Titiunik‡

November 20, 2019

Element prepared for

Cambridge Elements: Quantitative and Computational Methods for Social Science

Cambridge University Press

Published version:

https://doi.org/10.1017/9781108684606

*Department of Operations Research and Financial Engineering, Princeton University.
†Department of Political Science, University of Pennsylvania.
‡Department of Politics, Princeton University.

Abstract

In this Element and its accompanying Element, Matias D. Cattaneo, Nicolás Idrobo, and Rocío Titiunik provide an accessible and practical guide for the analysis and interpretation of Regression Discontinuity (RD) designs that encourages the use of a common set of practices and facilitates the accumulation of RD-based empirical evidence. In this Element, the authors discuss the foundations of the canonical Sharp RD design, which has the following features: (i) the score is continuously distributed and has only one dimension, (ii) there is only one cutoff, and (iii) compliance with the treatment assignment is perfect. In the accompanying Element, the authors discuss practical and conceptual extensions to the basic RD setup.

Contents

1 Introduction 3

2 The Sharp RD Design 8

3 RD Plots 20

4 The Continuity-Based Approach to RD Analysis 39

5 Validation and Falsification of the RD Design 88

6 Final Remarks 109

Bibliography 110

Acknowledgments

This Element, together with its accompanying Element (*A Practical Introduction to Regression Discontinuity Designs: Extensions*, Cattaneo, Idrobo, and Titiunik), collects and expands the instructional materials we prepared for more than 30 short courses and workshops on Regression Discontinuity (RD) methodology taught over the years 2014–2018. These teaching materials were used at various institutions and programs, including the Asian Development Bank, the Philippine Institute for Development Studies, the International Food Policy Research Institute, the ICPSR's Summer Program in Quantitative Methods of Social Research, the Abdul Latif Jameel Poverty Action Lab, the Inter-American Development Bank, the Georgetown Center for Econometric Practice, and the Universidad Católica del Uruguay's Winter School in Methodology and Data Analysis. The materials were also employed for teaching at the undergraduate and graduate level at Brigham Young University, Cornell University, Instituto Tecnológico Autónomo de México, Pennsylvania State University, Pontificia Universidad Católica de Chile, University of Michigan, and Universidad Torcuato Di Tella. We thank all these institutions and programs, as well as their many audiences, for the support, feedback, and encouragement we received over the years.

The work collected in our two Elements evolved and benefited from many insightful discussions with our present and former collaborators: Sebastián Calonico, Robert Erikson, Juan Carlos Escanciano, Max Farrell, Yingjie Feng, Brigham Frandsen, Sebastián Galiani, Michael Jansson, Luke Keele, Marko Klašnja, Xinwei Ma, Kenichi Nagasawa, Brendan Nyhan, Jasjeet Sekhon, Gonzalo Vazquez-Bare, and José Zubizarreta. Their intellectual contribution to our research program on RD designs has been invaluable, and certainly made our Elements much better than they would have been otherwise. We also thank Alberto Abadie, Joshua Angrist, Ivan Canay, Richard Crump, David Drukker, Sebastian Galiani, Guido Imbens, Patrick Kline, Justin McCrary, David McKenzie, Douglas Miller, Aniceto Orbeta, Zhuan Pei, and Andres Santos for the many stimulating discussions and criticisms we received from them over the years, which also shaped the work presented here in important ways. The co-editors Michael Alvarez and Nathaniel Beck offered useful and constructive comments on a preliminary draft of our manuscript, including the suggestion of splitting the content into two stand-alone Elements. We also thank the anonymous reviewers, who provided very valuable feedback. Last but not least, we gratefully acknowledge the support of the National Science Foundation through grant SES-1357561.

The goal of our two Elements is purposely practical and hence we focus on the empirical analysis of RD designs. We do not seek to provide a comprehensive literature review on RD

designs nor discuss theoretical aspects in detail. In this first Element, we employ the data of Meyersson (2014) as the main running example for empirical illustration. We thank this author for making his data and codes publicly available. We provide complete replication codes in both `R` and `Stata` for the entire empirical analysis discussed throughout the Element and, in addition, we provide replication codes for a second empirical illustration using the data of Cattaneo, Frandsen, and Titiunik (2015). This second empirical example is not discussed in the text to conserve space, and because it is already analyzed in our companion software articles.

`R` and `Stata` scripts replicating all the numerical results are available at `http://www.cambridge.org/introRDD`, and can be run interactively on-line via CODE-OCEAN (hyperlinks for each chapter are given below). Finally, the latest version of the general-purpose, open-source software we use, as well as other related materials, can be found at:

`https://sites.google.com/site/rdpackages/`

1 Introduction

An important goal in the social sciences is to understand the causal effect of a treatment on outcomes of interest. As social scientists, we are interested in questions as varied as the effect of minimum wage increases on unemployment, the role of information dissemination on political participation, the impact of educational reforms on student achievement, and the effects of conditional cash transfers on children's health. The analysis of such effects is relatively straightforward when the treatment of interest is randomly assigned, as this ensures the comparability of units assigned to the treatment and control conditions. However, by its very nature, many interventions of interest to social scientists cannot be randomly assigned for either ethical or practical reasons—often both.

In the absence of randomized treatment assignment, research designs that allow for the rigorous study of non-experimental interventions are particularly promising. One of these is the Regression Discontinuity (RD) design, which has emerged as one of the most credible non-experimental strategies for the analysis of causal effects. In the RD design, all units have a score, and a treatment is assigned to those units whose value of the score exceeds a known cutoff or threshold, and not assigned to those units whose value of the score is below the cutoff. The key feature of the design is that the probability of receiving the treatment changes abruptly at the known threshold. If units are unable to perfectly "sort" around this threshold, the discontinuous change in this probability can be used to learn about the local causal effect of the treatment on an outcome of interest, because units with scores barely below the cutoff can be used as a comparison group for units with scores barely above it.

The first step to employ the RD design in practice is to learn how to recognize it. There are three fundamental components in the RD design—a score, a cutoff, and a treatment. Without these three basic defining features, RD methodology cannot be employed. Therefore, an RD analysis is not always applicable to data, unlike other non-experimental methods such as those based on regression adjustments or more sophisticated selection-on-observables approaches, which can always be used to describe the conditional relationship between outcomes and treatments. The difference arises because RD is a research design, not an estimation strategy. In order to study causal effects with an RD design, the score, treatment, and cutoff must exist and be well defined, and the relationship between them must satisfy particular conditions that are objective and verifiable. The key defining feature of any canonical RD design is that the probability of treatment assignment as a function of the score changes discontinuously at the cutoff—a condition that is directly testable. In addition, the RD design comes with an extensive array of falsification tests and related empirical approaches that

can be used to offer empirical support for its validity, enhancing the credibility of particular applications. These features give the RD design an objective basis for implementation and testing that is usually lacking in other non-experimental empirical strategies, and endow it with superior credibility among non-experimental methods.

The popularity of the RD design has grown markedly over recent decades, and it is now used frequently in Economics, Political Science, Education, Epidemiology, Criminology, and many other disciplines. The RD design is also commonly used for impact and policy evaluation outside academia. This recent proliferation of RD applications has been accompanied by great disparities in how RD analysis is implemented, interpreted, and evaluated. RD applications often differ significantly in how authors estimate the effects of interest, make statistical inferences, present their results, evaluate the plausibility of the underlying assumptions, and interpret the estimated effects. The lack of consensus about best practices for validation, estimation, inference, and interpretation of RD results makes it hard for scholars and policymakers to judge the plausibility of the evidence and to compare results from different RD studies.

In both this Element and the accompanying Element, *A Practical Introduction to Regression Discontinuity Designs: Extensions* (Cattaneo, Idrobo, and Titiunik, forthcoming), our goal is to provide an accessible and practical guide for the analysis and interpretation of RD designs that encourages the use of a common set of practices and facilitates the accumulation of RD-based empirical evidence. In this Element, our focus is on the canonical RD setup that has the following features: (i) the score is continuously distributed and has only one dimension, (ii) there is only one cutoff, and (iii) compliance with treatment assignment is perfect, i.e., all units with score equal to or greater than the cutoff actually receive the treatment, and all units with score below the cutoff fail to receive the treatment and instead receive the control condition. We call this setup the Sharp RD design, and assume it throughout this Element. In the accompanying Element, we discuss extensions and departures from the basic Sharp RD design, including Fuzzy RD designs where compliance is imperfect, RD designs with multiple cutoffs, RD designs with multiple scores, geographic RD designs, and RD designs with discrete running variables.

In addition to the existence of a treatment assignment rule based on a score and a cutoff, the formal interpretation, estimation, and inference of RD treatment effects requires several other assumptions. First, we need to define the parameter of interest and provide assumptions under which this parameter is identifiable, i.e., conditions under which it is uniquely estimable in some objective sense (finite sample or super-population). Second, we must impose additional assumptions to ensure that the parameter can be estimated; these

assumptions will naturally vary according to the estimation/inference method employed and the parameter under consideration. There are two main frameworks for RD analysis, one based on continuity assumptions and another based on local randomization assumptions. Each of these defines different parameters of interest, relies on different identification assumptions, and employs different estimation and inference methods. These two alternative frameworks also generate different testable implications, which can be used to assess their validity in specific applications; see Cattaneo, Titiunik, and Vazquez-Bare (2017) for more discussion.

In this Element, we discuss the standard or *continuity-based* framework for RD analysis. This approach is based on conditions that ensure the smoothness of the regression functions, and is the framework most commonly employed in practice. We discuss the alternative *local randomization* framework for RD analysis in the second Element. This latter approach is based on conditions that ensure that the treatment can be interpreted as being randomly assigned for units near the cutoff. Both the continuity-based approach and the local randomization approach rely on the assumption that units that receive very similar score values on opposite sides of the cutoff are comparable to each other in all relevant aspects, except for their treatment status. The main distinction between these frameworks is how the idea of comparability is formalized: in the continuity-based framework, comparability is conceptualized as continuity of average (or some other feature of) potential outcomes near the cutoff, while in the local randomization framework, comparability is conceptualized as conditions that mimic a randomized experiment in a neighborhood around the cutoff.

Our upcoming discussion of the continuity-based approach focuses on the required assumptions, the adequate interpretation of the target parameters, the graphical illustration of the design, the appropriate methods to estimate treatment effects and conduct statistical inference, and the available strategies to evaluate the plausibility of the design. Our presentation of the topics is intentionally geared towards practitioners: our main goal is to clarify conceptual issues in the analysis of RD designs, and offer an accessible guide for applied researchers and policy-makers who wish to implement RD analysis. For this reason, we omit most technical discussions, but provide references for the technically inclined reader at the end of each section.

To ensure that our discussion is most useful to practitioners, we illustrate all methods by revisiting a study conducted by Meyersson (2014), who analyzed the effect of Islamic political representation in Turkey's municipal elections on the educational attainment of women. The score in this RD design is the margin of victory of the largest Islamic party in the municipality, a (nearly) continuous random variable, which makes the example suitable to illustrate both

the continuity-based methods in this Element, and also the local randomization methods in our second Element.

All the RD methods we discuss and illustrate are implemented using various general-purpose software packages, which are free and available for both `R` and `Stata`, two leading statistical software environments widely used in the social sciences. Each numerical illustration we present includes an `R` command with its output, and the analogous `Stata` command that reproduces the same analysis, though we omit the `Stata` output to avoid repetition. The local polynomial methods for continuity-based RD analysis are implemented in the package `rdrobust`, which is presented and illustrated in three companion software articles: Calonico, Cattaneo, and Titiunik (2014a), Calonico, Cattaneo, and Titiunik (2015b) and Calonico, Cattaneo, Farrell, and Titiunik (2017). This package has three functions specifically designed for continuity-based RD analysis: `rdbwselect` for data-driven bandwidth selection methods, `rdrobust` for local polynomial point estimation and inference, and `rdplot` for graphical RD analysis. In addition, the package `rddensity`, discussed by Cattaneo, Jansson, and Ma (2018), provides manipulation tests of density discontinuity based on local polynomial density estimation methods. The accompanying package `rdlocrand`, which is presented and illustrated by Cattaneo, Titiunik, and Vazquez-Bare (2016), implements the local randomization methods discussed in the second Element.

`R` and `Stata` software, replication codes, and other supplementary materials, are available at `https://sites.google.com/site/rdpackages/`. In that website, we also provide replication codes for two other empirical applications, both following closely our discussion. One employs the data on US Senate incumbency advantage originally analyzed by Cattaneo, Frandsen, and Titiunik (2015), while the other uses the Head Start data originally analyzed by Ludwig and Miller (2007) and employed in Cattaneo, Titiunik, and Vazquez-Bare (2017). Furthermore, a third distinct empirical illustration of the methods discussed in this Element, using the data of Klašnja and Titiunik (2017), is also available, and further discussed in Cattaneo, Titiunik, and Vazquez-Bare (2019).

To conclude, we emphasize that our main goal is to provide a succinct practical guide for empirical RD analysis, not to offer a comprehensive review of the literature on RD methodology—though we do offer references after each topic is presented for those interested in further reading. For early review articles see Imbens and Lemieux (2008) and Lee and Lemieux (2010), and for an edited volume with a contemporaneous overview of the RD literature see Cattaneo and Escanciano (2017). We are currently working on a literature review (Cattaneo and Titiunik, 2019) that complements these two practical Elements. See also Abadie and Cattaneo (2018) for an overview of program evaluation methods, and further

references on RD designs.

2 The Sharp RD Design

In the RD design, all units in the study receive a *score* (also known as *running variable*, *forcing variable*, or *index*), and a treatment is assigned to those units whose score is above a known cutoff and not assigned to those units whose score is below the cutoff. Our running example is based on the study by Meyersson (2014), who explored the effect of Islamic political representation in Turkey's municipal elections on the educational attainment of women. In this study, the units are municipalities and the score is the margin of victory of the (largest) Islamic party in the 1994 Turkish mayoral elections. The treatment is the Islamic party's electoral victory, and the cutoff is zero: municipalities elect an Islamic mayor when the Islamic vote margin is above zero, and elect a secular mayor otherwise.

These three components—score, cutoff, and treatment—define the RD design in general, and characterize its most important feature: in the RD design, unlike in other non-experimental studies, the assignment of the treatment follows a rule that is known (at least to the researcher) and hence empirically verifiable. To formalize, we assume that there are n units, indexed by $i = 1, 2, \ldots, n$, each unit has a score or running variable X_i, and c is a known cutoff. Units with $X_i \geq c$ are assigned to the treatment condition, and units with $X_i < c$ are assigned to the control condition. This treatment assignment, denoted T_i, is defined as $T_i = \mathbb{1}(X_i \geq c)$, where $\mathbb{1}(\cdot)$ is the indicator function, and it implies that the probability of treatment assignment as a function of the score changes discontinuously at the cutoff.

Being *assigned* to the treatment condition, however, is not the same as *receiving* or *complying with* the treatment. As in experimental and other non-experimental settings, this distinction is important in RD designs because non-compliance introduces complications and typically requires stronger assumptions to learn about treatment effects of interest. Following prior literature, we call Sharp RD design any RD design where the treatment condition assigned is identical to the treatment condition actually received for all units. Any RD design where compliance with treatment assignment is imperfect is referred to as Fuzzy RD design. In this Element, we focus exclusively on the Sharp RD design with a single score and a single cutoff. In the second Element (*A Practical Introduction to Regression Discontinuity Designs: Extensions*, Cattaneo, Idrobo, and Titiunik, forthcoming), we discuss and illustrate the Fuzzy RD design, extending the basic Sharp RD setup to settings where compliance with treatment is imperfect. (The second Element also discusses other extensions, including settings with multiple scores and multiple cutoffs.)

Regardless of whether we have perfect or imperfect compliance, a defining feature of

all RD designs is that the conditional probability of actually receiving treatment given the score changes discontinuously at the cutoff. We illustrate this for the Sharp RD design in Figure 1, where we plot the conditional probability of receiving treatment given the score, $\mathbb{P}(T_i = 1|X_i = x)$, for different values of the running variable X_i. As shown in the figure, in a Sharp RD design, this probability changes exactly from zero to one at the cutoff. Since in the Sharp RD design treatment assigned and treatment received are identical, this figure reflects both treatment assignment and treatment take-up.

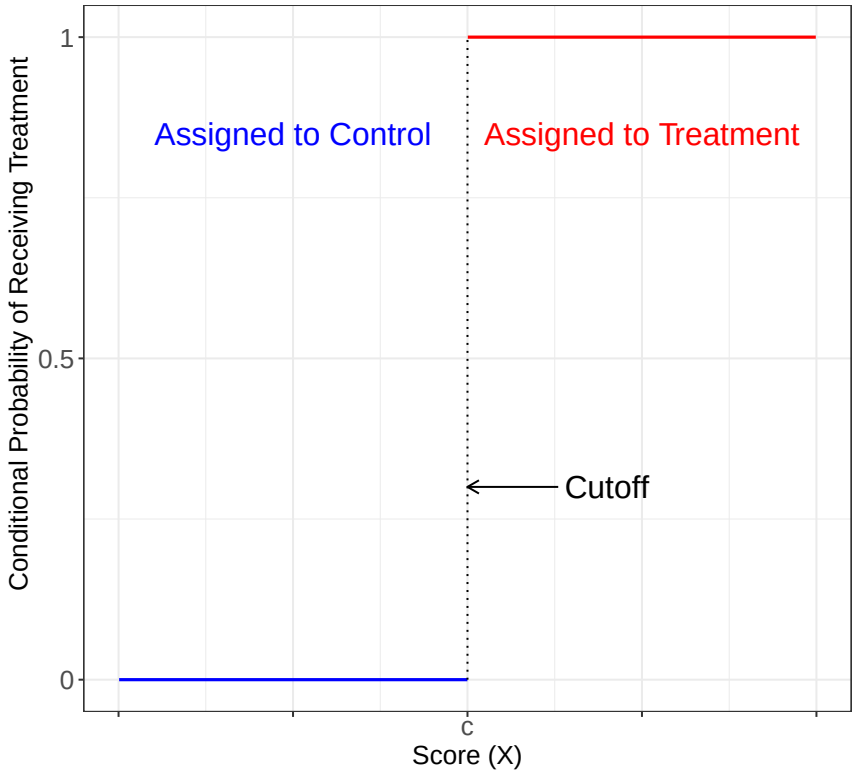


Figure 1: Conditional Probability of Receiving Treatment in the Sharp RD Design

Although it is common to use the language of experimental methods and talk about the RD treatment "assignment," in some RD applications units find themselves in different circumstances depending on their score value, and it is only ex post that the researcher interprets one of those circumstances as a treatment and the other as a control condition. This is different from an experiment, where the treatment and control conditions are always defined ex ante by the researcher, and units are explicitly assigned to one of these conditions. For example, in the RD design studied by Meyersson (2014), a municipality is treated when it elects a mayor from an Islamic party, and control when it elects a mayor from a secular party. In this case, there is no explicit assignment of municipalities to different ex ante ex-

perimental conditions; rather, depending on the outcome of the election, municipalities find themselves in different situations (with or without an Islamic mayor), which can be understood as treatment versus control for some analytic purposes. These conceptual distinctions between experimental and RD treatment assignments do not affect the validity of the RD mathematical expressions. But the reader should keep in mind these caveats when we employ the term "treatment assignment" in the RD context.

Following the causal inference literature, we adopt the potential outcomes framework and assume that each unit has two potential outcomes, $Y_i(1)$ and $Y_i(0)$, corresponding, respectively, to the outcomes that would be observed under the treatment or control conditions. In this framework, treatment effects are defined in terms of comparisons between features of (the distribution of) both potential outcomes, such as their means, variances or quantiles. Although every unit is assumed to have both $Y_i(1)$ and $Y_i(0)$, these outcomes are called potential because only one of them is observed. If unit i receives the treatment, we will observe $Y_i(1)$, the unit's outcome under treatment, and $Y_i(0)$ will remain latent or unobserved. Similarly, if i receives the control condition, we will observe $Y_i(0)$ but not $Y_i(1)$. This results in the so-called fundamental problem of causal inference, and implies that the treatment effect at the individual level is fundamentally unknowable.

The observed outcome is

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

Throughout this Element, we adopt the usual econometric perspective that sees the data $(Y_i, X_i)_{i=1}^n$ as a random sample from a larger population, taking the potential outcomes $(Y_i(1), Y_i(0))_{i=1}^n$ as random variables. We consider an alternative perspective in the second Element when we discuss inference in the context of the local randomization RD framework.

In the specific context of the Sharp RD design, the fundamental problem of causal inference occurs because we only observe the outcome under control, $Y_i(0)$, for those units whose score is below the cutoff, and we only observe the outcome under treatment, $Y_i(1)$, for those units whose score is above the cutoff. We illustrate this 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 score. In statistics, conditional expectation functions such as these are usually 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 $X_i \geq c$, 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), $X_i < c$, but unobserved for $X_i \geq c$ (dashed blue line). Thus, the observed average outcome given the score is

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

The Sharp RD design exhibits an extreme case of lack of common support, as units in the control ($T_i = \mathbb{1}(X_i \geq c) = 0$) and treatment ($T_i = \mathbb{1}(X_i \geq c) = 1$) groups cannot have the same value of the running variable (X_i). This feature sets aside RD designs from other non-experimental settings, and highlights that RD analysis fundamentally relies on extrapolation towards the cutoff point. As we discuss throughout this Element, a central goal of empirical RD analysis is to adequately perform (local) extrapolation in order to compare control and treatment units. This unique feature of the RD design also makes causal interpretation of some parameters potentially more difficult; see Cattaneo, Titiunik, and Vazquez-Bare (2017) for further discussion on this point.

As shown in Figure 2, the average treatment effect at a given value of the score, $\mathbb{E}[Y_i(1)|X_i = x] - \mathbb{E}[Y_i(0)|X_i = x]$, is the vertical distance between the two regression curves at that value. This distance cannot be directly estimated because we never observe both curves for the same value of x. However, a special situation occurs at the cutoff c: this is the only point at which we "almost" observe both curves. To see this, we imagine having units with score exactly equal to c, and units with score barely below c (that is, with score $c - \varepsilon$ 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 c are not abruptly different from their values at points near c, the units with $X_i = c$ and $X_i = c - \varepsilon$ would be very similar except for their treatment status, and we could approximately calculate the vertical distance at c using observed outcomes. In the figure, the vertical distance at c is $\mathbb{E}[Y_i(1)|X_i = c] - \mathbb{E}[Y_i(0)|X_i = c] \equiv \mu_+ - \mu_-$; this is precisely the treatment effect that can be estimated with a Sharp RD design. The *Sharp RD treatment effect* is thus formally defined as

$$\tau_{\texttt{SRD}} \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c].$$

This parameter captures the (reduced form) treatment effect for units with score values $X_i = c$. It answers the following question: what would be the average outcome change for units with score level $X_i = c$ if we switched their status from control to treated? As we

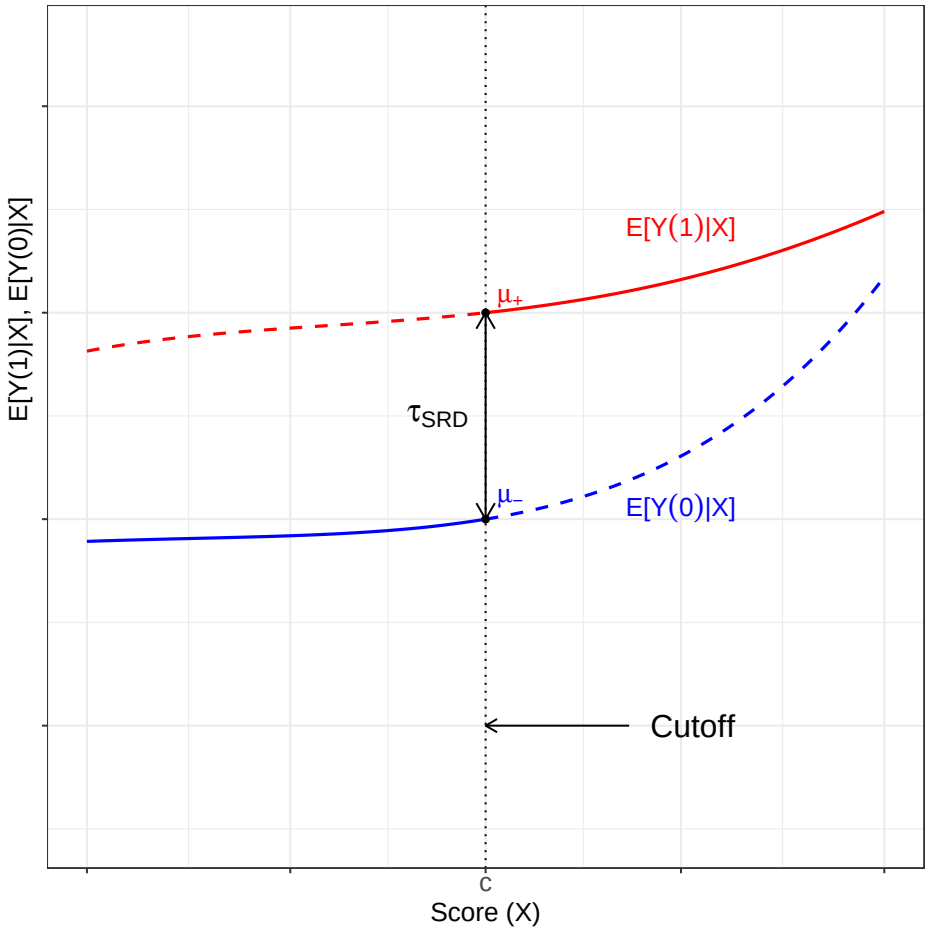


Figure 2: RD Treatment Effect in Sharp RD Design

discuss below, this treatment effect is, by construction, local in nature and, in the absence of additional assumptions, not informative about treatment effects at other levels of the score. Moreover, since the definition of a Sharp RD design implies that all units with $X_i = c$ are treated, $\tau_{\texttt{SRD}}$ can be interpreted as a (local, RD) average treatment effect on the treated.

The assumption 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 idea was first formalized by Hahn, Todd, and van der Klaauw (2001) using continuity assumptions. These authors showed that, among other conditions, if the regression functions $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$, seen as functions of x, are continuous at $x = c$, then in a Sharp RD design we have

$$\mathbb{E}[Y_i(1) - Y_i(0)|X_i = c] = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x]. \tag{2.1}$$

The result in Equation (2.1) says that, if the average potential outcomes are continuous

functions of the score at c, the difference between the limits of the treated and control average *observed* outcomes as the score converges to the cutoff is equal to the average treatment effect at the cutoff. Informally, a function $g(x)$ is continuous at the point $x = a$ if the values of $g(x)$ and $g(a)$ get close to each other as x gets close to a. In the RD context, continuity means that as the score x gets closer and closer to the cutoff c, the average potential outcome function $\mathbb{E}[Y_i(0)|X_i = x]$ gets closer and closer to its value at the cutoff, $\mathbb{E}[Y_i(0)|X_i = c]$ (and analogously for $\mathbb{E}[Y_i(1)|X_i = x]$). Thus, continuity gives a formal justification for estimating the Sharp RD effect by focusing on observations above and below the cutoff in a very small neighborhood around it. By virtue of being very close to the cutoff, the observations in this neighborhood will have very similar score values; and by virtue of continuity, their average potential outcomes will also be similar. Therefore, continuity offers one justification for using observations just below the cutoff to approximate the average outcome that units just above the cutoff would have had if they had received the control condition instead of the treatment.

2.1 The Effect of Islamic Political Representation on Women's Education

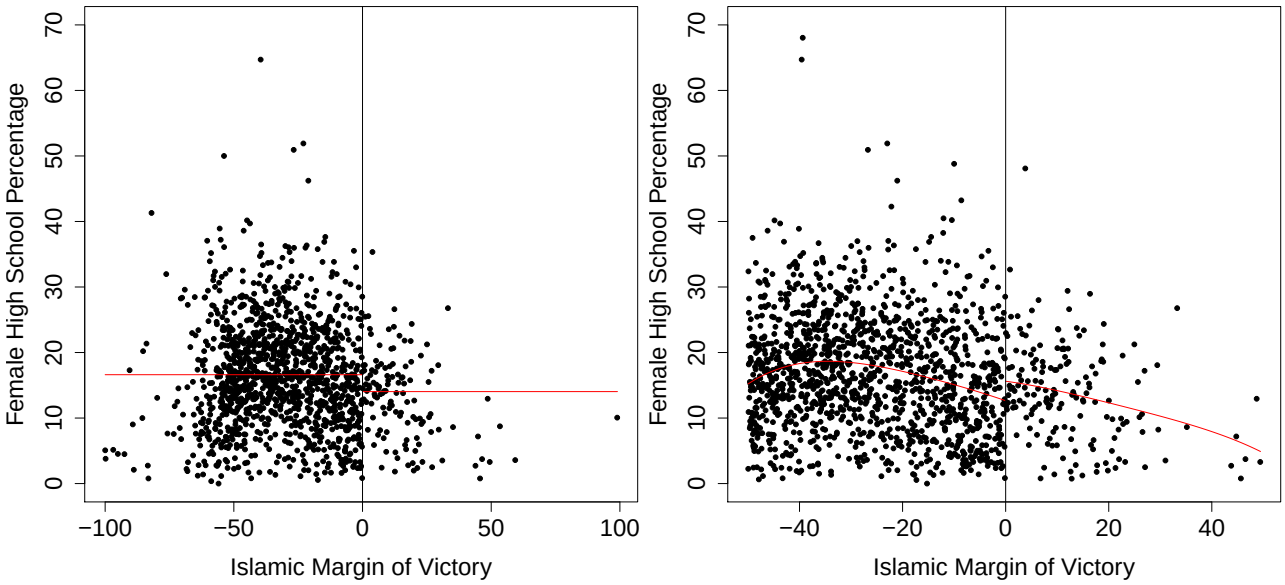
We now introduce in more detail the empirical example that we employ throughout this Element, originally analyzed by Meyersson (2014), henceforth Meyersson. This example employs a Sharp RD design, based on close elections in Turkey, to study the impact of having a mayor from an Islamic party on various outcomes. The running variable is based on vote shares, as popularized by the work of Lee (2008).

Meyersson is broadly interested in the effect of Islamic parties' control of local governments on women's rights, in particular on the educational attainment of young women. The methodological challenge is that municipalities where the support for Islamic parties is high enough to result in the election of an Islamic mayor may differ systematically from municipalities where the support for Islamic parties is more tenuous and results in the election of a secular mayor. (For brevity, we refer to a mayor who belongs to one of the Islamic parties as an "Islamic mayor," and to a mayor who belongs to a non-Islamic party as a "secular mayor.") If some of the characteristics on which both types of municipalities differ affect (or are correlated with) the educational outcomes of women, a simple comparison of municipalities with an Islamic versus a secular mayor will be misleading. For example, municipalities where an Islamic mayor wins in 1994 may be on average more religiously conservative than municipalities where a secular mayor is elected. If religious conservatism affects the educational outcomes of women, the naïve comparison between municipalities controlled by an

Islamic versus a secular mayor will not successfully isolate the effect of the Islamic party's control of the local government. Instead, the effect of interest will be contaminated by differences in the degree of religious conservatism between the two groups.

This challenge is illustrated in Figure 3, where we plot the percentage of young women who had completed high school by 2000 against the Islamic margin of victory in the 1994 mayoral elections (more information on these variables is given below). These figures are examples of so-called RD plots, which we discuss in detail in Section 3. In Figure 3(a), we show the scatter plot of the raw data (where each point is an observation), superimposing the overall sample mean for each group; treated observations (municipalities where an Islamic mayor is elected) are located to the right of zero, and control observations (municipalities where a secular mayor is elected) are located to the left of zero. The raw comparison reveals a negative average difference: municipalities with an Islamic mayor have, on average, lower educational attainment of women. Figure 3(b), shows the scatter plot for the subset of municipalities where the Islamic margin of victory is within 50 percentage points, a range that includes 83% of the total observations; this second figure superimposes a fourth-order polynomial fit separately on either side of the cutoff. Figure 3(b) reveals that the negative average effect in Figure 3(a) arises because there is an overall negative relationship or slope between Islamic vote percentage and educational attainment of women for the majority of the observations, so that the higher the Islamic margin of victory, the lower the educational attainment of women. Thus, a naïve comparison of treated and control municipalities, which differ systematically in the Islamic vote percentage, will mask systematic differences and may lead to incorrect inferences about the effect of electing an Islamic mayor.

The RD design can be used in cases such as these to isolate a treatment effect of interest from all other systematic differences between treated and control groups. Under appropriate assumptions, a comparison of municipalities where the Islamic party barely wins the election versus municipalities where the Islamic party barely loses will reveal the causal (local) effect of Islamic party control of the local government on the educational attainment of women. If parties cannot systematically manipulate the vote percentage they obtain, observations just above and just below the cutoff will tend to be comparable in terms of all characteristics with the exception of the party that won the 1994 election. Thus, right at the cutoff, the comparison is free of the complications introduced by systematic observed and unobserved differences between the groups. This strategy is illustrated in Figure 3(b), where we see that, despite the negative slope on either side, right near the cutoff the effect of an Islamic victory on the educational attainment of women is positive, in stark contrast to the negative difference-in-means in Figure 3(a).



(a) Raw Comparison of Means (b) Local Comparison of Means

Figure 3: Municipalities with Islamic Mayor versus Municipalities with Secular Mayor (Meyersson data)

Meyersson's original study employs an RD design to circumvent these methodological challenges and to estimate a causal effect of local Islamic rule. The design is focused exclusively on the 1994 Turkish mayoral elections. The unit of analysis is the municipality, and the score is the Islamic margin of victory, defined as the difference between the vote percentage obtained by the largest Islamic party, and the vote percentage obtained by the largest secular party opponent. Two Islamic parties competed in the 1994 mayoral elections, *Refah* and *Büyük Birlik Partisi* (BBP). However, the results essentially capture the effect of a victory by *Refah*, as the BBP received only 0.94% of the national vote and won in only 11 of the 329 municipalities where an Islamic mayor was elected. As defined, the Islamic margin of victory can be positive or negative, and the cutoff that determines an Islamic party victory is located at zero. Given this setup, the treatment group consists of municipalities that elected a mayor from an Islamic party in 1994, and the control group consists of municipalities that elected a mayor from a secular party. The outcome we re-analyze is the educational attainment of women who were (potentially) in high school during the period 1994-2000, calculated as the percentage of the cohort of women aged 15 to 20 in 2000 who had completed high school by 2000 according to the 2000 Turkish census. For brevity, we refer to this outcome as the educational attainment of women.

In order to streamline the computer code for our analysis, we rename the variables in the following way.

- `Y`: educational attainment of women, measured as the percentage of women aged 15 to 20 in 2000 who had completed high school by 2000.
- `X`: vote margin obtained by the Islamic party in the 1994 Turkish mayoral elections, measured as the vote percentage obtained by the Islamic party minus the vote percentage obtained by its strongest secular party opponent.
- `T`: electoral victory of the Islamic party in 1994, equal to 1 if the Islamic party won the mayoral election and 0 otherwise.

The Meyersson dataset also contains several predetermined covariates that we use in subsequent sections to investigate the plausibility of the RD design, and also to illustrate covariate-adjusted estimation methods. The covariates that we include in our analysis are the Islamic vote percentage in 1994 (`vshr_islam1994`), the number of parties receiving votes in 1994 (`partycount`), the logarithm of the population in 1994 (`lpop1994`), an indicator equal to one if the municipality elected an Islamic party in the previous election in 1989 (`i89`), a district center indicator (`merkezi`), a province center indicator (`merkezp`), a sub-metro center indicator (`subbuyuk`), and a metro center indicator (`buyuk`).

Table 1: Descriptive Statistics for Meyersson

Variable	Mean	Median	Std. Dev.	Min.	Max.
Y	16.306	15.523	9.584	0.000	68.038
X	-28.141	-31.426	22.115	-100.000	99.051
T	0.120	0.000	0.325	0.000	1.000
Percentage of men aged 15-20 with high school education	19.238	18.724	7.737	0.000	68.307
Islamic percentage of votes in 1994	13.872	7.029	15.385	0.000	99.526
Number of parties receiving votes 1994	5.541	5.000	2.192	1.000	14.000
Log population in 1994	7.840	7.479	1.188	5.493	15.338
Percentage of population below 19 in 2000	40.511	39.721	8.297	6.544	68.764
Percentage of population above 60 in 2000	9.222	8.461	3.960	1.665	27.225
Gender ratio in 2000	107.325	103.209	25.293	74.987	1033.636
Household size in 2000	5.835	5.274	2.360	2.823	33.634
District center	0.345	0.000	0.475	0.000	1.000
Province center	0.023	0.000	0.149	0.000	1.000
Sub-metro center	0.022	0.000	0.146	0.000	1.000

Note: the number of observations for all variables is 2,629

Table 1 presents descriptive statistics for the three RD variables (`Y`, `X`, and `T`), and the municipality-level predetermined covariates. The outcome of interest (`Y`) has a minimum of 0 and a maximum of 68.04, with a mean of 16.31. This implies that there is at least one municipality in 2000 where no women in the 15-to-20 age cohort had completed high school, and on average 16.31% of women in this cohort had completed high school by the year 2000. The Islamic vote margin (`X`) ranges from -100 (party receives zero votes) to 100

(party receives 100% of the vote), and it has a mean of −28.14, implying that on average the Islamic party loses by 28.14 percentage points. This explains why the mean of the treatment variable (T) is 0.120, since this indicates that in 1994 an Islamic mayor was elected in only 12.0% of the municipalities. This small proportion of victories is consistent with the finding that the average margin of victory is negative and thus leads to electoral loss.

2.2 The Local Nature of RD Effects

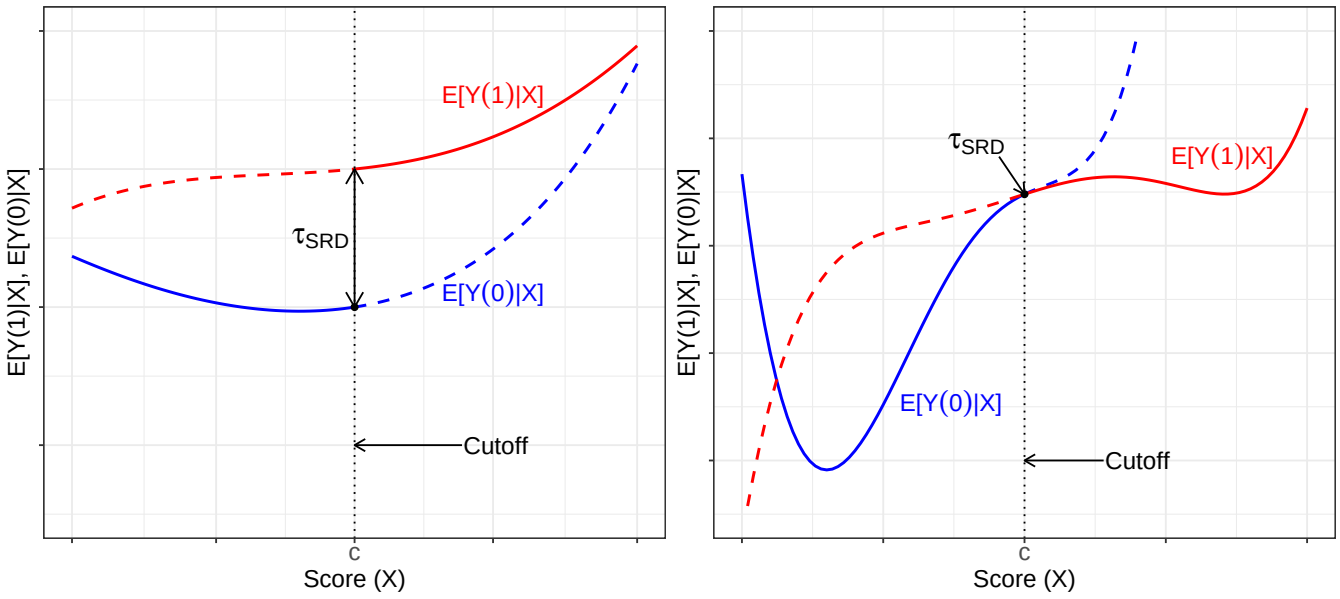
The Sharp RD parameter presented above can be interpreted as causal in the sense that it captures the average difference in potential outcomes under treatment versus control. However, in contrast to other causal parameters in the potential outcomes framework, this average difference is calculated at a single point on the support of a continuous random variable (the score X_i), and as a result captures a causal effect that is local in nature. According to some perspectives, this parameter cannot even be interpreted as causal because it cannot be reproduced via manipulation or experimentation.

Regardless of its status as a causal parameter, the RD treatment effect tends to have limited external validity, that is, the RD effect is generally not representative of the treatment effects that would occur for units with scores away from the cutoff. As discussed above, in the canonical Sharp RD design, the RD effect can be interpreted graphically as the vertical difference between $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ at the point where the score equals the cutoff, $x = c$. In the general case where the average treatment effect varies as a function of the score X_i, as is very common in applications, this effect may not be informative of the average effect of treatment at values of x different from c. For this reason, in the absence of specific (usually restrictive) assumptions about the global shape of the regression functions, the effect recovered by the RD design is only the average effect of treatment for units *local to the cutoff*, i.e., for units with score values $X_i = c$.

In the context of the Meyersson application, the lack of external validity is reflected in the focus on close, as opposed to all, elections. As illustrated in Figure 3(a), it seems that the educational attainment of women is higher in municipalities where the Islamic party barely wins than in municipalities where the party barely loses the election. By definition, the sample of municipalities near the cutoff comprises constituencies where the Islamic party is very competitive. It is likely that the political preferences and religious affiliation of Turkish citizens in these municipalities differ systematically from those in municipalities where the Islamic party wins or loses by very large margins. This means that, although the RD results indicate that Islamic mayors lead to an increase in the educational attainment of women

in competitive municipalities, it is not possible to know whether the same positive effect in female education would be seen if a mayor from the Islamic party governed a municipality with strong preferences for secular political parties. Figure 3 reveals that the vast majority of observations in the sample of 1994 Turkish mayoral elections is composed of municipalities where the Islamic party lost by a very large margin; without further assumptions, the RD effect is not informative about the educational effect of an Islamic party victory in these municipalities.

In general, the degree of representativeness or external validity of the RD treatment effect will depend on the specific application under consideration. For example, in the hypothetical scenario illustrated in Figure 4(a), the vertical distance between $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ at $x = c$ is considerably higher than at other points, but the effect is positive everywhere. A much more heterogeneous hypothetical 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 the counterfactual (dotted) regression functions are never observed in real examples, it is not possible to know with certainty the degree of external validity of any given RD application.



(a) Mild Heterogeneity (b) Severe Heterogeneity

Figure 4: Local Nature of the RD Effect

Increasing the external validity of RD estimates and estimands is a topic of active research and, regardless of the approach taken, necessarily requires more assumptions. For example, extrapolation of RD treatment effects can be done by imposing additional assumptions about

(i) the regression functions near the cutoff (Dong and Lewbel, 2015; Wing and Cook, 2013), (ii) local independence assumptions (Angrist and Rokkanen, 2015) (iii) exploiting specific features of the design such as imperfect compliance (Bertanha and Imbens, 2019), or (iv) the presence of multiple cutoffs (Cattaneo, Keele, Titiunik, and Vazquez-Bare, 2016, 2019). On this regard, RD designs are not different from randomized experiments: they both require additional assumptions to map internally valid estimates into externally valid ones.

2.3 Further Reading

For an introduction to causal inference based on potential outcomes see Imbens and Rubin (2015) and references therein. For a review on causal inference and program evaluation methods see Abadie and Cattaneo (2018) and references therein. The RD design was originally proposed by Thistlethwaite and Campbell (1960), and historical as well as early review articles are given by Cook (2008), Imbens and Lemieux (2008), and Lee and Lemieux (2010). Lee (2008) provided an influential contribution to the identification of RD effects; Lee (2008) and Pettersson-Lidbom (2008) were the first to apply the RD design to close elections. The edited volume by Cattaneo and Escanciano (2017) provides a recent overview of the RD literature and includes several recent methodological and practical contributions.

3 RD Plots

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, inference, and falsification discussed below, adds transparency to the analysis by displaying the observations used for estimation and inference. RD plots also allow researchers to readily summarize the main empirical findings as well as other important features of the work conducted. We now discuss the most transparent and effective methods to graphically illustrate the RD design.

At first glance, it seems that one should be able to illustrate the relationship between the outcome and the running variable by simply constructing a scatter plot of the observed outcome against the score, clearly identifying the points above and below the cutoff. However, this strategy is rarely useful, as it is often hard to see "jumps" or discontinuities in the outcome-score relationship by simply looking at the raw data. We illustrate this point with the Meyersson application, plotting the educational attainment of women against the Islamic vote margin using the raw observations. We create this scatter plot in `R` with the `plot` command.

R Snippet 1

```
> plot(X, Y, xlab = "Score", ylab = "Outcome", col = 1, pch = 20,
+ cex.axis = 1.5, cex.lab = 1.5)
> abline(v = 0)
```

Stata Snippet 1

```
. twoway (scatter Y X, ///
> mcolor(black) xline(0, lcolor(black))), ///
> graphregion(color(white)) ytitle(Outcome) ///
> xtitle(Score)
```

Each point in Figure 5 corresponds to one raw municipality-level observation in the dataset, so there are 2,629 points in the scatter plot (see Table 1). Although this plot is helpful to visualize the raw observations, detect outliers, etc., its effectiveness for visualizing the RD design is limited. In this application, there is empirical evidence that the Islamic party's victory translates into a small increase in women's educational attainment. Despite this evidence of a positive RD effect, a jump in the values of the outcome at the cutoff cannot be seen by simply looking at the raw cloud of points around the cutoff in Figure 5. In general, raw scatter plots do not allow for easy visualization of the RD effect even when the effect is large.

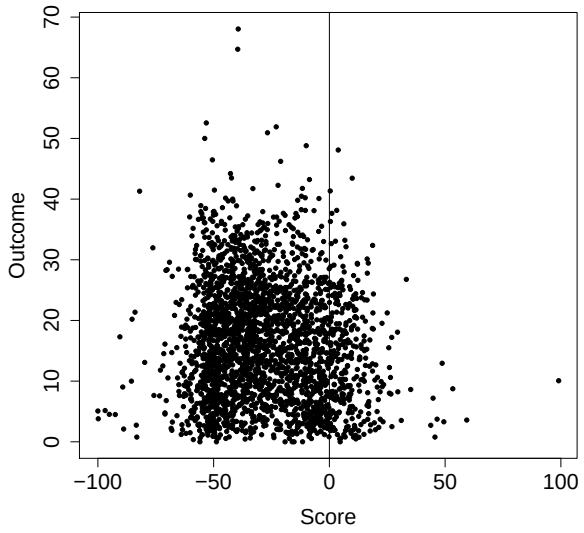


Figure 5: Scatter Plot (Meyersson Data)

A more useful approach is to aggregate or "smooth" the data before plotting. The typical RD plot presents two summaries: (i) a global polynomial fit, represented by a solid line, and (ii) local sample means, represented by dots. The global polynomial fit is simply a smooth approximation to the unknown regression functions based on a fourth- or fifth-order polynomial regression fit of the outcome on the score, fitted separately above and below the cutoff, and using the original raw data. In contrast, the local sample means are created by first choosing disjoint (i.e., non-overlapping) intervals or "bins" of the score, calculating the mean of the outcome for the observations falling within each bin, and then plotting the average outcome in each bin against the mid point of the bin. Local sample means can be interpreted as a non-smooth approximation to the unknown regression functions. The combination of these two ingredients in the same plot allows researchers to visualize the global or overall shape of the regression functions for treated and control observations, while at the same time retaining enough information about the local behavior of the data to observe the RD treatment effect and the variability of the data around the global fit. Note that, in the standard RD plot, the global polynomial is calculated using the original observations, not the binned observations.

For example, in the Meyersson application, if we use 20 bins of equal length on each side of the cutoff, we partition the support of the Islamic margin of victory into 40 disjoint intervals of length 5. Recall that a party's margin of victory ranges from -100 to 100, and that the Islamic margin of victory in the Meyersson data ranges from -100 to 99.051. Table 2 shows the bins and the corresponding average outcomes in this case, where we denote

the bins by $\mathcal{B}_{-,1}, \mathcal{B}_{-,2}, \ldots, \mathcal{B}_{-,20}$ (control group) and $\mathcal{B}_{+,1}, \mathcal{B}_{+,2}, \ldots, \mathcal{B}_{+,20}$ (treatment group), using the subscripts $-$ and $+$ to indicate, respectively, bins located to the left and right of the cutoff. In this table, each local sample average is computed as

$$\bar{Y}_{-,j} = \frac{1}{\#\{X_i \in \mathcal{B}_{-,j}\}} \sum_{i:X_i \in \mathcal{B}_{-,j}} Y_i \qquad \text{and} \qquad \bar{Y}_{+,j} = \frac{1}{\#\{X_i \in \mathcal{B}_{+,j}\}} \sum_{i:X_i \in \mathcal{B}_{+,j}} Y_i,$$

where $j = 1, 2, \cdots, 20$.

Table 2: Partition of Islamic Margin of Victory into 40 Bins of Equal Length (Meyersson Data)

Bin	Average Outcome in Bin	Number of Observations	Group Assignment
$\mathcal{B}_{-,1} = [-100, -95)$	$\bar{Y}_{-,1} = 4.6366$	4	Control
$\mathcal{B}_{-,2} = [-95, -90)$	$\bar{Y}_{-,2} = 10.8942$	2	Control
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{B}_{-,19} = [-10, -5)$	$\bar{Y}_{-,19} = 12.9518$	149	Control
$\mathcal{B}_{-,20} = [-5, 0)$	$\bar{Y}_{-,20} = 13.8267$	148	Control
$\mathcal{B}_{+,1} = [0, 5)$	$\bar{Y}_{+,1} = 15.3678$	109	Treatment
$\mathcal{B}_{+,2} = [5, 10)$	$\bar{Y}_{+,2} = 13.9640$	83	Treatment
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\mathcal{B}_{+,19} = [90, 95)$	$\bar{Y}_{+,19} = \text{NA}$	0	Treatment
$\mathcal{B}_{+,20} = [95, 100]$	$\bar{Y}_{+,20} = 10.0629$	1	Treatment

In Figure 6, we plot the binned outcome means shown in Table 2 against the score, adding a fourth-order global polynomial fit estimated separately for treated and control observations. (Below we show how to create this plot using the `rdplot` command.) The global fit reveals that the observed regression function seems to be non-linear, particularly on the control (left) side. At the same time, the binned means let us see the local behavior of the average response variable around the global fit. The plot also reveals a positive jump at the cutoff: the average educational attainment of women seems to be higher in those municipalities where the Islamic party obtained a barely positive margin of victory than in those municipalities where the Islamic party barely lost.

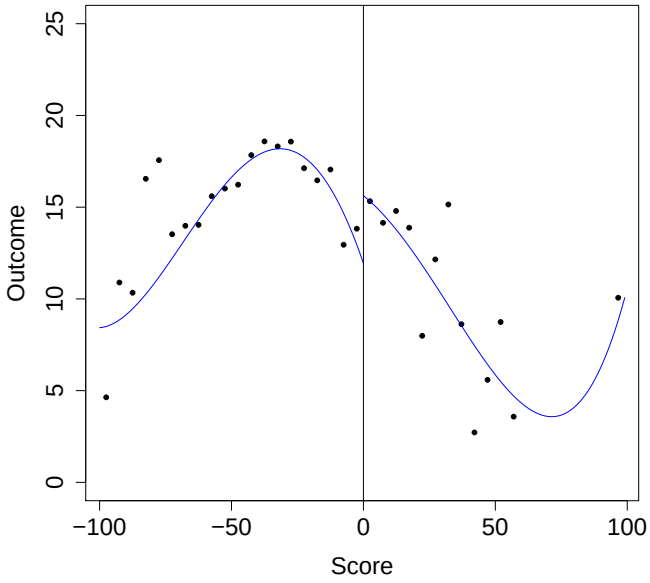


Figure 6: RD Plot for Meyersson Data Using 40 Bins of Equal Length

The types of information conveyed by Figures 5 and 6 are very different. In the raw scatter plot in Figure 5, it is difficult to see any systematic pattern, and there is no visible discontinuity in the average outcome at the cutoff. In contrast, when we bin the data and include a global polynomial fit in Figure 6, the plot now allows us to see a discontinuity at the cutoff and to better understand the shape of the underlying regression function over the whole support of the running variable. Binning the data may reveal striking patterns that can remain hidden in a simple scatter plot. Since binning often leads to drastically different patterns from those seen in the raw data, we now discuss how to choose the type and number of bins in a data-driven, transparent, and (sometimes) optimal way.

3.1 Choosing the Location of Bins

There are two different types of bins that can be used in the construction of RD plots: bins that have equal length, as in Table 2, or bins that contain (roughly) the same number of observations but whose length may differ. We refer to these two types as evenly-spaced and quantile-spaced bins, respectively.

In order to define the bins more precisely, we assume that the running variable takes values inside the interval $[x_l, x_u]$. In the Meyersson application, $x_l = -100$ and $x_u = 100$. We continue to use the subscripts $+$ and $-$ to denote treated and control observations, respec-

tively. The bins are constructed separately for treated and control observations, partitioning the support in non-overlapping intervals. We use J_- and J_+ to denote the total number of bins chosen to the left and right of the cutoff, respectively.

We define the bins generally as follows:

$$\texttt{Control Bins:} \quad \mathcal{B}_{-,j} = \begin{cases} [x_l \ , \ b_{-,1}) & j = 1 \\ [b_{-,j-1} \ , \ b_{-,j}) & j = 2, \cdots, J_- - 1 \\ [b_{-,J_--1} \ , \ c) & j = J_- \end{cases}$$

$$\texttt{Treated Bins:} \quad \mathcal{B}_{+,j} = \begin{cases} [c \ , \ b_{+,1}) & j = 1 \\ [b_{+,j-1} \ , \ b_{+,j}) & j = 2, \cdots, J_+ - 1 \\ [b_{+,J_+-1} \ , \ x_u] & j = J_+, \end{cases}$$

with $b_{-,0} < b_{-,1} < \cdots < b_{-,J_-}$ and $b_{+,0} < b_{+,1} < \cdots < b_{+,J_+}$. In other words, the union of the control and treated bins, $\mathcal{B}_{-,1} \cup \mathcal{B}_{-,2} \cup \ldots \cup \mathcal{B}_{J_-} \cup \mathcal{B}_{+,1} \cup \mathcal{B}_{+,2} \cup \ldots \cup \mathcal{B}_{J_+}$, forms a disjoint partition of the support of the running variable, $[x_l, x_u]$, centered at the cutoff c.

Letting $X_{-,(i)}$ and $X_{+,(i)}$ denote the ith quantiles of the control and treatment subsamples, respectively, and $\lfloor \cdot \rfloor$ denote the floor function, we can now formally define evenly-spaced (ES) and quantile-spaced (QS) bins.

- **Evenly-spaced (ES) bins**: non-overlapping intervals that partition the entire support of the running variable, all of the same length within each treatment assignment status:

$$b_{-,j} = x_l + \frac{j \ (c - x_l)}{J_-} \qquad \text{and} \qquad b_{+,j} = c + \frac{j \ (x_u - c)}{J_+}.$$

  Note that all ES bins in the control side have length $\frac{c - x_l}{J_-}$ and all bins in the treated side have length $\frac{x_u - c}{J_+}$.

- **Quantile-spaced (QS) bins**: non-overlapping intervals that partition the entire support of the running variable, all containing (roughly) the same number of observations within each treatment assignment status:

$$b_{-,j} = X_{-,(\lfloor j/J_- \rfloor)} \qquad \text{and} \qquad b_{+,j} = X_{+,(\lfloor j/J_+ \rfloor)}.$$

  Note that the length of QS bins may differ even within treatment assignment status; the bins will be larger in regions of the support where there are fewer observations.

In practical terms, the most important difference between ES and QS bins is the underlying variability of the local mean estimate in every bin. Although ES bins have equal length, if the observations are not uniformly distributed on $[x_l, x_u]$, each bin may contain a different number of observations. In an RD plot with ES bins, each of the local means represented by a dot may be computed using a different number of observations and thus may be more or less precisely calculated than the other local means in the plot, affecting comparability. For example, Table 2 shows that there are only 4 observations in $[-100, -95]$, and only 2 observations in $[-95, -90]$; thus, the variance of these local mean estimates is very high because they are constructed with very few observations.

In contrast, QS bins contain approximately the same number of observations by construction. Moreover, a quantile-spaced RD plot has the advantage of providing a quick visual representation of the density of observations over the support of the running variable. For example, if there are very few observations far from the cutoff, an RD plot with quantile-spaced bins will tend to be "empty" near the extremes of $[x_l, x_u]$, and will quickly convey the message that there are no observations with values of the score near x_l or x_u.

We now use the `rdplot` command to produce different RD plots and illustrate the differences between binning strategies, using the `binselect` option to choose between ES and QS methods. First, we reproduce the RD plot above, using 20 evenly-spaced bins on each side via the option `nbins`.

R Snippet 2

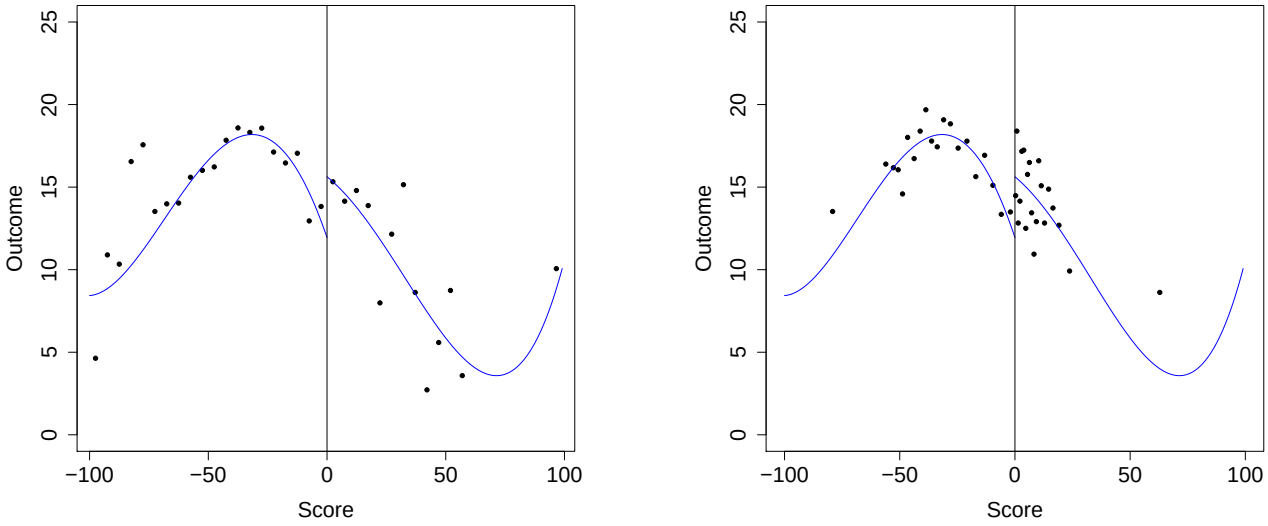
```
> out = rdplot(Y, X, nbins = c(20, 20), binselect = "es", y.lim = c(0,
+ 25), cex.axis = 1.5, cex.lab = 1.5)
> summary(out)
Call: rdplot

Number of Obs.                 2629
Kernel                      Uniform

Number of Obs.                2314          315
Eff. Number of Obs.           2314          315
Order poly. fit (p)              4            4
BW poly. fit (h)           100.000       99.051
Number of bins scale             1            1

Bins Selected                   20           20
Average Bin Length           5.000        4.953
Median Bin Length            5.000        4.953

IMSE-optimal bins               11            7
Mimicking Variance bins         40           75

Relative to IMSE-optimal:
Implied scale                1.818        2.857
WIMSE variance weight        0.143        0.041
WIMSE bias weight            0.857        0.959
```

Stata Snippet 2

```
. rdplot Y X, nbins(20 20) binselect(es) ///
> graph_options(graphregion(color(white)) ///
> xtitle(Score) ytitle(Outcome))
```

The full output of `rdplot` includes several descriptive statistics in addition to the actual plot, which is shown in Figure 7(a). The total number of observations is shown in the very top row, where we can also see the type of weights used to plot the observations. We have 2,629 observations in total, which by default are all given equal or uniform weight, as is indicated by the output `Kernel = Uniform`. The rest of the output is divided in two columns, corresponding to observations located to the left or right of the cutoff, respectively.

The output shows that there are 2,314 observations to the left of the cutoff, and 315 to the right, consistent with our descriptive analysis indicating that the Islamic party lost the



(a) 40 Evenly-Spaced Bins

(b) 40 Quantile-Spaced Bins

Figure 7: RD Plots—Meyersson Data

majority of these electoral races. The third row in the top panel indicates that the global polynomial fit used in the RD plot is of order 4 on both sides of the cutoff. The fourth row indicates the window or bandwidth h where the global polynomial fit was conducted; the global fit uses all observations in $[c-h, c)$ on the control side, and all observations in $[c, c+h]$ on the treated side. By default, all control and treated observations are included in the control and treated fit, respectively. Since the range of the Islamic margin of victory is $[-100, 99.051]$, the bandwidth on the right is slightly smaller than 100. Finally, the last row in the top panel shows the scale selected, which is an optional factor by which the chosen number of bins can be multiplied to either increase or decrease the original choice; by default, this factor is one and no scaling is performed.

The lower part of the output shows results on the number and type of bins selected. The top two rows show that we have selected 20 bins to the left of the cutoff, and 20 bins to the right of the cutoff. On the control side, the length of each bin is exactly $5 = \frac{c-x_l}{J_-} = \frac{0-(-100)}{20} = 100/20$. However, the actual length of the ES bins to the right of the cutoff is slightly smaller than 5, as the edge of the support on the treated side is 99.051 instead of 100. The actual length of the bins to the right of the cutoff is $\frac{x_u-c}{J_+} = \frac{99.051-0}{20} = 99.051/20 = 4.9526$. We postpone discussion of the five bottom rows until the next subsection where we discuss optimal bin number selection.

We now compare this plot to an RD plot that also uses 20 bins on each side, but with

quantile-spaced bins instead of evenly-spaced bins selected by setting the option `binselect = "qs"`. The resulting plot is shown in Figure 7(b).

R Snippet 3

```
> out = rdplot(Y, X, nbins = c(20, 20), binselect = "qs", x.lim = c(-100,
+ 100), y.lim = c(0, 25), cex.axis = 1.5, cex.lab = 1.5)
> summary(out)
Call: rdplot

Number of Obs.                  2629
Kernel                       Uniform

Number of Obs.                  2314            315
Eff. Number of Obs.             2314            315
Order poly. fit (p)                4              4
BW poly. fit (h)             100.000         99.051
Number of bins scale               1              1

Bins Selected                     20             20
Average Bin Length             4.995          4.957
Median Bin Length              2.950          1.011

IMSE-optimal bins                 21             14
Mimicking Variance bins           44             41

Relative to IMSE-optimal:
Implied scale                  0.952          1.429
WIMSE variance weight          0.537          0.255
WIMSE bias weight              0.463          0.745
```

Stata Snippet 3

```
. rdplot Y X, nbins(20 20) binselect(qs) ///
> graph_options(graphregion(color(white)) ///
> xtitle(Score) ytitle(Outcome))
```

A comparison of the two RD plots in Figure 7 reveals where the observations are located. In the evenly-spaced RD plot in Figure 7(a), there are five bins in the interval [-100,-75] of the running variable. In contrast, in the quantile-spaced RD plot in Figure 7(b), this interval is entirely contained in the first bin. The vast difference in the length of QS and ES bins occurs because, as shown in Table 2, there are very few observations near -100, which leads

to local mean estimates with high variance. This problem is avoided when we choose QS bins, which ensures that each bin has the same number of observations.

3.2 Choosing the Number of Bins

Once the positioning of the bins has been decided by choosing either QS or ES bins, the only remaining choice is the total number of bins on either side of the cutoff—the quantities J_- and J_+. Below we discuss two methods to produce data-driven, automatic RD plots by selecting J_- and J_+, given a choice of QS or ES bins.

3.2.1 Integrated Mean Squared Error (IMSE) Method

The first method we discuss selects the values of J_- and J_+ that minimize an asymptotic approximation to the integrated mean-squared error (IMSE) of the local means estimator, that is, the sum of the expansions of the (integrated) variance and squared bias. If we choose a large number of bins, we have a small bias because the bins are smaller and the local constant fit is better; but this reduction in bias comes at a cost, as increasing the number of bins leads to fewer observations per bin and thus more variability within bin. The IMSE-optimal J_- and J_+ are the numbers of bins that balance squared-bias and variance so that the IMSE is (approximately) minimized.

By construction, choosing an IMSE-optimal number of bins will result in binned sample means that "trace out" the underlying regression function; this is useful to assess the overall shape of the regression function, perhaps to identify potential discontinuities in these functions that occur far from the cutoff. However, the IMSE-optimal method often results in a very smooth plot where the local means nearly overlap with the global polynomial fit, and may not be appropriate to capture the local variability of the data near the cutoff.

The IMSE-optimal values of J_- and J_+ are, respectively,

$$J_-^{\texttt{IMSE}} = \left\lceil \mathscr{C}_-^{\texttt{IMSE}}\, n^{1/3} \right\rceil \qquad \text{and} \qquad J_+^{\texttt{IMSE}} = \left\lceil \mathscr{C}_+^{\texttt{IMSE}}\, n^{1/3} \right\rceil,$$

where n is the total number of observations, $\lceil \cdot \rceil$ denotes the ceiling operator, and the exact form of the constants $\mathscr{C}_-^{\texttt{IMSE}}$ and $\mathscr{C}_+^{\texttt{IMSE}}$ depends on whether ES or QS bins are used (and some features of the underlying data generating process). In practice, the unknown constants $\mathscr{C}_-^{\texttt{IMSE}}$ and $\mathscr{C}_+^{\texttt{IMSE}}$ are estimated using preliminary, objective data-driven procedures.

In order to produce an RD plot that uses an IMSE-optimal number of evenly-spaced bins,

we use the command `rdplot` with the option `binselect = "es"`, but this time omitting the `nbins = c(20 20)` option. When the number of bins is omitted, `rdplot` automatically chooses the number of bins according to the criterion specified with `binselect`. We now produce an RD plot that uses ES bins and chooses the total number of bins on either side of the cutoff to be IMSE-optimal.

R Snippet 4

```
> out = rdplot(Y, X, binselect = "es", x.lim = c(-100, 100), y.lim = c(0,
+ 25), cex.axis = 1.5, cex.lab = 1.5)
> summary(out)
Call: rdplot

Number of Obs.                  2629
Kernel                       Uniform

Number of Obs.                  2314            315
Eff. Number of Obs.             2314            315
Order poly. fit (p)                4              4
BW poly. fit (h)             100.000         99.051
Number of bins scale               1              1

Bins Selected                     11              7
Average Bin Length             9.091         14.150
Median Bin Length              9.091         14.150

IMSE-optimal bins                 11              7
Mimicking Variance bins           40             75

Relative to IMSE-optimal:
Implied scale                  1.000          1.000
WIMSE variance weight          0.500          0.500
WIMSE bias weight              0.500          0.500
```

Stata Snippet 4

```
. rdplot Y X, binselect(es) ///
> graph_options(graphregion(color(white)) ///
> xtitle(Score) ytitle(Outcome))
```

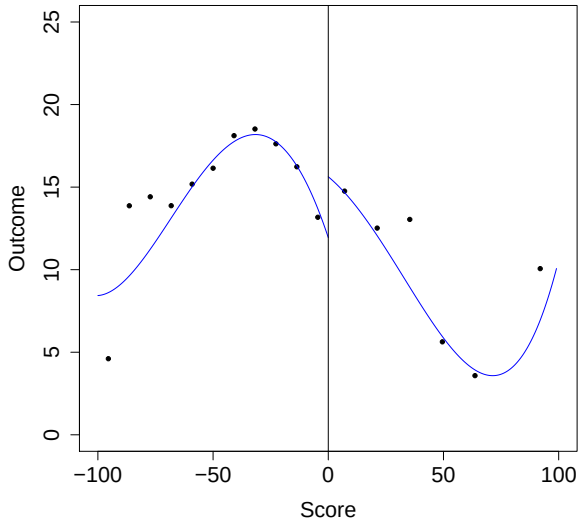


Figure 8: IMSE RD Plot with Evenly-Spaced Bins (Meyersson Data)

The plot is shown in Figure 8. The output reports both the average and the median length of the bins. In the ES case, since each bin has the same length, each bin has length equal to both the average and the median length on each side. The IMSE criterion leads to different numbers of ES bins above and below the cutoff. As shown in the `Bins Selected` row, the IMSE-optimal number of bins is 11 below the cutoff and 7 above it. As a result, the lengths of the bins above and below the cutoff are different: above the cutoff, each bin has a length of 14.150 percentage points, while below the cutoff the bins are smaller, with a length of 9.091. The middle rows show the optimal number of bins according to both the IMSE criterion and the mimicking variance criterion (we discuss the latter in the next subsection). The bottom three rows show the bias and variance weights implied by the chosen number of bins in the IMSE objective function. When the IMSE criterion is used, these weights are always equal to 1/2.

To produce an RD plot that uses an IMSE-optimal number of quantile-spaced bins, we use the option `binselect = "qs"` instead of `binselect = "es"`.

R Snippet 5

```
> out = rdplot(Y, X, binselect = "qs", x.lim = c(-100, 100), y.lim = c(0,
+ 25), cex.axis = 1.5, cex.lab = 1.5)
> summary(out)
Call: rdplot

Number of Obs.                 2629
Kernel                      Uniform

Number of Obs.                 2314          315
Eff. Number of Obs.            2314          315
Order poly. fit (p)               4            4
BW poly. fit (h)            100.000       99.051
Number of bins scale              1            1

Bins Selected                    21           14
Average Bin Length            4.757        7.082
Median Bin Length             2.833        1.429

IMSE-optimal bins                21           14
Mimicking Variance bins          44           41

Relative to IMSE-optimal:
Implied scale                 1.000        1.000
WIMSE variance weight         0.500        0.500
WIMSE bias weight             0.500        0.500
```

Stata Snippet 5

```
. rdplot Y X, binselect(qs) ///
> graph_options(graphregion(color(white)) ///
> xtitle(Score) ytitle(Outcome))
```

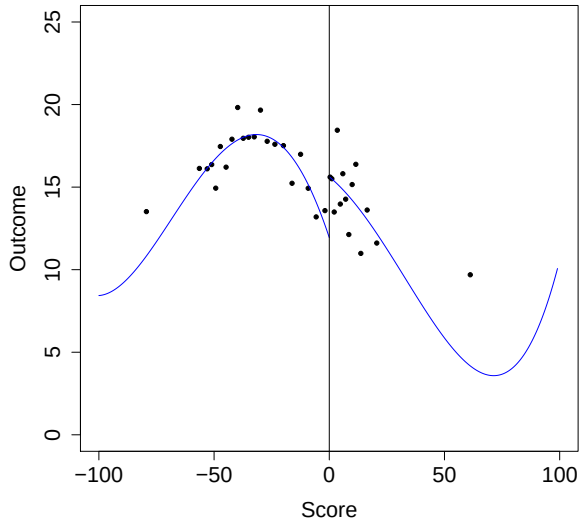


Figure 9: IMSE RD Plot with Quantile-Spaced Bins (Meyersson Data)

The resulting plot is shown in Figure 9. Note that the IMSE-optimal number of QS bins is much larger on both sides, with 21 bins below the cutoff and 14 above it, versus 11 and 7 in the analogous ES plot in Figure 8. The average bin length is 4.7572 below the cutoff, and 7.0821 above it. As expected, the median length of the bins is much smaller than the average length on both sides of the cutoff, particularly above. Since there are very few observations where the Islamic vote margin is above 50%, the length of the last bin above the cutoff must be very large in order to ensure that it contains $315/14 \approx 22$ observations.

3.2.2 Mimicking Variance Method

The second method to select the number of bins chooses the vales of J_- and J_+ so that the binned means have an asymptotic (integrated) variability that is approximately equal to the variability of the raw data. In other words, the number of bins is chosen so that the overall variability of the binned means "mimics" the overall variability in the raw scatter plot of the data. In the Meyersson application, this method involves choosing J_- and J_+ so that the binned means have a total variability approximately equal to the variability illustrated in Figure 5. We refer to this choice of total number of bins as a mimicking variance (MV) choice.

The mimicking-variance values of J_- and J_+ are

$$J_-^{\texttt{MV}} = \left\lceil \mathscr{C}_-^{\texttt{MV}} \frac{n}{\log(n)^2} \right\rceil, \qquad \text{and} \qquad J_+^{\texttt{MV}} = \left\lceil \mathscr{C}_+^{\texttt{MV}} \frac{n}{\log(n)^2} \right\rceil,$$

where again n is the sample size and the exact form of the constants $\mathscr{C}_-^{\texttt{MV}}$ and $\mathscr{C}_+^{\texttt{MV}}$ depends on whether ES or QS bins are used (and some features of the underlying data generating process). These constants are different from those appearing in the IMSE-optimal choices and, in practice, are also estimated using preliminary, objective data-driven procedures.

In general, $J_-^{\texttt{MV}} > J_-^{\texttt{ES}}$ and $J_+^{\texttt{MV}} > J_+^{\texttt{ES}}$. That is, the MV method leads to a larger number of bins than the IMSE method, resulting in an RD plot with more dots representing local means and thus giving a better sense of the variability of the data. In order to produce an RD plot with ES bins and an MV total number of bins on either side, we use the option `binselect ="esmv"`.

R Snippet 6

```
> out = rdplot(Y, X, binselect = "esmv", cex.axis = 1.5, cex.lab = 1.5)
> summary(out)
Call: rdplot

Number of Obs.                  2629
Kernel                       Uniform

Number of Obs.                  2314          315
Eff. Number of Obs.             2314          315
Order poly. fit (p)                4            4
BW poly. fit (h)             100.000       99.051
Number of bins scale               1            1

Bins Selected                     40           75
Average Bin Length             2.500        1.321
Median Bin Length              2.500        1.321

IMSE-optimal bins                 11            7
Mimicking Variance bins           40           75

Relative to IMSE-optimal:
Implied scale                  3.636       10.714
WIMSE variance weight          0.020        0.001
WIMSE bias weight              0.980        0.999
```

Stata Snippet 6

```
. rdplot Y X, binselect(esmv) ///
> graph_options(graphregion(color(white)) ///
> xtitle(Score) ytitle(Outcome))
```

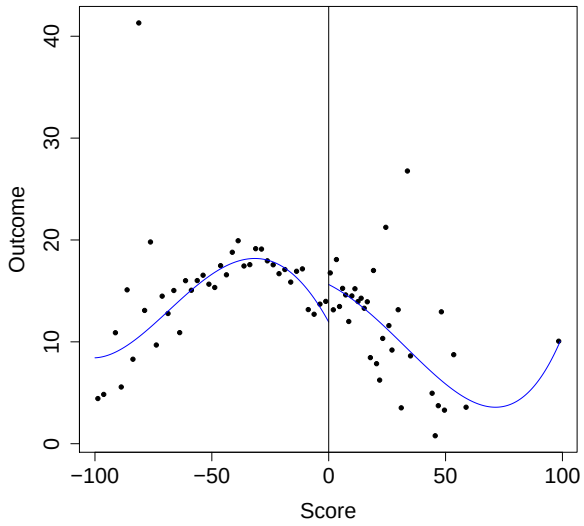


Figure 10: Mimicking Variance RD Plot with Evenly-Spaced Bins (Meyersson Data)

As shown in the output and illustrated in Figure 10, this produces a much higher number of bins than we obtained with the IMSE criterion for both ES and QS bins. The MV total number of bins is 40 below the cutoff and 75 above the cutoff, with length 2.5 and 1.321, respectively. The difference in the chosen number of bins between the IMSE and the MV criteria is dramatic. The middle rows show the number of bins that would have been produced according to the IMSE criterion (11 and 7) and the number of bins that would have been produced according to the MV criterion (40 and 75). This allows for a quick comparison between both methods. Finally, the bottom rows indicate that the chosen number of MV bins on both sides of the cutoff is equivalent to the number of bins that would have been chosen according to an IMSE criterion where, instead of giving the bias and the variance each a weight of 1/2, these weights had been, respectively, 0.020 and 0.980 below the cutoff, and 0.001 and 0.999 above the cutoff. Thus, we see that if we want to justify the MV choice in terms of the IMSE criterion, we must weigh the bias much more than the variance.

Finally, to create an RD plot that chooses the total number of bins according to the MV criterion but uses QS bins, we use the option `binselect = "qsmv"`.

R Snippet 7

```
> out = rdplot(Y, X, binselect = "qsmv", x.lim = c(-100, 100),
+ y.lim = c(0, 25), cex.axis = 1.5, cex.lab = 1.5)
> summary(out)
Call: rdplot

Number of Obs.                  2629
Kernel                       Uniform

Number of Obs.                  2314            315
Eff. Number of Obs.             2314            315
Order poly. fit (p)                4              4
BW poly. fit (h)             100.000         99.051
Number of bins scale               1              1

Bins Selected                     44             41
Average Bin Length             2.270          2.418
Median Bin Length              1.376          0.506

IMSE-optimal bins                 21             14
Mimicking Variance bins           44             41

Relative to IMSE-optimal:
Implied scale                  2.095          2.929
WIMSE variance weight          0.098          0.038
WIMSE bias weight              0.902          0.962
```

Stata Snippet 7

```
. rdplot Y X, binselect(qsmv) ///
> graph_options(graphregion(color(white)) ///
> xtitle(Score) ytitle(Outcome))
```

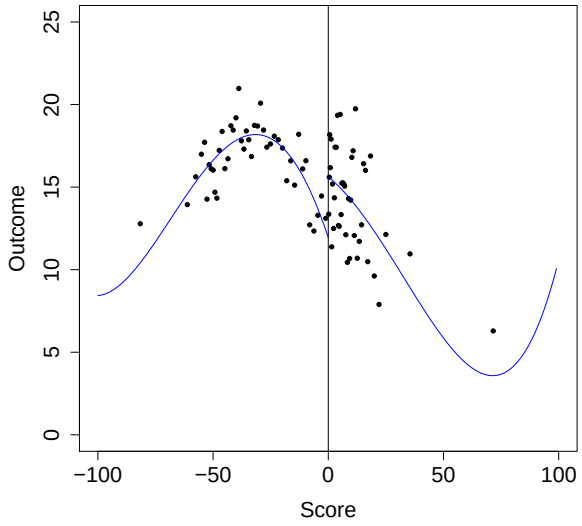


Figure 11: Mimicking Variance RD Plot with Quantile-Spaced Bins—Meyersson Data

The resulting plot is shown in Figure 11. Below the cutoff, the MV number of QS bins is very similar to the MV choice for ES bins (44 versus 40). However, above the cutoff, the MV number of QS bins is much lower than for ES bins (41 versus 75). This occurs because, although the range of the running variable is $[-100, 99.051]$, there are very few observations in the intervals $[-100, -50]$ and $[50, 100]$ far from the cutoff. Since ES bins force the length of the bins to be the same everywhere in the support, the number of ES bins has to be large in order to produce small enough bins to adequately mimic the overall variability of the scatter plot in regions with few observations. In contrast, QS bins can be short near the cutoff and long away from the cutoff, so they can mimic the overall variability by adapting their length to the density of the data.

In sum, bins can be chosen in many different ways. Which method of implementation is most appropriate depends on the researcher's particular goal, for example, illustrating/testing for the overall functional form versus showing the variability of the data. We recommend to start with MV bins to better illustrate the variability of the outcome as a function of the score, ideally comparing ES to QS bins to highlight the distributional features of the score. Then, if needed, the researcher can select the number of bins to be IMSE-optimal in order to explore the global features of the regression function.

3.3 Further Reading

A detailed discussion of RD plots and formal methods for automatic data-driven bin selection are given by Calonico, Cattaneo, and Titiunik (2015a). This paper formalized the commonly used RD plots with evenly-spaced binning, introduced RD plots with quantile-spaced binning, and developed optimal choices for the number of bins in terms of both integrated mean squared error and mimicking variance targets. See also Calonico, Cattaneo, Farrell, and Titiunik (2017) for other features of RD plots, including confidence intervals for the local means in each bin. RD plots are special cases of nonparametric partitioning estimators—see Cattaneo and Farrell (2013), Cattaneo, Farrell, and Feng (2019), and references therein. Finally, see Cattaneo, Crump, Farrell, and Feng (2019) for closely related binscatter methods.

4 The Continuity-Based Approach to RD Analysis

We now discuss empirical methods for estimation and inference in RD designs based on continuity assumptions and extrapolation towards the cutoff point, which rely on large-sample approximations with random potential outcomes under repeated sampling. These methods offer tools useful not only for the analysis of main treatment effects, but also for falsification and validation of the design, which we discuss in Section 5. The approach discussed here is based on formal statistical methods and hence leads to disciplined and objective empirical analysis, which typically has two related but distinct goals: point estimation of RD treatment effect (i.e., give a scalar estimate of the vertical distance between the regression functions at the cutoff), and statistical inference about the RD treatment effect (i.e., construct valid statistical hypothesis tests and confidence intervals).

The methods discussed in this section are based on the continuity conditions underlying Equation (2.1), and generalizations thereof. This framework for RD analysis, which we call the *continuity-based* RD framework, uses methodological tools that directly rely on continuity (and differentiability) assumptions and define $\tau_{\texttt{SRD}}$ as the parameter of interest. In this framework, estimation typically proceeds by using (local to the cutoff) polynomial methods to approximate the regression function $\mathbb{E}[Y_i|X_i = x]$ on each side of the cutoff separately. In practical terms, this involves using least-squares methods to fit a polynomial of the observed outcome on the score. When all the observations are used for estimation, these polynomial fits are global or parametric in nature, like those used in the default RD plots discussed in the previous section. In contrast, when estimation employs only observations with scores near the cutoff, the polynomial fits are local, "flexible," or "non-parametric." Our upcoming discussion focuses exclusively on local polynomial methods, which are by now the standard framework for RD empirical analysis because they offer a good compromise between flexibility and simplicity.

In the second Element (*A Practical Introduction to Regression Discontinuity Designs: Extensions*; Cattaneo, Idrobo, and Titiunik, forthcoming), we discuss an alternative framework for RD analysis that relies on assumptions of local random assignment of the treatment near the cutoff, and employs tools and ideas from the literature on the analysis of experiments. This alternative approach offers a complement to, and a robustness check for, the local polynomial methods based on continuity assumptions that we discuss in the remainder of this Element. Furthermore, the local randomization RD approach can be used in cases where local polynomial methods are invalid or difficult to justify.

4.1 Local Polynomial Approach: Overview

A fundamental feature of the RD design is that, in general, there are no observations for which the score X_i is exactly equal to the cutoff value c: because the running variable is assumed continuous, there are no (or sometimes in practice very few) observations whose score is c or very nearly so. Thus, local extrapolation in RD designs is unavoidable in general. In other words, in order to form estimates of the average control response at the cutoff, $\mathbb{E}[Y_i(0)|X_i = c]$, and of the average treatment response at the cutoff, $\mathbb{E}[Y_i(1)|X_i = c]$, we must rely on observations further away from the cutoff. In the Sharp RD design, for example, the treatment effect $\tau_{\texttt{SRD}}$ is the vertical distance between the $\mathbb{E}[Y_i(1)|X_i = x]$ and $\mathbb{E}[Y_i(0)|X_i = x]$ at $x = c$, as shown in Figure 2, and thus estimation and inference proceed by first *approximating* these unknown regression functions, and then computing the estimated treatment effect and/or the statistical inference procedure of interest. In this context, the key practical issue in RD analysis is how the approximation of the unknown regression functions is done, as this will directly affect the robustness and credibility of the empirical findings.

The problem of approximating an unknown function is well understood: any sufficiently smooth function can be well approximated by a polynomial function, locally or globally, up to an error term. A large literature in statistics has used this principle to develop non-parametric methods based on polynomials or other bases of approximation, relaxing strong parametric assumptions and relying instead on more flexible approximations of the unknown regression function. Applied to the RD point estimation problem, this principle suggests that the unknown regression functions $\mathbb{E}[Y_i(0)|X_i = x]$ and $\mathbb{E}[Y_i(1)|X_i = x]$ can be approximated by a polynomial function of the score. The available statistical results have to be adapted to the RD case, considering the complications that arise because the approximation must occur at the cutoff, which is a boundary point.

Early empirical work employed the idea of polynomial approximation globally, that is, tried to approximate these functions using flexible higher-order polynomials, usually of fourth or fifth order, over the entire support of the data. This global approach is still used in RD plots, as illustrated in the previous section, because the goal there is to illustrate the *entire* unknown regression functions. However, it is now widely recognized that a global polynomial approach does not deliver point estimators and inference procedures with good properties for the RD treatment effect, the main object of interest. The reason is that global polynomial approximations tend to deliver a good approximation overall, but a poor approximation at boundary points—a problem known as Runge's phenomenon in approximation theory. Moreover, global approximations can induce counter-intuitive weighting schemes, for example, when the point estimator is heavily influenced by observations far from the boundary.

Since the RD point estimator is defined at a boundary point, global polynomial methods can lead to unreliable RD point estimators, and thus the conclusions from a global parametric RD analysis can be highly misleading. For these reasons, we recommend against using global polynomial methods for formal RD analysis.

Modern RD empirical work employs local polynomial methods, which focus on approximating the regression functions only near the cutoff. Because this approach localizes the polynomial fit to the cutoff (discarding observations sufficiently far away) and employs a low-order polynomial approximation (usually linear or quadratic), it is substantially more robust and less sensitive to boundary and overfitting problems. Furthermore, this approach can be viewed formally as a non-parametric local polynomial approximation, which has also aided the development of a comprehensive toolkit of statistical and econometric results for estimation and inference. In contrast to global higher-order polynomials, local lower-order polynomial approximations can be viewed as intuitive approximations with a potential misspecification of the functional form of the regression function near the cutoff, which can be modeled and understood formally, with the advantage that they are less sensitive to outliers or other extreme features of the data generating process far from the cutoff. Local polynomial methods employ only observations close to the cutoff, and interpret the polynomial used as a local approximation, not necessarily as a correctly specified model.

The statistical properties of local polynomial estimation and inference depend crucially on the accuracy of the approximation near the cutoff, which is controlled by the size of the neighborhood around the cutoff where the local polynomial is fit. In the upcoming sections, we discuss the modern local polynomial methods for RD analysis, and explain all the steps involved in their implementation for both estimation and inference. We also discuss several extensions and modifications, including the inclusion of predetermined covariates and the use of cluster-robust standard errors.

4.2 Local Polynomial Point Estimation

Local polynomial methods implement linear regression fits using only observations near the cutoff point, separately for control and treatment units. Specifically, this approach uses only observations that are between $c - h$ and $c + h$, where $h > 0$ is a so-called bandwidth that determines the size of the neighborhood around the cutoff where the empirical RD analysis is conducted. Within this bandwidth, it is common to adopt a weighting scheme to ensure that the observations closer to c receive more weight than those further away; the weights are determined by a kernel function $K(\cdot)$. The local polynomial approach can be understood and

analyzed formally as non-parametric, in which case the fit is taken as an approximation to the unknown underlying regression functions within the region determined by the bandwidth.

Local polynomial estimation consists of the following basic steps.

1. Choose a polynomial order p and a kernel function $K(\cdot)$.

2. Choose a bandwidth h.

3. For observations above the cutoff (i.e., observations with $X_i \geq c$), fit a weighted least squares regression of the outcome Y_i on a constant and $(X_i - c), (X_i - c)^2, \ldots, (X_i - c)^p$, where p is the chosen polynomial order, with weight $K(\frac{X_i - c}{h})$ for each observation. The estimated intercept from this local weighted regression, $\hat{\mu}_+$, is an estimate of the point $\mu_+ = \mathbb{E}[Y_i(1)|X_i = c]$:

$$\hat{\mu}_+ \quad : \quad \hat{Y}_i = \hat{\mu}_+ + \hat{\mu}_{+,1}(X_i - c) + \hat{\mu}_{+,2}(X_i - c)^2 + \cdots + \hat{\mu}_{+,p}(X_i - c)^p.$$

4. For observations below the cutoff (i.e., observations with $X_i < c$), fit a weighted least-squares regression of the outcome Y_i on a constant and $(X_i - c), (X_i - c)^2, \ldots, (X_i - c)^p$, where p is the chosen polynomial order, with weight $K(\frac{X_i - c}{h})$ for each observation. The estimated intercept from this local weighted regression, $\hat{\mu}_-$, is an estimate of the point $\mu_- = \mathbb{E}[Y_i(0)|X_i = c]$:

$$\hat{\mu}_- \quad : \quad \hat{Y}_i = \hat{\mu}_- + \hat{\mu}_{-,1}(X_i - c) + \hat{\mu}_{-,2}(X_i - c)^2 + \cdots + \hat{\mu}_{-,p}(X_i - c)^p.$$

5. Calculate the Sharp RD point estimate: $\hat{\tau}_{\texttt{SRD}} = \hat{\mu}_+ - \hat{\mu}_-$.

A graphical representation of local polynomial RD point estimation is given in Figure 12, where a polynomial of order one ($p = 1$) is fit within bandwidth h_1; observations outside this bandwidth are not used in the estimation. The RD effect is $\tau_{\texttt{SRD}} = \mu_+ - \mu_-$ and the local polynomial estimator of this effect is $\hat{\mu}_+ - \hat{\mu}_-$. Local polynomial methods produce the fit employing the raw data, not the binned data typically reported in the RD plots.

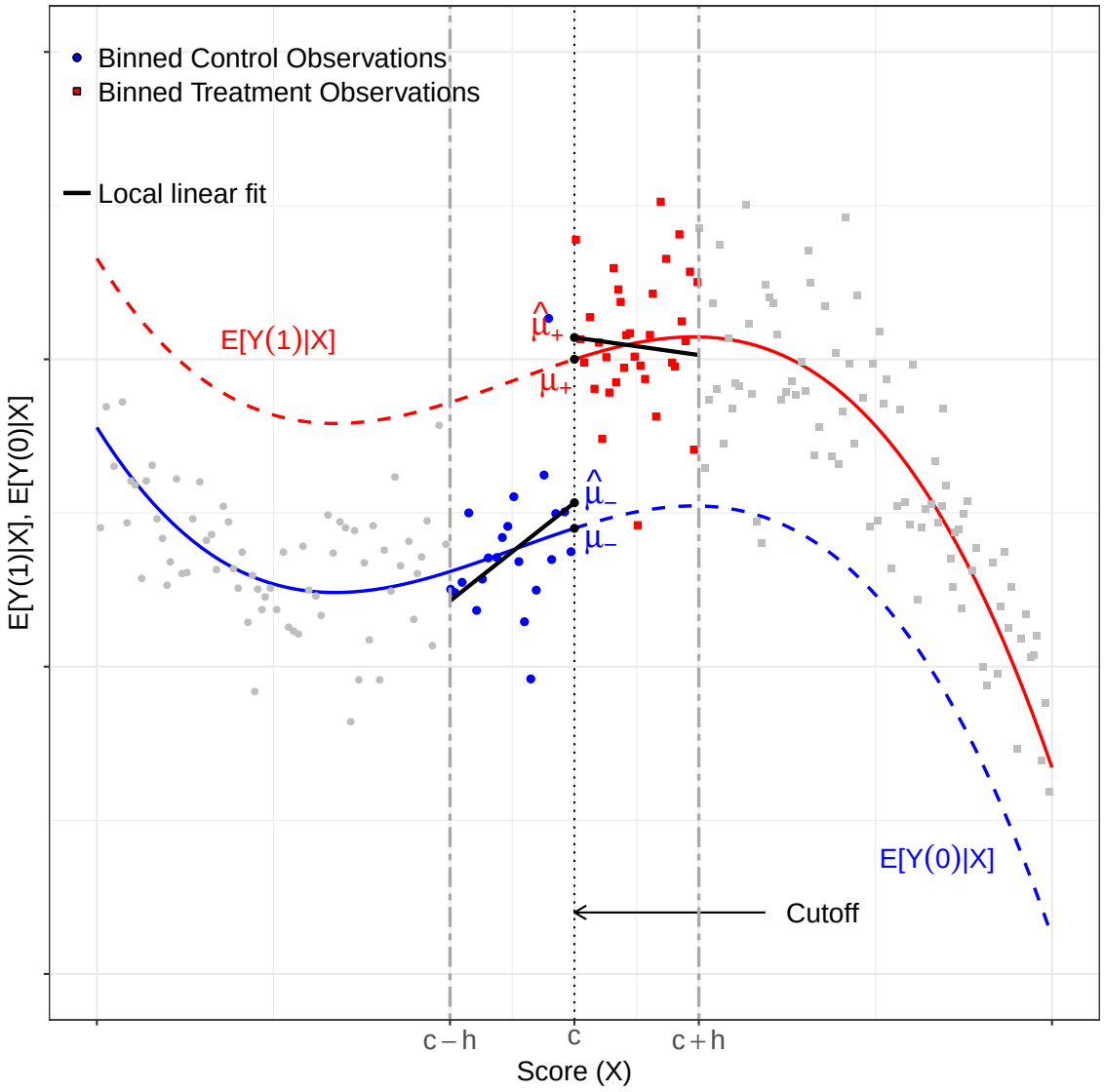


Figure 12: RD Estimation with Local Polynomial

The implementation of the local polynomial approach thus requires the choice of three main ingredients: the kernel function $K(\cdot)$, the order of the polynomial p, and the bandwidth h. We now turn to a discussion of each of these choices.

4.2.1 Choice of Kernel Function and Polynomial Order

The kernel function $K(\cdot)$ assigns non-negative weights to each transformed observation $\frac{X_i - c}{h}$, based on the distance between the observation's score X_i and the cutoff c. The recommended choice is the triangular kernel function, $K(u) = (1 - |u|)\mathbb{1}(|u| \leq 1)$, because when used in

conjunction with a bandwidth that optimizes the mean squared error (MSE), it leads to a point estimator with optimal properties (more details about MSE-optimal bandwidths are given below). As illustrated in Figure 13, the triangular kernel function assigns zero weight to all observations with score outside the interval $[c-h, c+h]$, and positive weights to all observations within this interval. The weight is maximized at $X_i = c$, and declines symmetrically and linearly as the value of the score gets farther from the cutoff.

Despite the desirable asymptotic optimality properties of the triangular kernel, researchers sometimes prefer to use the more simple uniform kernel $K(u) = \mathbb{1}(|u| \leq 1)$, which also gives zero weight to observations with score outside $[c-h, c+h]$, but equal weight to all observations whose scores are within this interval, see Figure 13. Employing a local linear estimation with bandwidth h and uniform kernel is therefore equivalent to estimating a simple linear regression without weights using only observations whose distance from the cutoff is at most h. A uniform kernel minimizes the asymptotic variance of the local polynomial estimator under some technical conditions. A third weighting scheme sometimes encountered in practice is the Epanechnikov kernel, $K(u) = (1-u^2)\mathbb{1}(|u| \leq 1)$, also depicted in Figure 13, which gives a quadratic decaying weight to observations with $X_i \in [c-h, c+h]$ and zero weight to the rest. In practice, estimation and inference results are typically not very sensitive to the particular choice of kernel used.

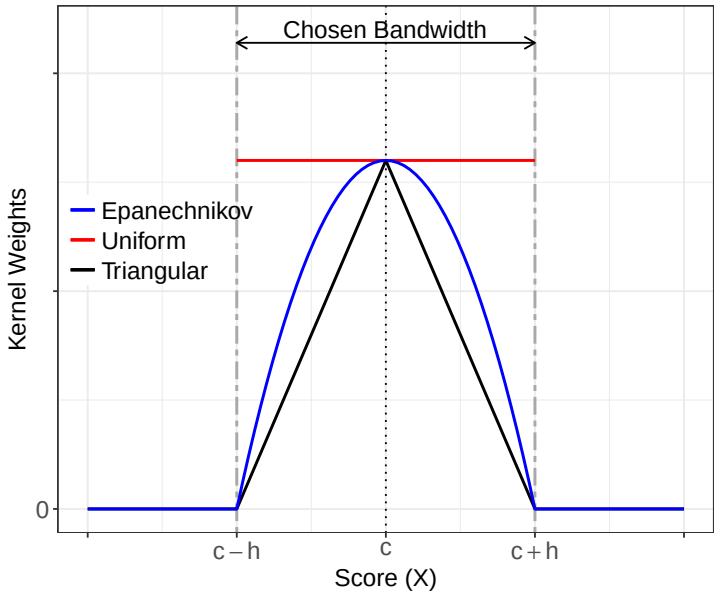


Figure 13: Different Kernel Weights for RD Estimation

A more consequential decision is the choice of the local polynomial order, which must consider various factors. First, a polynomial of order zero—a constant fit—has undesirable

theoretical properties at 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 also increases the variability of the treatment effect estimator. Third, as mentioned above, higher-order polynomials tend to produce overfitting of the data and lead to unreliable results near boundary points. Combined, these factors have led researchers to prefer the local linear RD estimator, which by now is the default point estimator in most applications. In finite samples, of course, the ranking between different local polynomial estimators may be different, but in general the local linear estimator seems to deliver a good trade-off between simplicity, precision, and stability in RD settings.

Although it may seem at first that a linear polynomial is not flexible enough, an appropriately chosen bandwidth will adjust to the chosen polynomial order so that the linear approximation to the unknown regression functions is reliable. We turn to this issue below.

4.2.2 Bandwidth Selection and Implementation

The bandwidth h controls the width of the neighborhood around the cutoff that is used to fit the local polynomial that approximates the unknown regression functions. The choice of h is fundamental for the analysis and interpretation of RD designs, as h directly affects the properties of local polynomial estimation and inference procedures, and empirical findings are often sensitive to its particular value.

Figure 14 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 polynomial. Indeed, inside the interval $[c-h_2, c+h_2]$, a linear approximation yields an estimated RD effect equal to $\hat{\mu}_+(h_2) - \hat{\mu}_-(h_2)$ (distance between points `c` and `d`), which is considerably different from the true effect $\tau_{\texttt{SRD}}$. Thus, a linear fit within bandwidth h_2 results in a poor approximation because of misspecification error. However, reducing the bandwidth from h_2 to h_1 improves the linear approximation considerably, as now the estimated RD effect $\hat{\mu}_+(h_1) - \hat{\mu}_-(h_1)$ (distance between points `a` and `b`) is much closer to the population treatment effect $\tau_{\texttt{SRD}}$. The reason is that the regression functions are nearly linear in the interval $[c-h_1, c+h_1]$, and therefore the linear approximation results in a smaller misspecification error when the bandwidth shrinks from h_2 to h_1. This illustrates the general principle that, given a polynomial order, the accuracy of the approximation can always be improved by reducing the bandwidth.

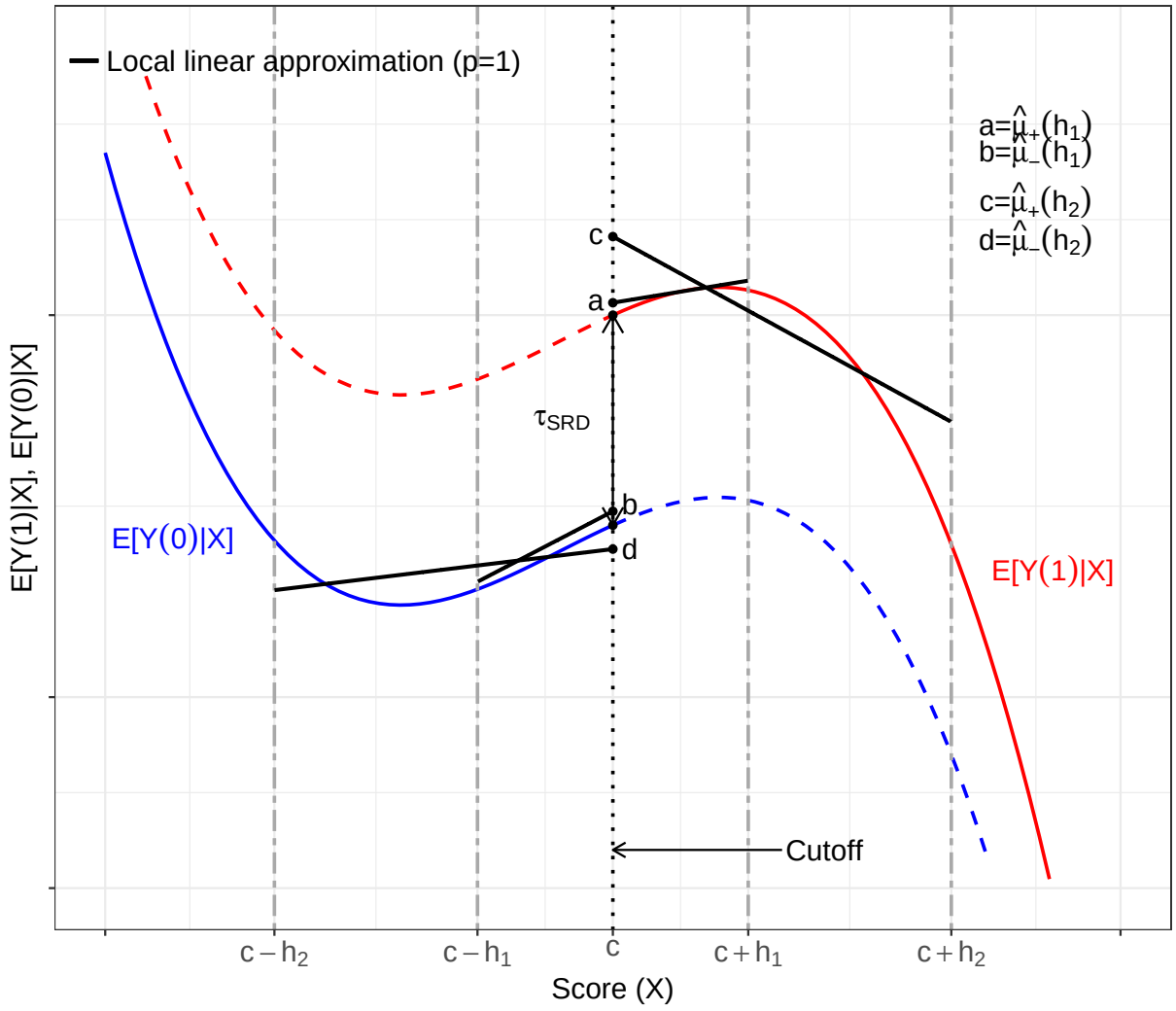


Figure 14: Bias in Local Approximations

Choosing a smaller h will reduce the misspecification error (also known as "smoothing bias") of the local polynomial approximation, but will simultaneously tend to increase the variance of the estimated coefficients because fewer observations will be available for estimation. On the other hand, a larger h will result in more smoothing bias if the unknown function differs considerably from the polynomial model used for approximation, but will reduce the variance because the number of observations in the interval $[c-h, c+h]$ will be larger. For this reason, the choice of bandwidth is said to involve a "bias-variance trade-off."

Since RD empirical results are often sensitive to the choice of bandwidth, it is important to select h in a data-driven, automatic way to avoid specification searching and ad hoc decisions. Most bandwidth selection methods try to balance some form of bias-variance trade-off (sometimes involving other features of the estimator, inference procedure, and data

generating process). The most popular approach in practice seeks to minimize the MSE of the local polynomial RD point estimator, $\hat{\tau}_{\texttt{SRD}}$, given a choice of polynomial order and kernel function. Since the MSE of an estimator is the sum of its squared bias and its variance, this approach effectively chooses h to optimize a bias-variance trade-off. The precise procedure involves deriving an asymptotic approximation to the MSE of $\hat{\tau}_{\texttt{SRD}}$, optimizing it with respect to h, and using data-driven methods to estimate the unknown quantities in the resulting formula of the optimal h.

The general form of the approximate (conditional) MSE for the RD treatment effect is

$$\mathsf{MSE}(\hat{\tau}_{\texttt{SRD}}) = \mathsf{Bias}^2(\hat{\tau}_{\texttt{SRD}}) + \mathsf{Variance}(\hat{\tau}_{\texttt{SRD}}) = \mathscr{B}^2 + \mathscr{V},$$

where the approximate (conditional) bias and variance of the estimator are

$$\mathscr{B} = h^{2(p+1)}\,\mathcal{B} \qquad \text{and} \qquad \mathscr{V} = \frac{1}{nh}\,\mathcal{V},$$

respectively. The quantities $\mathcal{B}$ and $\mathcal{V}$ represent, respectively, the (leading) bias and variance of the RD point estimator $\hat{\tau}_{\texttt{SRD}}$, not including the rates controlled by the sample size and bandwidth choice. Although we omit the technical details, we present the general form of $\mathscr{B}$ and $\mathscr{V}$ to clarify the most important trade-offs involved in the choice of an MSE-optimal bandwidth for the local polynomial RD estimate, and because these quantities will be used for inference below.

The general form of the bias $\mathscr{B}$ is determined by the bandwidth $h^{2(p+1)}$ and the quantities

$$\mathcal{B} = \mathcal{B}_+ - \mathcal{B}_-, \qquad \mathcal{B}_- \approx \mu_-^{(p+1)} B_-, \qquad \mathcal{B}_+ \approx \mu_+^{(p+1)} B_+$$

where the derivatives

$$\mu_+^{(p+1)} = \lim_{x \downarrow c} \frac{d^{p+1}\mathbb{E}[Y_i(1)|X=x]}{dx^{p+1}} \qquad \text{and} \qquad \mu_-^{(p+1)} = \lim_{x \uparrow c} \frac{d^{p+1}\mathbb{E}[Y_i(0)|X=x]}{dx^{p+1}}$$

are related to the "curvature" of the unknown regression functions for treatment and control units, respectively, and the known constants B_+ and B_- are related to the kernel function and the order of the polynomial used. These calculations assume a common bandwidth h, but the expressions can be extended to allow for different bandwidths on the left and right of the cutoff.

The bias term $\mathscr{B}$ associated with the local polynomial RD point estimator of order p, $\hat{\tau}_{\texttt{SRD}}$, depends on the $(p+1)$th derivatives of the regression functions $\mathbb{E}[Y_i(1)|X=x]$ and

$\mathbb{E}[Y_i(0)|X = x]$ with respect to the running variable. This is a more formal characterization of the phenomenon we illustrated in Figure 14. When we approximate $\mathbb{E}[Y_i(1)|X = x]$ and $\mathbb{E}[Y_i(0)|X = x]$ with a local polynomial of order p, that approximation has an error (unless $\mathbb{E}[Y_i(1)|X = x]$ and $\mathbb{E}[Y_i(0)|X = x]$ happen to be polynomials of at most order p). The leading term of the approximation error is related to the derivative of order $p + 1$, that is, the order following the polynomial order used to estimate $\tau_{\texttt{SRD}}$. For example, as illustrated in Figure 14, if we use a local linear polynomial to estimate $\tau_{\texttt{SRD}}$, our approximation by construction ignores the second-order term (which depends on the second derivative of the function), and all higher-order terms (which depend on the higher-order derivatives). Thus, the leading bias associated with a local linear estimator depends on the second derivatives of the regression functions, which are the leading terms in the error of approximation incurred when we set $p = 1$. Similarly, if we use a local quadratic polynomial to estimate $\hat{\tau}_{\texttt{SRD}}$, the leading bias will depend on the third derivatives of the regression function.

The variance term $\mathscr{V}$ depends on the sample size and bandwidth through the expression $\frac{1}{nh}$ and also involves the quantities

$$\mathcal{V} = \mathcal{V}_- + \mathcal{V}_+, \qquad \mathcal{V}_- \approx \frac{\sigma_-^2}{f} V_-, \qquad \mathcal{V}_+ \approx \frac{\sigma_+^2}{f} V_+$$

where

$$\sigma_+^2 = \lim_{x \downarrow c} \mathbb{V}[Y_i(1)|X_i = x] \qquad \text{and} \qquad \sigma_-^2 = \lim_{x \uparrow c} \mathbb{V}[Y_i(0)|X_i = x]$$

capture the conditional variability of the outcome given the score at the cutoff for treatment and control units, respectively, f denotes the density of the score variable at the cutoff, and the known constants V_- and V_+ are related to the kernel function and the order of the polynomial used.

As the number of observations near the cutoff decreases (e.g., as the density f decreases), the contribution of the variance term to the MSE increases, and vice versa as the number of observations near the cutoff increases. This captures the intuition that the variability of the RD point estimator will partly depend on the density of observations near the cutoff. Similarly, an increase (decrease) in the conditional variability of the outcome given the score will increase (decrease) the MSE of the RD point estimators.

In order to obtain an MSE-optimal point estimator $\hat{\tau}_{\texttt{SRD}}$, we choose the bandwidth that minimizes the MSE approximation:

$$\min_{h>0} \left(h^{2(p+1)} \, \mathcal{B}^2 + \frac{1}{nh} \, \mathcal{V} \right),$$

which leads to the MSE-optimal bandwidth choice

$$h_{\texttt{MSE}} = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2} \right)^{1/(2p+3)} n^{-1/(2p+3)}.$$

This formula formally incorporates the bias-variance trade-off mentioned above. It follows that $h_{\texttt{MSE}}$ is proportional to $n^{-1/(2p+3)}$, and that this MSE-optimal bandwidth increases with $\mathcal{V}$ and decreases with $\mathcal{B}$. In other words, a larger asymptotic variance will lead to a larger MSE-optimal bandwidth; this is intuitive, as a larger bandwidth will include more observations in the estimation and thus reduce the variance of the resulting point estimator. In contrast, a larger asymptotic bias will lead to a smaller bandwidth, as a smaller bandwidth will reduce the approximation error and reduce the bias of the resulting point estimator.

Another way to see this trade-off is to note that if we chose a bandwidth $h > h_{\texttt{MSE}}$, decreasing h would lead to a reduction in the approximation error and an increase in the variability of the point estimator, but the MSE reduction caused by the decrease in bias would be larger than the MSE increase caused by the variance increase, leading to a smaller MSE overall. In other words, when $h > h_{\texttt{MSE}}$, it is possible to reduce the misspecification error without increasing the MSE. In contrast, when we set $h = h_{\texttt{MSE}}$, both increasing and decreasing the bandwidth necessarily lead to a higher MSE.

Given the quantities $\mathcal{V}$ and $\mathcal{B}$, increasing the sample size n leads to a smaller optimal $h_{\texttt{MSE}}$. This is also intuitive: as a larger sample becomes available, both bias and variance are reduced, because it is possible to reduce the error in the approximation by reducing the bandwidth without paying a penalty in added variability (as the larger number of available observations compensates for the bandwidth reduction).

In some applications, it may be useful to choose different bandwidths on each side of the cutoff. Since the RD treatment effect $\tau_{\texttt{SRD}} = \mu_+ - \mu_-$ is simply the difference of two (one-sided) estimates, allowing for two distinct bandwidth choices can be accomplished by considering an MSE approximation for each estimate separately. In other words, two different bandwidths can be selected for $\hat{\mu}_+$ and $\hat{\mu}_-$, and then used to form the RD treatment effect estimator. Practically, this is equivalent to choosing an asymmetric neighborhood near the cutoff of the form $[c - h_-, c + h_+]$, where h_- and h_+ denote the control (left) and treatment (right) bandwidths, respectively. In this case, the MSE-optimal choices are given by

$$h_{\texttt{MSE},-} = \left(\frac{\mathcal{V}_-}{2(p+1)\mathcal{B}_-^2}\right)^{1/(2p+3)} n_-^{-1/(2p+3)} \tag{4.1}$$

$$h_{\texttt{MSE},+} = \left(\frac{\mathcal{V}_+}{2(p+1)\mathcal{B}_+^2}\right)^{1/(2p+3)} n_+^{-1/(2p+3)}. \tag{4.2}$$

These bandwidth choices will be most practically relevant when the bias and/or variance of the control and treatment groups differ substantially, for example, because of different curvature of the unknown regression functions, or different conditional variance of the outcome given the score near the cutoff.

In practice, the optimal bandwidth selectors described above (and variants thereof) are implemented by constructing preliminary plug-in estimates of the unknown quantities entering their formulas. For example, given a bandwidth choice and sample size, the misspecification biases $\mathscr{B}_+$ and $\mathscr{B}_-$ are estimated by forming preliminary "curvature" estimates $\hat{\mu}_-^{(p+1)}$ and $\hat{\mu}_+^{(p+1)}$, which are constructed using a local polynomial of order $q \geq p+1$ with bias bandwidth b, not necessarily equal to h. The resulting estimators take the form

$$\hat{\mathscr{B}} = h^{2(p+1)}\,\hat{\mathcal{B}}, \qquad \hat{\mathcal{B}} = \hat{\mathcal{B}}_+ - \hat{\mathcal{B}}_-, \qquad \hat{\mathcal{B}}_+ = \hat{\mu}_+^{(p+1)} B_+, \qquad \hat{\mathcal{B}}_- = \hat{\mu}_-^{(p+1)} B_-,$$

where the quantities B_- and B_+ are readily implementable given the information available (e.g., data, bandwidth choices, kernel choice, etc.). Similarly, a variance estimator is

$$\hat{\mathscr{V}} = \frac{1}{nh}\,\hat{\mathcal{V}}, \qquad \hat{\mathcal{V}} = \hat{\mathcal{V}}_+ + \hat{\mathcal{V}}_-,$$

where the estimators $\hat{\mathcal{V}}_-$ and $\hat{\mathcal{V}}_+$ are usually constructed using plug-in pre-asymptotic formulas capturing the asymptotic variance of the estimates on the left and right of the cutoff, respectively. Natural choices are some version of heteroskedasticity-consistent standard error formulas or modifications thereof allowing for clustered data, all of which are implemented in the `rdrobust` software

Given these ingredients, data-driven MSE-optimal bandwidth selectors are easily constructed for the RD treatment effect (i.e., one common bandwidths on both sides of the cutoff) or for each of the two regression function estimators at the cutoff (i.e., two distinct bandwidths). For example, once a preliminary bandwidth choice is available to construct the

above estimators, the MSE-optimal bandwidth choice is

$$\hat{h}_{\texttt{MSE}} = \left(\frac{\hat{\mathcal{V}}}{2(p+1)\hat{\mathcal{B}}^2} \right)^{1/(2p+3)} n^{-1/(2p+3)},$$

and similarly for $\hat{h}_{\texttt{MSE},+}$ and $\hat{h}_{\texttt{MSE},-}$.

A potential drawback of the MSE bandwidth selection approach is that in some applications the estimated biases may be close to zero, leading to poor behavior of the resulting bandwidth selectors. To handle this computational issue, it is common to include a "regularization" term $\mathcal{R}$ to avoid small denominators in small samples. For example, in the case of a common bandwidth, the alternative formula is

$$h_{\texttt{MSE}} = \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2 + \mathcal{R}} \right)^{1/(2p+3)} n^{-1/(2p+3)},$$

where the extra term $\mathcal{R}$ can be justified theoretically but requires additional preliminary estimators when implemented. Empirically, since $\mathcal{R}$ is in the denominator, including a regularization term will always lead to a smaller $h_{\texttt{MSE}}$. This idea is also used in the case of $h_{\texttt{MSE},-}$ and $h_{\texttt{MSE},+}$, and other related bandwidth selection procedures. We discuss how to include and exclude a regularization term in practice in Section 4.2.4.

4.2.3 Optimal Point Estimation

Given the choice of polynomial order p and kernel function $K(\cdot)$, the local polynomial RD point estimator $\hat{\tau}_{\texttt{SRD}}$ is implemented for a choice of bandwidth h. Selecting either a common MSE-optimal bandwidth for $\hat{\tau}_{\texttt{SRD}} = \hat{\mu}_- - \hat{\mu}_+$, or two distinct MSE-optimal bandwidths for its ingredients $\hat{\mu}_-$ and $\hat{\mu}_+$, leads to an RD point estimator that is both consistent and MSE-optimal, in the sense that it achieves the fastest rate of decay in an MSE sense. Furthermore, it can be shown that the triangular kernel is the MSE-optimal choice for point estimation. Because of these optimality properties, and the fact that the procedures are data driven and objective, modern RD empirical work routinely employs some form of automatic MSE-optimal bandwidth selection with triangular kernel, and reports the resulting MSE-optimal point estimator of the RD treatment effect.

4.2.4 Point Estimation in Practice

We now return to the Meyersson application to illustrate RD point estimation using local polynomials. First, we use standard least-squares commands to emphasize that local polynomial point estimation is simply a weighted least-squares fit.

We start by choosing an ad hoc bandwidth $h = 20$, postponing the illustration of optimal bandwidth selection until the following section. Within this arbitrary bandwidth choice, we can construct the local linear RD point estimation with a uniform kernel using standard least-squares routines. As mentioned above, a uniform kernel simply means that all observations outside $[c - h, c + h]$ are excluded, and all observations inside this interval are weighted equally.

R Snippet 8

```
> out = lm(Y[X < 0 & X >= -20] ~ X[X < 0 & X >= -20])
> left_intercept = out$coefficients[1]
> print(left_intercept)
(Intercept)
   12.62254
> out = lm(Y[X >= 0 & X <= 20] ~ X[X >= 0 & X <= 20])
> right_intercept = out$coefficients[1]
> print(right_intercept)
(Intercept)
   15.54961
> difference = right_intercept - left_intercept
> print(paste("The RD estimator is", difference, sep = " "))
[1] "The RD estimator is 2.92707507543107"
```

Stata Snippet 8

```
. reg Y X if X < 0 & X >= -20
. matrix coef_left = e(b)
. local intercept_left = coef_left[1, 2]
. reg Y X if X >= 0 & X <= 20
. matrix coef_right = e(b)
. local intercept_right = coef_right[1, 2]
. local difference = `intercept_right' - `intercept_left'
The RD estimator is `difference'
The RD estimator is 2.92707507543108
```

The results indicate that within this ad hoc bandwidth of 20 percentage points, the percentage of women aged 15 to 20 who completed high school increases by about 2.927 percentage points with an Islamic victory: about 15.55% of women in this age group had

completed high school by 2000 in municipalities where the Islamic party barely won the 1994 mayoral elections, while the analogous percentage in municipalities where the Islamic party was barely defeated is about 12.62%.

We now show that the same point estimator can be obtained by fitting a single linear regression that includes an interaction between the treatment indicator and the score—both approaches are algebraically equivalent.

R Snippet 9

```
> T_X = X * T
> out = lm(Y[X >= -20 & X <= 20] ~ X[X >= -20 & X <= 20] + T[X >=
+ -20 & X <= 20] + T_X[X >= -20 & X <= 20])
> summary(out)

Call:
lm(formula = Y[X >= -20 & X <= 20] ~ X[X >= -20 & X <= 20] +
    T[X >= -20 & X <= 20] + T_X[X >= -20 & X <= 20])

Residuals:
    Min      1Q  Median      3Q     Max
-17.373  -7.718  -0.755   6.384  33.697

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              12.62254    0.77459  16.296  < 2e-16 ***
X[X >= -20 & X <= 20]    -0.24807    0.06723  -3.690 0.000238 ***
T[X >= -20 & X <= 20]     2.92708    1.23529   2.370 0.018024 *
T_X[X >= -20 & X <= 20]   0.12612    0.12459   1.012 0.311667
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.316 on 884 degrees of freedom
Multiple R-squared:  0.01721,        Adjusted R-squared:  0.01387
F-statistic: 5.159 on 3 and 884 DF,  p-value: 0.00154
```

Stata Snippet 9

```
. gen T_X = X * T
. reg Y X T T_X if X >= -20 & X <= 20
```

The coefficient on the treatment indicator is 2.92708, the same value we obtained by subtracting the intercepts in the two separate regressions.

To produce the same point estimation with a triangular kernel instead of a uniform kernel, we simply use a least-squares routine with weights. First, we create the weights according to the triangular kernel formula.

R Snippet 10

```
> w = NA
> w[X < 0 & X >= -20] = 1 - abs(X[X < 0 & X >= -20]/20)
> w[X >= 0 & X <= 20] = 1 - abs(X[X >= 0 & X <= 20]/20)
```

Stata Snippet 10

```
. gen weights = .
. replace weights = (1 - abs(X / 20)) if X < 0 & X >= -20
. replace weights = (1 - abs(X / 20)) if X >= 0 & X <= 20
```

Then, we use the weights in the least-squares fit.

R Snippet 11

```
> out = lm(Y[X < 0] ~ X[X < 0], weights = w[X < 0])
> left_intercept = out$coefficients[1]
> out = lm(Y[X >= 0] ~ X[X >= 0], weights = w[X >= 0])
> right_intercept = out$coefficients[1]
> difference = right_intercept - left_intercept
> print(paste("The RD estimator is", difference, sep = " "))
[1] "The RD estimator is 2.93731873078712"
```

Stata Snippet 11

```
. reg Y X [aw = weights] if X < 0 & X >= -20
. matrix coef_left = e(b)
. local intercept_left = coef_left[1, 2]
. reg Y X [aw = weights] if X >= 0 & X <= 20
. matrix coef_right = e(b)
. local intercept_right = coef_right[1, 2]
. local difference = `intercept_right' - `intercept_left'
The RD estimator is `difference'
The RD estimator is 2.937318684658599
```

Note that, with h and p fixed, changing the kernel from uniform to triangular alters the point estimator only slightly, from about 2.9271 to 2.9373. This is typical; point estimates tend to be relatively stable with respect to the choice of kernel.

Although using standard least-squares estimation routines is useful to clarify the algebraic mechanics behind local polynomial point estimation, the confidence intervals and standard

errors provided by these routines will be generally invalid for our purposes, a point we discuss extensively in the upcoming sections. Thus, from this point on, we employ the `rdrobust` software package, which is specifically tailored to RD designs and includes several functions to conduct local polynomial bandwidth selection, RD point estimation, and RD inference using a fully non-parametric and internally coherent methodology.

To replicate the previous point estimators using the command `rdrobust`, we use the options `p` to set the order of the polynomial, `kernel` to set the kernel, and `h` to choose the bandwidth manually. By default, `rdrobust` sets the cutoff value to zero, but this can be changed with the option `c`. We first use `rdrobust` to implement a local linear RD point estimator with $h = 20$ and uniform kernel.

R Snippet 12

```
> out = rdrobust(Y, X, kernel = "uniform", p = 1, h = 20)
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                       Manual
Kernel                       Uniform
VCE method                        NN

Number of Obs.                2314         315
Eff. Number of Obs.            608         280
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                 20.000      20.000
BW bias (b)                 20.000      20.000
rho (h/b)                    1.000       1.000

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     2.927     1.235     2.371     0.018     [0.507 , 5.347]
        Robust         -         -     1.636     0.102    [-0.582 , 6.471]
================================================================================
```

Stata Snippet 12

```
. rdrobust Y X, kernel(uniform) p(1) h(20)
```

The output includes many details. The four uppermost rows indicate that the total number of observations is 2,629, the bandwidth is chosen manually, and the observations

are weighed with a uniform kernel. The final line indicates that the variance-covariance estimator (`VCE`) is constructed using nearest-neighbor (NN) estimators instead of sums of squared residuals (this default behavior can be changed with the option `vce`); we discuss details on variance estimation further below in the context of RD inference.

The middle rows resemble the output of `rdplot` in that they are divided in two columns that give information separately for the observations above (`Right`) and below (`Left`) the cutoff. The first row shows that the $2,629$ observations are split into $2,314$ (control) observations below the cutoff, and 315 (treated) observations above the cutoff. The second row shows the effective number of observations that are used for estimation of the RD effect, that is, the number of observations whose scores are within distance h from the cutoff, $X_i \in [c-h, c+h]$. The output indicates that there are 608 observations with $X_i \in [c-h, c)$, and 280 observations with $X_i \in [c, c+h]$. The third line shows the order of the local polynomial used to estimate the main RD effect, $\tau_{\texttt{SRD}}$, which in this case is equal to $p = 1$. The bandwidth used to estimate $\tau_{\texttt{SRD}}$ is shown on the fifth line, `BW est. (h)`, where we see that the same bandwidth $h = 20$ was used to the left and right of the cutoff. We defer discussion of `Order Bias (q)`, `BW bias (b)`, and `rho (h/b)` until we discuss inference methods.

The bottom rows show the estimation results. The RD point estimator, reported in the first row of the `Coef.` column, is $\hat{\tau}_{\texttt{SRD}} = 2.927$, indicating that in municipalities where the Islamic party barely won, the educational attainment of women is roughly 3 percentage points higher than in municipalities where the party barely lost. As expected, this number is identical to the number we obtained with the least-squares function `lm` in `R` or the command `reg` in `Stata`.

The `rdrobust` routine also allows us to easily estimate the RD effect using triangular instead of uniform kernel weights.

R Snippet 13

```
> out = rdrobust(Y, X, kernel = "triangular", p = 1, h = 20)
> summary(out)
Call: rdrobust

Number of Obs.                  2629
BW type                       Manual
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 2314         315
Eff. Number of Obs.             608         280
Order est. (p)                    1           1
Order bias  (p)                   2           2
BW est. (h)                  20.000      20.000
BW bias (b)                  20.000      20.000
rho (h/b)                     1.000       1.000

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     2.937     1.343     2.187     0.029     [0.305 , 5.569]
        Robust         -         -     1.379     0.168    [-1.117 , 6.414]
================================================================================
```

Stata Snippet 13

```
. rdrobust Y X, kernel(triangular) p(1) h(20)
```

Once again, this produces the same coefficient of 2.937 that we found when we used the weighted least-squares command with triangular weights. We postpone the discussion of standard errors, confidence intervals, and the distinction between the `Conventional` versus `Robust` results until we discuss inference methods.

Finally, if we wanted to reduce the approximation error in the estimation of the RD effect, we could increase the order of the polynomial and use a local quadratic fit instead of a local linear one. This can be implemented in `rdrobust` setting `p=2`.

R Snippet 14

```
> out = rdrobust(Y, X, kernel = "triangular", p = 2, h = 20)
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                      Manual
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2314         315
Eff. Number of Obs.            608         280
Order est. (p)                   2           2
Order bias  (p)                  3           3
BW est. (h)                 20.000      20.000
BW bias (b)                 20.000      20.000
rho (h/b)                    1.000       1.000

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     2.649     1.921     1.379     0.168    [-1.117 , 6.414]
        Robust         -         -     0.420     0.674    [-3.969 , 6.135]
================================================================================
```

Stata Snippet 14

```
. rdrobust Y X, kernel(triangular) p(2) h(20)
```

Note that the estimated effect changes from 2.937 with $p = 1$, to 2.649 with $p = 2$. It is not unusual to observe a change in the point estimate as one changes the polynomial order used in the estimation. Unless the higher-order terms in the approximation are exactly zero, incorporating those terms in the estimation will reduce the approximation error and thus lead to changes in the estimated effect. The relevant practical question is whether such changes in the point estimator change the conclusions of the study. For that, we need to consider inference as well as estimation procedures, a topic we discuss in the upcoming sections.

In general, choosing an ad hoc bandwidth (as done in the previous commands) is not advisable. It is unclear what the value $h = 20$ means in terms of bias and variance properties, or whether this is the best approach for estimation and inference. The command `rdbwselect`, which is part of the `rdrobust` package, implements optimal, data-driven bandwidth selection methods. We illustrate the use of `rdbwselect` by selecting an MSE-optimal bandwidth for

the local linear estimator of τ_{SRD}.

R Snippet 15

```
> out = rdbwselect(Y, X, kernel = "triangular", p = 1, bwselect = "mserd")
> summary(out)
Call: rdbwselect

Number of Obs.                  2629
BW type                        mserd
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 2314         315
Order est. (p)                    1           1
Order bias  (q)                   2           2

=============================================================
                   BW est. (h)    BW bias (b)
            Left of c Right of c  Left of c Right of c
=============================================================
     mserd    17.239     17.239     28.575     28.575
=============================================================
```

Stata Snippet 15

```
. rdbwselect Y X, kernel(triangular) p(1) bwselect(mserd)
```

The MSE-optimal bandwidth choice depends on the choice of polynomial order and kernel function, which is why both have to be specified in the call to `rdbwselect`. The first output line indicates the type of bandwidth selector; in this case, it is MSE-optimal (`mserd`). The type of kernel used is also reported, as is the total number of observations. The middle rows report the number of observations on each side of the cutoff, and the order of polynomial chosen for estimation of the RD effect, the `Order est. (p)` row.

In the bottom rows, we see the estimated optimal bandwidth choices. The bandwidth `h` refers to the bandwidth used to estimate the RD effect τ_{SRD}; we sometimes refer to it as the *main bandwidth*. The bandwidth `b` is an additional bandwidth used to estimate a bias term that is needed for robust inference; we omit discussion of `b` until the following sections.

The estimated MSE-optimal bandwidth for the local-linear RD point estimator with triangular kernel weights is 17.239. The option `bwselect = "mserd"` imposes the same bandwidth `h` on each side of the cutoff, that is, uses the neighborhood $[c-h, c+h]$. This is why the columns `Left of c` and `Right of c` have the same value 17.239. If instead we wish to allow

the bandwidth to be different on each side of the cutoff, we can choose two MSE-optimal bandwidths by using the `bwselect = "msetwo"` option. This leads to a bandwidth of 19.967 on the control side, and a bandwidth of 17.359 on the treated side, as shown below.

R Snippet 16

```
> out = rdbwselect(Y, X, kernel = "triangular", p = 1, bwselect = "msetwo")
> summary(out)
Call: rdbwselect

Number of Obs.                 2629
BW type                      msetwo
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2314         315
Order est. (p)                   1           1
Order bias  (q)                  2           2

=====================================================
                  BW est. (h)    BW bias (b)
            Left of c Right of c  Left of c Right of c
=====================================================
    msetwo    19.967     17.359     32.278     29.728
=====================================================
```

Stata Snippet 16

```
. rdbwselect Y X, kernel(triangular) p(1) bwselect(msetwo)
```

Once we select the MSE-optimal bandwidth(s), we can pass them to the function `rdrobust` using the option `h`. But it is much easier to use the option `bwselect` in `rdrobust`. When we use this option, `rdrobust` calls `rdbwselect` internally, selects the bandwidth as requested, and then uses the optimally chosen bandwidth to estimate the RD effect.

We now use the `rdrobust` command to perform bandwidth selection and point estimation in one step, using $p = 1$ and triangular kernel weights.

R Snippet 17

```
> out = rdrobust(Y, X, kernel = "triangular", p = 1, bwselect = "mserd")
> summary(out)
Call: rdrobust

Number of Obs.                  2629
BW type                        mserd
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 2314         315
Eff. Number of Obs.             529         266
Order est. (p)                    1           1
Order bias  (p)                   2           2
BW est. (h)                  17.239      17.239
BW bias (b)                  28.575      28.575
rho (h/b)                     0.603       0.603

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     3.020     1.427     2.116     0.034     [0.223 , 5.817]
        Robust         -         -     1.776     0.076    [-0.309 , 6.276]
================================================================================
```

Stata Snippet 17

```
. rdrobust Y X, kernel(triangular) p(1) bwselect(mserd)
```

As we can see, when the same MSE-optimal bandwidth is used on both sides of the cutoff, the effect of a bare Islamic victory on the educational attainment of women is 3.020, slightly larger than the 2.937 effect that we found above when we used the ad hoc bandwidth of 20.

We can also explore the `rdrobust` output to obtain the estimates of the average outcome at the cutoff separately for treated and control observations.

R Snippet 18

```
> rdout = rdrobust(Y, X, kernel = "triangular", p = 1, bwselect = "mserd")
> print(names(rdout)[1:7])
[1] "Estimate" "bws"      "coef"     "se"       "z"        "pv"       "ci"
> print(names(rdout)[8:15])
[1] "beta_p_l" "beta_p_r" "V_cl_l"   "V_cl_r"   "V_rb_l"   "V_rb_r"   "N"        "Nh"
> print(names(rdout)[16:23])
[1] "Nb"     "tau_cl" "tau_bc" "c"      "p"      "q"      "bias"   "kernel"
> print(names(rdout)[24:27])
[1] "all"      "vce"      "bwselect" "level"
> print(rdout$beta_p_r)
           [,1]
[1,] 15.6649438
[2,] -0.1460846
> print(rdout$beta_p_l)
           [,1]
[1,] 12.6454218
[2,] -0.2477231
```

Stata Snippet 18

```
. rdrobust Y X
. ereturn list
```

We see that the RD effect of 3.020 percentage points in the female high school attainment percentage is the difference between a percentage of 15.6649438% in municipalities where the Islamic party barely wins and a percentage of 12.6454218% in municipalities where the Islamic party barely loses, that is, $15.6649438 - 12.6454218 \approx 3.020$. By accessing the control mean at the cutoff in this way, we learn that the RD effect represents an increase of $(3.020/12.6454218) \times 100 = 23.88\%$ relative to the control mean.

This effect, together with the means at either side of the cutoff, can be easily illustrated with `rdplot`, using the options `h`, `p`, and `kernel`, to set exactly the same specification used in `rdrobust` and produce an exact illustration of the RD effect. We illustrate the commands below, and show the resulting plot in Figure 15.

R Snippet 19

```
> bandwidth = rdrobust(Y, X, kernel = "triangular", p = 1, bwselect = "mserd")$bws[1,
+ 1]
> out = rdplot(Y[abs(X) <= bandwidth], X[abs(X) <= bandwidth],
+ p = 1, kernel = "triangular", cex.axis = 1.5, cex.lab = 1.5)
> summary(out)
Call: rdplot

Number of Obs.                  795
Kernel                   Triangular

Number of Obs.                  529            266
Eff. Number of Obs.             528            265
Order poly. fit (p)               1              1
BW poly. fit (h)             17.225         17.048
Number of bins scale              1              1

Bins Selected                    19             17
Average Bin Length            0.907          1.003
Median Bin Length             0.907          1.003

IMSE-optimal bins                 5              3
Mimicking Variance bins          19             17

Relative to IMSE-optimal:
Implied scale                 3.800          5.667
WIMSE variance weight         0.018          0.005
WIMSE bias weight             0.982          0.995
```

Stata Snippet 19

```
. rdrobust Y X, p(1) kernel(triangular) bwselect(mserd)
. local bandwidth = e(h_l)
. rdplot Y X if abs(X) <= `bandwidth', p(1) h(`bandwidth') kernel(triangular)
```

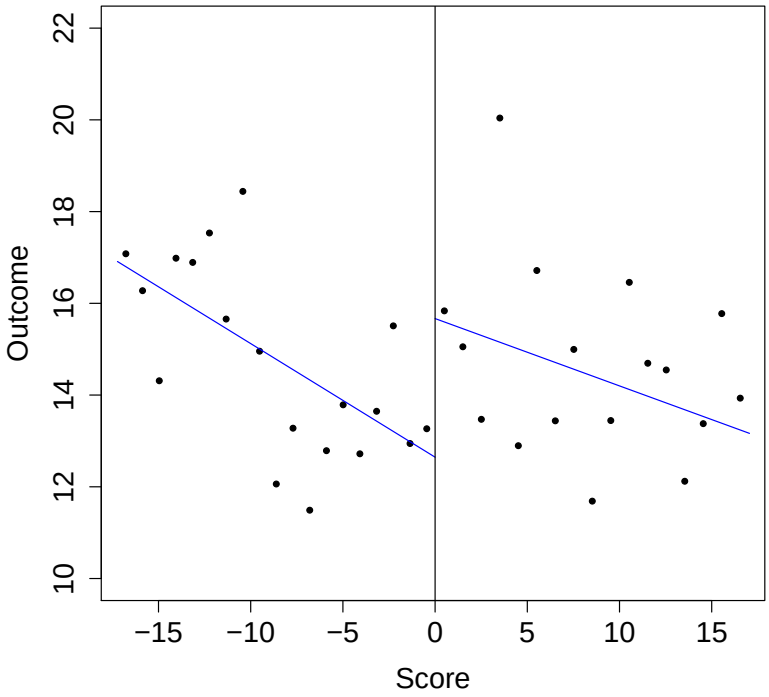


Figure 15: Local Polynomial RD Effect Illustrated with `rdplot` (Meyersson Data)

Finally, we note that by default, all MSE-optimal bandwidth selectors in `rdrobust` include the regularization term that we discussed in subsection 4.2.2. We can exclude the regularization term with the option `scaleregul=0` in the `rdrobust` (or `rdbwselect`) call.

R Snippet 20

```
> out = rdrobust(Y, X, kernel = "triangular", scaleregul = 0, p = 1,
+ bwselect = "mserd")
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 2314         315
Eff. Number of Obs.            1152         305
Order est. (p)                    1           1
Order bias  (p)                   2           2
BW est. (h)                  34.983      34.983
BW bias (b)                  46.233      46.233
rho (h/b)                     0.757       0.757

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     2.843     1.110     2.562     0.010     [0.668 , 5.018]
        Robust         -         -     2.384     0.017     [0.596 , 6.104]
================================================================================
```

Stata Snippet 20

```
. rdrobust Y X, kernel(triangular) p(1) bwselect(mserd) scaleregul(0)
```

In this application, excluding the regularization term has a very large impact on the estimated $h_{\texttt{MSE}}$. With regularization, $\hat{h}_{\texttt{MSE}}$ is 17.239, while excluding regularization increases it to 34.983, an increase of roughly 100%. Nevertheless, the point estimate remains relatively stable, moving from 3.020 with regularization to 2.843 without regularization.

4.3 Local Polynomial Inference

In addition to providing a local polynomial point estimator of the RD treatment effect, we are interested in testing hypotheses and constructing confidence intervals. Although, at first glance, it seems that we could employ ordinary least-squares (OLS) inference methods, these methods would treat the local polynomial regression model as correctly specified (i.e.,

parametric), and de facto disregard its fundamental approximation (i.e., non-parametric) nature. Thus, it would be intellectually and methodologically incoherent to simultaneously select a bandwidth according to a bias-variance trade-off and then proceed as if the bias were zero, that is, as if the local polynomial fit were exact and no misspecification error existed.

These considerations imply that valid inference should take into account the effect of misspecification. In particular, the MSE-optimal bandwidths discussed previously ($h_{\texttt{MSE}}$, $h_{\texttt{MSE},-}$, and $h_{\texttt{MSE},+}$) result in an RD point estimator that is both consistent and optimal in an MSE sense, but inferences based on these bandwidth choices pose a problem. The challenge is that these bandwidths are not "small" enough to remove the leading bias term in the standard distributional approximations used to conduct statistical inference. The root of the problem is that these bandwidth choices are developed for point estimation purposes, and as such they pay no attention to their effects in terms of distributional properties of typical t-tests or related statistics. Thus, constructing confidence intervals using standard OLS large-sample results using the data with $X_i \in [c - h_{\texttt{MSE}}, c + h_{\texttt{MSE}}]$ will generally result in invalid inferences.

There are two general approaches that can be used to address this key problem. One approach is to use the bandwidth $h_{\texttt{MSE}}$ for both estimation and inference, but modify the usual t-statistic to account for the effects of misspecification due to the large bandwidth, as well as for the additional sampling error introduced by such modification. The other is to use $h_{\texttt{MSE}}$ only for point estimation, and then choose a different bandwidth for inference purposes. We elaborate on these issues next.

4.3.1 Using the MSE-Optimal Bandwidth for Inference

We first discuss how to make valid inferences when the bandwidth choice is $h_{\texttt{MSE}}$ (or some data-driven implementation thereof). The local polynomial RD point estimator $\hat{\tau}_{\texttt{SRD}}$ has an approximate large-sample distribution

$$\frac{\hat{\tau}_{\texttt{SRD}} - \tau_{\texttt{SRD}} - \mathscr{B}}{\sqrt{\mathscr{V}}} \overset{a}{\sim} \mathcal{N}(0,1)$$

where $\mathscr{B}$ and $\mathscr{V}$ are, respectively, the asymptotic bias and variance of the RD local polynomial estimator of order p, discussed previously in the context of MSE expansions and bandwidth selection. This distributional result is similar to those encountered, for example, in standard linear regression problems—with the important distinction that now the bias term $\mathscr{B}$ features explicitly; this term highlights the trade-off between bandwidth choice and misspecification bias locally to the cutoff. The variance term $\mathscr{V}$ can be calculated as

in (weighted) least-squares problems, for instance accounting for heteroskedasticity and/or clustered data. We do not provide the exact formulas for variance estimation, to save space and notation, but they can be found in the references given at the end of this section and are all implemented in `rdrobust`.

Given the distributional approximation for the RD local polynomial estimator, an asymptotic 95% confidence interval for $\tau_{\texttt{SRD}}$ is approximately given by

$$\texttt{CI} = \left[(\hat{\tau}_{\texttt{SRD}} - \mathscr{B}) \pm 1.96 \cdot \sqrt{\mathscr{V}}\right].$$

This confidence interval depends on the unknown bias or misspecification error $\mathscr{B}$, and any practical procedure that ignores it will lead to incorrect inferences unless this term is negligible (i.e., unless the local linear regression model is close-to-correctly specified). The bias term arises because the local polynomial approach is a non-parametric approximation: instead of *assuming* that the underlying regression functions are pth order polynomials (as would occur in OLS estimation), this approach uses the polynomial to *approximate* the unknown regression functions.

We now discuss different strategies that are often employed to make inferences for $\tau_{\texttt{SRD}}$ based on asymptotic distributional approximations in the presence of non-parametric misspecification biases, and explain why some of them are invalid. Our recommendation is to use a robust bias correction approach, which is theoretically valid, enjoys some optimality properties, and performs well in practice.

Conventional Inference and Undersmoothing

A strategy sometimes found in RD empirical work is to ignore the misspecification error even when an MSE-optimal bandwidth is used. This is not only invalid but also methodologically incoherent: an MSE-optimal bandwidth cannot be selected in the absence of misspecification error (zero bias), and statistical inference based on standard OLS methods (ignoring the bias) cannot be valid when an MSE-optimal bandwidth is employed.

This naïve approach to statistical inference treats the local polynomial approach as parametric within the neighborhood around the cutoff and de facto ignores the bias term, a procedure that leads to invalid inferences in all cases except when the approximation error is so small that it can be ignored. When the bias term is zero, the approximate distribution

of the RD estimator is $\hat{\tau}_{\texttt{SRD}} - \tau_{\texttt{SRD}}/\sqrt{\mathscr{V}} \overset{a}{\sim} \mathcal{N}(0,1)$ and the confidence interval is

$$\texttt{CI}_{\texttt{us}} = \left[\hat{\tau}_{\texttt{SRD}} \pm 1.96 \cdot \sqrt{\mathscr{V}} \right].$$

Since this is the same confidence interval that follows from parametric least-squares estimation, we refer to it as *conventional*. Using the conventional confidence interval $\texttt{CI}_{\texttt{us}}$ implicitly assumes that the chosen polynomial gives an exact approximation to the true functions $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$. Since these functions are unknown, this assumption is not verifiable and will rarely be credible. If researchers use $\texttt{CI}_{\texttt{us}}$ when in fact the approximation error is non-negligible, all inferences will be incorrect, leading to under-coverage of the true treatment effect or, equivalently, over-rejection of the null hypothesis of zero treatment effect. For this reason, we strongly discourage researchers from using conventional inference when using local polynomial methods, unless the misspecification bias can credibly be assumed small (ruling out, in particular, the use of MSE-optimal bandwidth choices).

A theoretically sound but ad hoc alternative procedure is to use these conventional confidence intervals with a smaller or "undersmoothed" bandwidth relative to the MSE-optimal one used for construction of the point estimator $\hat{\tau}_{\texttt{SRD}}$. Practically, this procedure involves first selecting the MSE-optimal bandwidth, then selecting a bandwidth smaller than the MSE-optimal choice, and finally constructing the conventional confidence intervals $\texttt{CI}_{\texttt{us}}$ with this smaller bandwidth—note that the latter step requires estimating both a new point estimator and a new standard error with the smaller bandwidth. The theoretical justification is that, for bandwidths smaller than the MSE-optimal choice, the bias term will become negligible in the large-sample distributional approximation. (This is why we use the subscript "`us`" to refer to the conventional confidence interval.)

The main drawback of this undersmoothing 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 choose to divide by three, etc. 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 fewer observations used for estimation and inference. Finally, from a substantive perspective, some researchers prefer to avoid using different observations for estimation and inference, which is required by the undersmoothing approach.

Standard Bias Correction

As an alternative to undersmoothing (i.e., to choosing a bandwidth smaller than the MSE-optimal bandwidth), inference could be based on the MSE-optimal bandwidth so long as the induced misspecification error is manually estimated and removed from the distributional approximation. This approach, known as *bias correction*, first estimates the bias term $\mathscr{B}$ with the estimator $\hat{\mathscr{B}}$ (which in fact is already estimated for implementation of MSE-optimal bandwidth selection), and then constructs confidence intervals that are centered at the bias-corrected point estimate:

$$\texttt{CI}_{\texttt{bc}} = \left[\ \left(\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}\right) \pm 1.96 \cdot \sqrt{\mathscr{V}} \ \right].$$

As explained above, the bias term depends on the "curvature" of the unknown regression functions captured via their derivative of order $p+1$ at the cutoff. These unknown derivatives can be estimated with a local polynomial of order $q = p + 1$ or higher, which requires another choice of bandwidth, denoted b. Therefore, the RD point estimate $\hat{\tau}_{\texttt{SRD}}$ employs the bandwidth h, while the bias estimate $\hat{\mathscr{B}}$ employs the additional bandwidth b. The ratio $\rho = h/b$ is important, as it relates to the variability of the bias correction estimate relative to the RD point estimator. Standard bias correction methods require $\rho = h/b \to 0$, that is, a small ρ. In particular, note this rules out $\rho = h/b = 1$, that is, standard bias correction does not allow $h = b$.

The bias-corrected confidence intervals $\texttt{CI}_{\texttt{bc}}$ allow for a wider range of bandwidths h and, in particular, result in valid inferences when the MSE-optimal bandwidth is used. However, they typically have poor performance in applications because the variability introduced in the bias estimation step is not incorporated in the variance term used. Despite employing the additional estimated term $\hat{\mathscr{B}}$, $\texttt{CI}_{\texttt{bc}}$ employs the same variance as $\texttt{CI}_{\texttt{us}}$, essentially ignoring the variability that is introduced when $\mathscr{B}$ is estimated. This results in a poor distributional approximation and hence considerable coverage distortions in practice.

Robust Bias Correction

A superior strategy that is both theoretically sound and leads to improved coverage in finite samples is to use *robust* bias correction for constructing confidence intervals. This approach leads to demonstrably superior inference procedures, with smaller coverage error and shorter average length than those associated with either $\texttt{CI}_{\texttt{us}}$ or $\texttt{CI}_{\texttt{bc}}$. Furthermore, the robust bias correction approach delivers valid inferences even when the MSE-optimal bandwidth for

point estimation is used—no undersmoothing is necessary—and remains valid even when $\rho = h/b = 1$ ($h = b$), which implies that exactly the same data can be used for both point estimation and inference.

Robust bias-corrected confidence intervals are based on the bias correction procedure described above, by which the estimated bias term $\hat{\mathscr{B}}$ is removed from the RD point estimator. However, in contrast to $\texttt{CI}_{\texttt{bc}}$, the derivation 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 a new asymptotic variance $\mathscr{V}_{\texttt{bc}}$ that, unlike the variance $\mathscr{V}$ used in $\texttt{CI}_{\texttt{us}}$ and $\texttt{CI}_{\texttt{bc}}$, incorporates the contribution of the bias correction step to the variability of the bias-corrected point estimator. Because the new variance $\mathscr{V}_{\texttt{bc}}$ incorporates the extra variability introduced in the bias estimation step, it is larger than the conventional OLS variance $\mathscr{V}$ when the same bandwidth is used.

This approach leads to the robust bias-corrected confidence interval:

$$\texttt{CI}_{\texttt{rbc}} = \left[\left(\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}\right) \pm 1.96 \cdot \sqrt{\mathscr{V}_{\texttt{bc}}} \right],$$

which is constructed by subtracting the bias estimate from the local polynomial estimator and using the new variance formula for Studentization. Note that, like $\texttt{CI}_{\texttt{bc}}$, $\texttt{CI}_{\texttt{rbc}}$ is centered around the bias-corrected point estimate, $\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}$, not around the uncorrected estimate $\hat{\tau}_{\texttt{SRD}}$. This robust confidence interval results in valid inferences when the MSE-optimal bandwidth is used, because it has smaller coverage errors and is therefore less sensitive to tuning parameter choices. In practice, the confidence interval can be implemented by setting $\rho = h/b = 1$ ($h = b$) and choosing $h = h_{\texttt{MSE}}$, or by selecting both h and b to be MSE-optimal for the corresponding estimators, in which case ρ is set to $h_{\texttt{MSE}}/b_{\texttt{MSE}}$ or their respective data-driven implementations.

We summarize the differences between the three types of confidence intervals discussed in Table 3. The conventional OLS confidence interval $\texttt{CI}_{\texttt{us}}$ ignores the bias term and is thus centered at the local polynomial point estimator $\hat{\tau}_{\texttt{SRD}}$, and uses the conventional standard error $\sqrt{\hat{\mathscr{V}}}$. The bias-corrected confidence interval $\texttt{CI}_{\texttt{bc}}$ removes the bias estimate from the conventional point estimator, and is therefore centered at $\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}$; this confidence interval, however, ignores the variability introduced in the bias correction step and thus continues to use the standard error $\sqrt{\hat{\mathscr{V}}}$, which is the same standard error used by $\texttt{CI}_{\texttt{us}}$. The robust bias-corrected confidence interval $\texttt{CI}_{\texttt{rbc}}$ is also centered at the bias-corrected point estimator $\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}$ but, in contrast to $\texttt{CI}_{\texttt{bc}}$, it employs a different standard error, $\sqrt{\hat{\mathscr{V}}_{\texttt{bc}}}$, which is larger than the conventional standard error $\sqrt{\hat{\mathscr{V}}}$ when the same bandwidth h is used. Thus, relative

to the conventional confidence interval, the robust bias-corrected confidence interval is both recentered and rescaled. As discussed above, when $h = h_{\texttt{MSE}}$, the conventional confidence interval $\texttt{CI}_{\texttt{us}}$ is invalid.

Table 3: Local Polynomial Confidence Intervals

	Centered at	Standard Error
Conventional: $\texttt{CI}_{\texttt{us}}$	$\hat{\tau}_{\texttt{SRD}}$	$\sqrt{\hat{\mathscr{V}}}$
Bias-Corrected: $\texttt{CI}_{\texttt{bc}}$	$\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}$	$\sqrt{\hat{\mathscr{V}}}$
Robust bias-corrected: $\texttt{CI}_{\texttt{rbc}}$	$\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}$	$\sqrt{\hat{\mathscr{V}}_{\texttt{bc}}}$

From a practical perspective, the most important feature of the robust bias-corrected confidence interval $\texttt{CI}_{\texttt{rbc}}$ is that it can be used with the MSE-optimal point estimator $\hat{\tau}_{\texttt{SRD}}$ when this estimator is constructed using the MSE-optimal bandwidth choice $h_{\texttt{MSE}}$. In other words, using the robust bias-corrected confidence interval allows researchers to use the same observations with score $X_i \in [c - h_{\texttt{MSE}}, c + h_{\texttt{MSE}}]$ for both optimal point estimation and valid statistical inference.

4.3.2 Using Different Bandwidths for Point Estimation and Inference

Conceptually, the invalidity of the conventional confidence interval $\texttt{CI}_{\texttt{us}}$ based on the MSE-optimal bandwidth $h_{\texttt{MSE}}$ stems from using for inference a bandwidth that is optimally chosen for point estimation purposes. Using $h_{\texttt{MSE}}$ for estimation of the RD effect $\tau_{\texttt{SRD}}$ results in a point estimator $\hat{\tau}_{\texttt{SRD}}$ that is not only consistent but also has minimal asymptotic MSE. Thus, from a point estimation perspective, $h_{\texttt{MSE}}$ leads to highly desirable properties. In contrast, serious methodological challenges arise when researchers attempt to use $h_{\texttt{MSE}}$ for building confidence intervals and making inferences in the standard parametric way, because the MSE-optimal bandwidth choice is not designed with the goal of ensuring good (or even valid) distributional approximations. As shown above, robust bias correction restores a valid standard normal distributional approximation when $h_{\texttt{MSE}}$ is used by recentering and rescaling the usual t-statistic, allowing researchers to use the same bandwidth $h_{\texttt{MSE}}$ for both point estimation and inference.

While employing the MSE-optimal bandwidth for both optimal point estimation and valid statistical inference is very useful in practice, it may be important to also consider statistical inference that is optimal. A natural optimality criterion associated with robustness properties of confidence intervals is the minimization of their coverage error, that is, the discrepancy

between the empirical coverage of the confidence interval and its nominal level. For example, if a 95% confidence interval contains the true parameter 80% of the time, the coverage error is 15 percentage points. Minimization of coverage error for confidence intervals is an idea analogous to minimization of MSE for point estimators.

Thus, an alternative approach to RD inference is to decouple the goal of point estimation from the goal of inference, using a different bandwidth for each case. In particular, this strategy involves estimating the RD effect with $h_{\texttt{MSE}}$, and constructing confidence intervals using a different bandwidth, where the latter is specifically chosen to minimize an approximation to the coverage error (CER) of the confidence interval $\texttt{CI}_{\texttt{rbc}}$, leading to the choice $h = h_{\texttt{CER}}$. Just like $h_{\texttt{MSE}}$ minimizes the asymptotic MSE of the point estimator $\hat{\tau}_{\texttt{SRD}}$, the CER-optimal bandwidth $h_{\texttt{CER}}$ minimizes the asymptotic coverage error rate of the robust bias-corrected confidence interval for $\tau_{\texttt{SRD}}$. This bandwidth cannot be obtained in closed form, but it can be shown to have a faster rate of decay than $h_{\texttt{MSE}}$, which implies that for all practically relevant sample sizes $h_{\texttt{CER}} < h_{\texttt{MSE}}$. By design, constructing $\texttt{CI}_{\texttt{rbc}}$ using the CER-optimal bandwidth choice $h_{\texttt{CER}}$ leads to confidence intervals that are not only valid but also have the fastest rate of coverage error decay.

Note that using $h_{\texttt{CER}}$ for point estimation will result in an RD point estimator that has too much variability relative to its bias and is therefore not MSE-optimal (but is nonetheless consistent). Thus, we recommend that practitioners continue to use $h_{\texttt{MSE}}$ for point estimation of $\tau_{\texttt{SRD}}$, and use either $h_{\texttt{MSE}}$ or $h_{\texttt{CER}}$ to build the robust bias-corrected confidence interval $\texttt{CI}_{\texttt{rbc}}$ for inference purposes, where $\texttt{CI}_{\texttt{rbc}}$ will be either valid (if $h_{\texttt{MSE}}$ is used) or valid and CER-optimal (if $h_{\texttt{CER}}$ is used).

4.3.3 RD Local Polynomial Inference in Practice

We can now discuss the full output of our previous call to `rdrobust` with $p = 1$ and triangular kernel, which we reproduce below.

R Snippet 21

```
> out = rdrobust(Y, X, kernel = "triangular", p = 1, bwselect = "mserd")
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2314         315
Eff. Number of Obs.            529         266
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                 17.239      17.239
BW bias (b)                 28.575      28.575
rho (h/b)                    0.603       0.603

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     3.020     1.427     2.116     0.034     [0.223 , 5.817]
        Robust         -         -     1.776     0.076     [-0.309 , 6.276]
================================================================================
```

Stata Snippet 21

```
. rdrobust Y X, kernel(triangular) p(1) bwselect(mserd)
```

As reported before, the local linear RD effect estimate is 3.020, estimated within the MSE-optimal bandwidth of 17.239. The last output provides all the necessary information to make inferences. The row labeled `Conventional` reports, in addition to the point estimator $\hat{\tau}_{\text{SRD}}$, the conventional standard error $\sqrt{\hat{\mathscr{V}}}$, the standardized test statistic $(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}})/\sqrt{\hat{\mathscr{V}}}$, the corresponding p-value, and the 95% conventional confidence interval $\texttt{CI}_{\text{us}}$. This confidence interval ranges from 0.223 to 5.817 percentage points, suggesting a positive effect of an Islamic victory on the educational attainment of women. Note that $\texttt{CI}_{\text{us}}$ is centered around the conventional point estimator $\hat{\tau}_{\text{SRD}}$:

$$3.020 + 1.427 \times 1.96 = 5.81692 \approx 5.817$$
$$3.020 - 1.427 \times 1.96 = 0.22308 \approx 0.223.$$

The row labeled `Robust` reports the robust bias-corrected confidence interval $\texttt{CI}_{\texttt{rbc}}$. In contrast to $\texttt{CI}_{\texttt{us}}$, $\texttt{CI}_{\texttt{rbc}}$ is centered around the point estimator $\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}}$ (which is by default not reported), and scaled by the robust standard error $\sqrt{\hat{\mathscr{V}}_{\texttt{bc}}}$ (not reported either). $\texttt{CI}_{\texttt{rbc}}$ ranges from -0.309 to 6.276; in contrast to the conventional confidence interval, it does include zero. As expected, $\texttt{CI}_{\texttt{rbc}}$ is not centered at $\hat{\tau}_{\texttt{SRD}}$.

For a fixed common bandwidth, the length of $\texttt{CI}_{\texttt{rbc}}$ is always greater than the length of $\texttt{CI}_{\texttt{us}}$ because $\sqrt{\hat{\mathscr{V}}_{\texttt{bc}}} > \sqrt{\hat{\mathscr{V}}}$. We can see this in our example:

$$\text{Length of } \texttt{CI}_{\texttt{us}} = 5.817 - 0.223 = 5.594$$
$$\text{Length of } \texttt{CI}_{\texttt{rbc}} = 6.276 - (-0.309) = 6.585.$$

However, this will not necessarily be true if different bandwidths are used to construct each confidence interval.

The omission of the bias-corrected point estimator that is at the center of $\texttt{CI}_{\texttt{rbc}}$ from the `rdrobust` output is intentional: when the MSE-optimal bandwidth for $\hat{\tau}_{\texttt{SRD}}$ is used, the bias-corrected estimator is suboptimal in terms of MSE relative to $\hat{\tau}_{\texttt{SRD}}$. (Although the bias-corrected estimator is consistent and valid whenever $\hat{\tau}_{\texttt{SRD}}$ is.) Practically, it is usually desirable to report an MSE-optimal point estimator and then form valid confidence intervals either with the same MSE-optimal bandwidth or with some other optimal choice specifically tailored for inference.

In order to see all the ingredients that go into building the robust confidence interval, we can use the `all` option in `rdrobust`.

R Snippet 22

```
> out = rdrobust(Y, X, kernel = "triangular", p = 1, bwselect = "mserd",
+ all = TRUE)
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2314         315
Eff. Number of Obs.            529         266
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                 17.239      17.239
BW bias (b)                 28.575      28.575
rho (h/b)                    0.603       0.603

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     3.020     1.427     2.116     0.034     [0.223 , 5.817]
Bias-Corrected     2.983     1.427     2.090     0.037     [0.186 , 5.780]
        Robust     2.983     1.680     1.776     0.076    [-0.309 , 6.276]
================================================================================
```

Stata Snippet 22

```
. rdrobust Y X, kernel(triangular) p(1) bwselect(mserd) all
```

The three rows at the bottom of the output are analogous to the the rows in Table 3: the `Conventional` row reports $\texttt{CI}_{\texttt{us}}$, the `Bias-Corrected` row reports $\texttt{CI}_{\texttt{bc}}$, and the `Robust` row reports $\texttt{CI}_{\texttt{rbc}}$. We can see that the standard error used by $\texttt{CI}_{\texttt{us}}$ and $\texttt{CI}_{\texttt{bc}}$ is the same ($\sqrt{\hat{\mathscr{V}}} = 1.427$), while $\texttt{CI}_{\texttt{rbc}}$ uses a different standard error ($\sqrt{\hat{\mathscr{V}}_{\texttt{bc}}} = 1.680$). We also see that the conventional confidence interval is centered at the conventional, non-bias-corrected point estimator 3.020, while both $\texttt{CI}_{\texttt{bc}}$ and $\texttt{CI}_{\texttt{rbc}}$ are centered at the bias-corrected point estimator 2.983. Since we know that $\hat{\tau}_{\texttt{SRD}} = 3.020$ and $\hat{\tau}_{\texttt{SRD}} - \hat{\mathscr{B}} = 2.983$, we can deduce that the bias estimate is $\hat{\mathscr{B}} = 3.020 - 2.983 = 0.037$.

Finally, we investigate the properties of robust bias-corrected inference when employing a CER-optimal bandwidth choice. This is implemented via `rdrobust` with the option `bwselect="cerrd"`.

R Snippet 23

```
> out = rdrobust(Y, X, kernel = "triangular", p = 1, bwselect = "cerrd")
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                       cerrd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2314         315
Eff. Number of Obs.            360         216
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                 11.629      11.629
BW bias (b)                 28.575      28.575
rho (h/b)                    0.407       0.407

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     2.430     1.682     1.444     0.149    [-0.868 , 5.727]
        Robust         -         -     1.324     0.186    [-1.158 , 5.979]
================================================================================
```

Stata Snippet 23

```
. rdrobust Y X, kernel(triangular) p(1) bwselect(cerrd)
```

The common CER-optimal bandwidth for both control and treatment units is $h_{\texttt{CER}} = 11.629$, which is smaller than the MSE-optimal bandwidth calculated previously, $h_{\texttt{MSE}} = 17.239$. The results are qualitatively similar, but now with a larger p-value as the nominal 95% robust bias-corrected confidence interval changes from $[-0.309, 6.276]$ with MSE-optimal bandwidth to $[-1.158, 5.979]$ with CER-optimal bandwidth. The RD point estimator changes from the MSE-optimal value 3.020 to the undersmoothed value 2.43, where the latter RD estimate can be interpreted as having less bias but more variability than the former.

Since both the change in bandwidth choice from MSE-optimal to CER-optimal and the change from one common bandwidth to two different bandwidths are practically important, we conclude this section with a report of all the bandwidth choices. This is obtained using the `all` option in the `rdbwselect` command.

R Snippet 24

```
> out = rdbwselect(Y, X, kernel = "triangular", p = 1, all = TRUE)
> summary(out)
Call: rdbwselect

Number of Obs.                  2629
BW type                          All
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 2314         315
Order est. (p)                    1           1
Order bias  (q)                   2           2

=====================================================
                  BW est. (h)    BW bias (b)
            Left of c Right of c  Left of c Right of c
=====================================================
     mserd    17.239     17.239     28.575     28.575
    msetwo    19.967     17.359     32.278     29.728
    msesum    17.772     17.772     30.153     30.153
  msecomb1    17.239     17.239     28.575     28.575
  msecomb2    17.772     17.359     30.153     29.728
     cerrd    11.629     11.629     28.575     28.575
    certwo    13.468     11.710     32.278     29.728
    cersum    11.988     11.988     30.153     30.153
  cercomb1    11.629     11.629     28.575     28.575
  cercomb2    11.988     11.710     30.153     29.728
=====================================================
```

Stata Snippet 24

```
. rdbwselect Y X, kernel(triangular) p(1) all
```

There are five MSE-optimal bandwidths reported. The row labeled `mserd` reports the bandwidth that minimizes the MSE of the RD point estimator under the constraint that the bandwidth to the left of the cutoff is the same as the bandwidth to the right of it, while the row labeled `msetwo` reports the bandwidth that minimizes the same MSE but allowing the left and right bandwidths to be different. In contrast to the `mserd` and `msetwo` bandwidths, which optimize the MSE of the RD point estimator, $\hat{\tau}_{\texttt{SRD}} = \hat{\mu}_+ - \hat{\mu}_-$, the `msesum` row reports the common bandwidth that minimizes the MSE of the sum of the regression coefficients, not their difference, that is, the MSE of $\hat{\mu}_+ + \hat{\mu}_-$. The fourth and fifth

rows report a combination of the prior MSE-optimal bandwidths: `msecomb1` is the minimum between `mserd` and `msesum`, while `msecomb2` is the median of `msetwo`, `mserd`, and `msesum`. The CER bandwidths reported in the last five rows are analogous to the prior five, with the only difference that the bandwidths reported are optimal with respect to the CER of the confidence interval for $\tau_{\texttt{SRD}}$, not its MSE.

4.4 Extensions

Up to this point, our discussion has considered local polynomials that included only the running variable as a regressor, in a setting where all the observations were assumed to be independent. We now discuss how local polynomial methods can be generalized to accommodate both additional covariates in the model specification, and clustering of observations.

4.4.1 Adding Covariates to the Analysis

The simplest way to implement RD local polynomial analysis is to fit the outcome on the score alone. Although this basic specification is sufficient to analyze most applications, some researchers may want to augment it by including other covariates in addition to the score. Local polynomial methods can easily accommodate additional covariates, but the latter must satisfy an important condition. Unless researchers are willing to invoke parametric assumptions or redefine the parameter of interest, the covariates used to augment the analysis must be balanced at the cutoff. In general, covariate adjustment cannot be used to restore identification of standard RD design treatment effects when treated and control observations differ systematically at the cutoff. When the empirical evidence shows that important predetermined covariates differ systematically at the cutoff, the assumption of continuity of the potential outcomes is implausible, and thus the non-parametric continuity-based RD framework discussed in this Element is no longer appropriate without further (restrictive) assumptions about the data generating process.

We let $\mathbf{Z}_i(1)$ and $\mathbf{Z}_i(0)$ denote two vectors of potential covariates, where $\mathbf{Z}_i(1)$ represents the value taken by the covariates above the cutoff (i.e., under treatment), and $\mathbf{Z}_i(0)$ represents the value taken below the cutoff (i.e., under control). We assume that these covariates are predetermined, that is, that their values are determined prior to, or independently from, the treatment assignment and therefore that the treatment effect on them is zero by construction.

For adjustment, researchers use the observed covariates, $\mathbf{Z}_i$, defined as

$$\mathbf{Z}_i = \begin{cases} \mathbf{Z}_i(0) & \text{if } X_i < c \\ \mathbf{Z}_i(1) & \text{if } X_i \geq c. \end{cases}$$

Predetermined covariates can be included in different ways to augment the basic RD estimation and inference methods. The two most natural approaches are (i) conditioning or subsetting, which makes the most sense when only a few discrete covariates are used, and (ii) partialling out via local polynomial methods. The first approach amounts to employing all the methods we discussed so far, after subsetting the data along the different subclasses generated by the interacted values of the covariates being used. For example, researchers may want to conduct separate analyses for men and women in the sample to study whether the estimated effects and confidence intervals differ between the two subgroups. The implementation of this conditioning approach does not require any modifications to the methods discussed above; they can be applied directly.

The second approach is based on augmenting the local polynomial model to include several additional covariates, which can be discrete or continuous. In this case, the idea is to directly include as many predetermined covariates as possible without affecting the validity of the point estimator, while at the same time improving its efficiency.

Our recommended covariate adjustment strategy is to augment the local polynomial fit by adding the covariates in a linear and additive-separable way. This involves fitting a weighted least-squares regression of the outcome Y_i on (i) a constant, (ii) the treatment indicator T_i, (iii) a p-order polynomial on the running variable, $(X_i - c), (X_i - c)^2, \ldots, (X_i - c)^p$, (iv) a p-order polynomial on the running variable interacted with the treatment, $T_i(X_i - c), T_i(X_i - c)^2, \ldots, T_i(X_i - c)^p$, and (v) the covariates $\mathbf{Z}_i$, using the weights $K((X_i - c)/h)$. This defines the covariate-adjusted RD estimator:

$$\begin{aligned} \tilde{\tau}_{\texttt{SRD}} : \tilde{Y}_i = \tilde{\alpha} + \tilde{\tau}_{\texttt{SRD}} T_i + \tilde{\mu}_{-,1}(X_i - c) + \cdots + \tilde{\mu}_{-,p}(X_i - c)^p \\ + \tilde{\mu}_{+,1} T_i(X_i - c) + \cdots + \tilde{\mu}_{+,p} T_i(X_i - c)^p + \mathbf{Z}_i' \tilde{\boldsymbol{\gamma}}. \end{aligned} \tag{4.3}$$

The estimator $\tilde{\tau}_{\texttt{SRD}}$ captures the average outcome jump at the cutoff in a fully interacted local polynomial regression fit, after partialling out the effect of the covariates $\mathbf{Z}_i$. This approach reduces to the standard RD estimation when no covariates are included.

A very important question is whether the covariate-adjusted estimator $\tilde{\tau}_{\texttt{SRD}}$ estimates the same parameter as the unadjusted estimator $\hat{\tau}_{\texttt{SRD}}$. It can be shown that under mild regularity

conditions, a sufficient condition for $\tilde{\tau}_{\texttt{SRD}}$ to be consistent for the average treatment effect at the cutoff, $\tau_{\texttt{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]$, is that the RD treatment effect on the covariates is zero, that is, that the averages of the covariates under treatment and control at the cutoff are equal to each other, $\mathbb{E}[\mathbf{Z}_i(0)|X_i = c] = \mathbb{E}[\mathbf{Z}_i(1)|X_i = c]$. This condition is analogous to the "covariate balance" requirement in randomized experiments, and will hold naturally when the covariates are truly predetermined.

Thus, when predetermined covariates are included in the estimation as in Equation (4.3), the covariate-adjusted estimator estimates the standard RD treatment effect, $\tau_{\texttt{SRD}}$. This result, however, depends crucially on the particular way in which the covariates are included in (4.3): linearly, additive-separably, and without interacting the covariates with the treatment. If, instead of adding $\mathbf{Z}_i'\tilde{\boldsymbol{\gamma}}$, we interacted the covariates with the treatment and included the terms $(1 - T_i)\mathbf{Z}_i'\check{\boldsymbol{\gamma}}_- + T_i\mathbf{Z}_i'\check{\boldsymbol{\gamma}}_+$, a zero RD treatment effect on the covariates would no longer be sufficient for $\tilde{\tau}_{\texttt{SRD}}$ to be a consistent estimator of $\tau_{\texttt{SRD}}$. We therefore recommend including covariates without interacting them with the treatment indicator, as shown in (4.3).

In sum, if the goal is to estimate the RD treatment effect $\tau_{\texttt{SRD}}$, the covariate adjustment should only include predetermined covariates, as including posttreatment or imbalanced covariates will change the parameter being estimated. It follows that, in general, it is not possible to include imbalanced covariates in the estimation to "fix" an RD design in which predetermined covariates are discontinuous at the cutoff and the required continuity assumptions are called into question. For the inclusion of covariates to "control for" unexpected imbalances, researchers will either need to invoke parametric assumptions on the regression functions to enable extrapolation, or redefine the parameter of interest. Therefore, analogously to the case of randomized experiments, the generally valid justification for including covariates in RD analysis is the potential for efficiency gains, not the promise to fix implausible identification assumptions. In many RD applications, a covariate-adjusted local polynomial estimation strategy will lead to shorter confidence intervals for the RD treatment effect, increasing the precision of statistical inferences.

Practical Implementation of Covariate-Adjusted RD Analysis

Including covariates in a linear-in-parameters way as in Equation (4.3) requires the same type of choices as in the standard, unadjusted case: a polynomial order p, a kernel function $K(\cdot)$, and a bandwidth h. Once again, the bandwidth is a crucial choice, and we recommend using optimal data-driven methods to select it. Since the covariate-adjusted point estimator $\tilde{\tau}_{\texttt{SRD}}$ is a function of the covariates, its MSE will also be a function of the covariates. Thus, the optimal

bandwidth choices for $\tilde{\tau}_{\texttt{SRD}}$ will depend on the covariates and will be in general different from the previously discussed bandwidths $h_{\texttt{MSE}}$ and $h_{\texttt{CER}}$. As a consequence, the principled implementation of covariate-adjusted local polynomial methods requires employing an MSE-optimal bandwidth that accounts for the inclusion of covariates in the bandwidth selection step. Although we omit the technical details here, both the MSE-optimal and the CER-optimal bandwidth choices that account for covariate adjustment have been theoretically derived; and they are both implemented in the `rdrobust` software.

We illustrate the inclusion of covariates with the Meyersson application, using the predetermined covariates introduced in Section 2.1: variables from the 1994 election (`vshr_islam1994`, `partycount`, `lpop1994`), and the geographic indicators (`merkezi`, `merkezp`, `subbuyuk`, `buyuk`). In order to keep the same number of observations as in the analysis without covariates, we exclude the indicator for electing an Islamic party in 1989 (`i89`) because this variable has many missing values.

We start by using `rdbwselect` to choose an MSE-optimal bandwidth with covariates, using the default options: a polynomial of order one, a triangular kernel, and the same bandwidth on each side of the cutoff (`mserd` option). We include covariates using the option `covs`.

R Snippet 25

```
> Z = cbind(data$vshr_islam1994, data$partycount, data$lpop1994,
+ data$merkezi, data$merkezp, data$subbuyuk, data$buyuk)
> colnames(Z) = c("vshr_islam1994", "partycount", "lpop1994", "merkezi",
+ "merkezp", "subbuyuk", "buyuk")
> out = rdbwselect(Y, X, covs = Z, kernel = "triangular", scaleregul = 1,
+ p = 1, bwselect = "mserd")
> summary(out)
Call: rdbwselect

Number of Obs.                  2629
BW type                        mserd
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 2314         315
Order est. (p)                    1           1
Order bias  (q)                   2           2

=============================================================
                   BW est. (h)    BW bias (b)
            Left of c Right of c  Left of c Right of c
=============================================================
     mserd    14.409     14.409     23.731     23.731
=============================================================
```

Stata Snippet 25

```
. global covariates "vshr_islam1994 partycount lpop1994 merkezi merkezp subbuyuk buyuk"
. rdbwselect Y X, covs($covariates) p(1) kernel(triangular) bwselect(mserd) scaleregul(1)
```

The MSE-optimal bandwidth including covariates is 14.409, considerably different from the value of 17.239 that we found before in the absence of covariate adjustment. This illustrates the general principle that covariate adjustment will generally change the values of the optimal bandwidths, which in turn will change the point estimates. (Note, however, that the covariate-adjusted local polynomial RD estimate would be different from the unadjusted estimate even if the same bandwidth were employed, as in finite samples the inclusion of covariates will change the estimated coefficients in the local polynomial fit.)

To perform covariate-adjusted local polynomial estimation and inference, we use the `rdrobust` command using the `covs` option.

R Snippet 26

```
> Z = cbind(data$vshr_islam1994, data$partycount, data$lpop1994,
+ data$merkezi, data$merkezp, data$subbuyuk, data$buyuk)
> colnames(Z) = c("vshr_islam1994", "partycount", "lpop1994", "merkezi",
+ "merkezp", "subbuyuk", "buyuk")
> out = rdrobust(Y, X, covs = Z, kernel = "triangular", scaleregul = 1,
+ p = 1, bwselect = "mserd")
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2314         315
Eff. Number of Obs.            448         241
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                 14.409      14.409
BW bias (b)                 23.731      23.731
rho (h/b)                    0.607       0.607

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     3.108     1.284     2.421     0.015     [0.592 , 5.624]
        Robust         -         -     2.088     0.037     [0.194 , 6.132]
================================================================================
```

Stata Snippet 26

```
. global covariates "vshr_islam1994 partycount lpop1994 merkezi merkezp subbuyuk buyuk"
. rdrobust Y X, covs($covariates) p(1) kernel(triangular) bwselect(mserd) scaleregul(1)
```

The estimated RD effect is now 3.108, similar to the unadjusted estimate of 3.020 that we found before. This similarity is reassuring because, if the included covariates are truly predetermined, the unadjusted estimator and the covariate-adjusted estimator are estimating the same parameter and should result in similar estimates. In terms of inference, with the inclusion of covariates, the 95% robust confidence interval is now $[0.194, 6.132]$. The unadjusted robust confidence interval we estimated in the previous section is $[-0.309, 6.276]$. Thus, including covariates reduces the length of the confidence interval from $6.276-(-0.309) = 6.585$

to $6.132 - 0.194 = 5.938$, a reduction of $(|5.938 - 6.585|/6.585) \times 100 = 9.82\%$. The shorter confidence interval obtained with covariate adjustment (and the slight increase in the point estimate) results in the robust p-value decreasing from 0.076 to 0.037.

This exercise illustrates the main benefit of covariate adjustment in local polynomial RD estimation: when successful, the inclusion of covariates in the analysis decreases the length of the confidence interval while simultaneously leaving the point estimate (roughly) unchanged.

4.4.2 Clustering the Standard Errors

Another issue commonly encountered by practitioners is the clustering of observations in groups, such as individuals inside households, municipalities inside counties, or households inside villages. When the units of analysis are clustered into groups and the researcher suspects that the errors are correlated within (but not across) groups, it may be appropriate to employ variance estimators that are robust to the clustered nature of the data.

Using ideas from least-squares estimation and inference methods, it is possible to adjust the local polynomial variance estimators to account for clustering. Since the CER- and MSE-optimal bandwidth selectors depend on the variance estimators, employing cluster-robust variance estimators changes the optimal bandwidth relative to the case of no clustering. Consequently, in the local polynomial RD setting (and in contrast to the ordinary least-squares setting) employing cluster-robust variance estimators leads not only to different estimated standard errors relative to the unclustered case, but also to different point estimates. In general, cluster-robust variance estimators can be smaller or larger than variance estimators that do not account for clustering. This fact, combined with the associated change in the point estimator that results from the change in the optimal bandwidth when cluster-robust variance estimators are employed, means that cluster-robust standard errors can lead to recentered confidence intervals that can be either shorter or longer in length.

The cluster-robust variance estimation formulas are beyond the scope of this practical guide, but we do illustrate how to employ these estimators in practice using `rdrobust`. We provide further illustration of these methods in the discussion of RD designs with discrete running variables in the accompanying Element (Cattaneo, Idrobo, and Titiunik, forthcoming).

In the Meyersson application, we estimate the effect of Islamic victory on the educational attainment of women, clustering each individual observation (which corresponds to a municipality) by province. In `rdrobust`, we use the option `cluster` to pass the variable that contains the cluster information for every observation.

R Snippet 27

```
> out = rdrobust(Y, X, kernel = "triangular", scaleregul = 1, p = 1,
+ bwselect = "mserd", cluster = data$prov_num)
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2314         315
Eff. Number of Obs.            584         277
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                 19.035      19.035
BW bias (b)                 29.873      29.873
rho (h/b)                    0.637       0.637

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     2.969     1.604     1.851     0.064    [-0.175 , 6.113]
        Robust         -         -     1.635     0.102    [-0.583 , 6.460]
================================================================================
```

Stata Snippet 27

```
. rdrobust Y X, p(1) kernel(triangular) bwselect(mserd) ///
> scaleregul(1) vce(nncluster prov_num)
```

Using a cluster-robust variance estimator leads to a point estimator of 2.969 percentage points, slightly smaller than the point estimator of 3.020 that we obtained without clustering in Section 4.2.4. This change in the point estimate occurs because the MSE-optimal bandwidth is now 19.035, larger than the 17.239 bandwidth estimated in the absence of clustering. In addition, the cluster-robust variance estimator is larger than the unclustered variance estimator—for example, comparing to the prior results in the absence of clustering, the conventional standard error changes from 1.427 to 1.604 when a cluster-robust estimator is used. The decrease in the point estimator, together with the increase in the variance, lead to a wider confidence interval; with a cluster-robust variance estimator, the robust confidence interval is [-0.583,6.460], wider and with center closer to zero than the [-0.309,6.276] robust confidence interval in the absence of clustering.

We can also combine a cluster-robust variance estimator with covariate adjustment in the local polynomial fit, using the `covs` and `cluster` options simultaneously.

R Snippet 28

```
> Z = cbind(data$vshr_islam1994, data$partycount, data$lpop1994,
+ data$merkezi, data$merkezp, data$subbuyuk, data$buyuk)
> colnames(Z) = c("vshr_islam1994", "partycount", "lpop1994", "merkezi",
+ "merkezp", "subbuyuk", "buyuk")
> out = rdrobust(Y, X, covs = Z, kernel = "triangular", scaleregul = 1,
+ p = 1, bwselect = "mserd", cluster = data$prov_num)
> summary(out)
Call: rdrobust

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2314         315
Eff. Number of Obs.            481         254
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                 15.675      15.675
BW bias (b)                 24.663      24.663
rho (h/b)                    0.636       0.636

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     3.146     1.301     2.419     0.016     [0.597 , 5.696]
        Robust         -         -     2.121     0.034     [0.243 , 6.154]
================================================================================
```

Stata Snippet 28

```
. global covariates "vshr_islam1994 partycount lpop1994 merkezi merkezp subbuyuk buyuk"
. rdrobust Y X, covs($covariates) p(1) kernel(triangular) bwselect(mserd) ///
> scaleregul(1) vce(nncluster prov_num)
```

Relative to the unadjusted, unclustered case, adding covariates and employing cluster-robust variance estimators leads to a different optimal bandwidth of 15.675, which changes the point estimate to 3.146. Once again, these changes translate into a different confidence interval, equal to [0.243,6.154]. Relative to the case of cluster-robust variance estimator

without covariate adjustment, adding covariates reduces the length of the confidence interval and shifts it to the right. As a result, the cluster-robust covariate-adjusted 95% confidence interval does not include zero.

4.5 Further Reading

A textbook discussion of non-parametric local polynomial methods can be found in Fan and Gijbels (1996), and their application to RD estimation and inference is discussed by Hahn, Todd, and van der Klaauw (2001). Calonico, Cattaneo, and Titiunik (2015a) and Gelman and Imbens (2019) discuss the role of global polynomial estimation for RD analysis. MSE-optimal bandwidth selection for the local polynomial RD point estimator is developed in Imbens and Kalyanaraman (2012), Calonico, Cattaneo, and Titiunik (2014b), Bartalotti and Brummet (2017), Calonico, Cattaneo, Farrell, and Titiunik (2019), and Arai and Ichimura (2018). Robust bias corrected confidence intervals were proposed by Calonico, Cattaneo, and Titiunik (2014b), and their higher-order properties as well as CER-optimal bandwidth selection were developed by Calonico, Cattaneo, and Farrell (2018, 2019a,b). See also Cattaneo and Vazquez-Bare (2016) for an overview of RD bandwidth selection methods. Bootstrap methods based on robust bias correction are developed in Bartalotti, Calhoun, and He (2017). RD analysis with the inclusion of predetermined covariates and cluster-robust inference is discussed in Calonico, Cattaneo, Farrell, and Titiunik (2019), and other extensions of estimation and inference using robust bias correction are discussed in Xu (2017), Dong (2019), and Dong, Lee, and Gou (2019). Hyytinen, Meriläinen, Saarimaa, Toivanen, and Tukiainen (2018) offer an empirical example assessing the performance of robust bias correction inference methods. Cattaneo, Titiunik, and Vazquez-Bare (2018) discuss power calculations using local polynomial menthods in RD designs. Further related results and references are given in the edited volume by Cattaneo and Escanciano (2017).

5 Validation and Falsification of the RD Design

A main advantage of the RD design is that the mechanism by which treatment is assigned is known and based on observable features, 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, the known rule that assigns treatment based on the score and cutoff is not by itself enough to guarantee that the assumptions needed to recover the RD effect are met.

For example, a scholarship may be assigned based on whether students receive an exam grade above a cutoff, but if the cutoff is known to the students' parents and there are mechanisms to appeal the grade, the RD design may be invalid if systematic differences among students are present due to the appeal process. Formally, the presence of an appeal process might invalidate the assumption that the average potential outcomes are continuous at the cutoff. If the parents who are successful in appealing the grade when their child is barely below the cutoff are systematically different from the parents who choose not to appeal in ways that affect the outcome of interest, then the RD design based on the final grade assigned to each student would be invalid (while the RD design based on the original grade would not). For instance, if the outcome of interest is performance on a future exam and parent involvement is positively correlated with students' future academic achievement, the average potential outcomes of students at or just above the cutoff will be much higher than the average potential outcomes of students just below the cutoff, leading to a discontinuity at the cutoff and thus invalidating the RD design.

If the RD cutoff 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. Thus, the first type of information that should be provided is whether an institutionalized mechanism to appeal the score exists, and if so, how often (and by whom) it is used to successfully change the score. 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. For example, social programs are commonly assigned based on some poverty index; if program officers moved units with index barely below the cutoff to the treatment group in a systematic way (e.g., all households with small children), then the RD design would be invalid whenever the systematic differences between treated and control units near the cutoff were correlated with outcome differences. This type of behavior can typically be identified collecting qualitative information (such as interviews, internal rules and memos, etc.) from the

program administration officers.

In many cases, however, qualitative information will be limited, and researchers will be unable to completely rule out the possibility of units manipulating their score. More importantly, the fact that there are no institutionalized or known mechanisms to appeal and change the score does not imply the absence of informal mechanisms by which this may happen. Thus, an essential step in evaluating the plausibility of the RD assumptions is to provide empirical evidence supporting the validity of the design. Naturally, the continuity assumptions that guarantee the validity of the RD design are about unobservable features and as such are inherently untestable. Nonetheless, the RD design offers an array of empirical methods that, under reasonable assumptions, can provide useful evidence about the plausibility of its assumptions. These so-called validation methods are based on various empirical implications of the unobservable RD assumptions that can be expected to hold in most cases, and can provide indirect evidence about its validity.

We now discuss five empirical validation tests based on (i) the null treatment effect on predetermined covariates or placebo outcomes, (ii) the continuity of the score density around the cutoff, (iii) the treatment effect at artificial cutoff values, (iv) the exclusion of observations near the cutoff, and (v) the sensitivity to bandwidth choices.

5.1 Predetermined Covariates and Placebo Outcomes

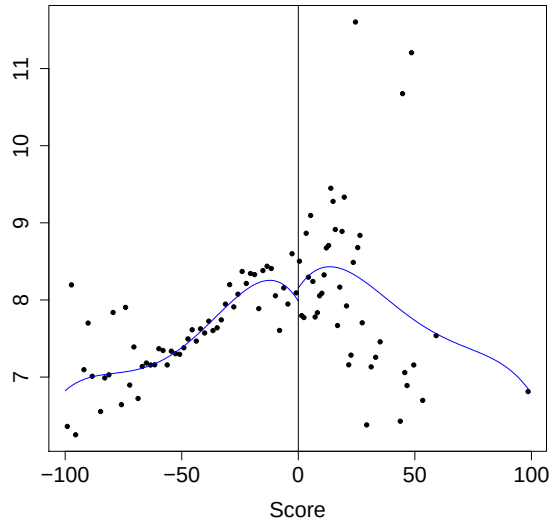
One of the most important RD falsification tests involves examining whether, near the cutoff, treated units are similar to control units in terms of observable characteristics. The idea is simply that, if units lack the ability to precisely manipulate the score value they receive, there should be no 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 variables 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*.

Note that predetermined covariates can be unambiguously defined, but placebo outcomes are always specific to each application. For example, any characteristic that is determined before the moment when treatment is assigned is generally a predetermined covariate. In contrast, whether a variable is a placebo outcome depends on the particular treatment under consideration. For example, if the treatment is access to clean water and the outcome of

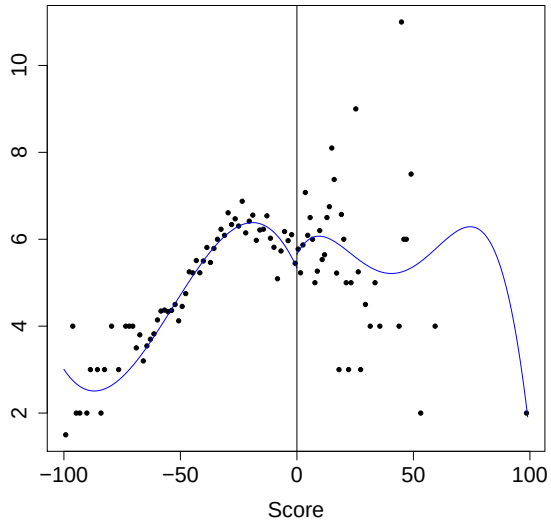
interest is child mortality, a treatment effect is expected on mortality due to water-bone illnesses but not on mortality due to other causes such as car accidents (see Galiani, Gertler, and Schargrodsky, 2005). Thus, mortality from road accidents would be a reasonable placebo outcome in this example. However, child mortality from road accidents would not be an adequate placebo outcome to validate an RD design that studies the effects of a safety program aimed at increasing the use of car seats.

Regardless of whether the analysis is based on predetermined covariates or placebo outcomes, the fundamental principle behind this type of falsification analysis is always the same: 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 or placebo outcome, researchers should first choose an optimal bandwidth, and then use local polynomial techniques within that bandwidth to estimate the "treatment effect" and employ valid inference procedures such as the robust bias-corrected methods discussed previously. The fundamental idea behind this test is that, since the predetermined covariate (or placebo outcome) could not have been affected by the treatment, the null hypothesis of no treatment effect should not be rejected if the RD design is valid. The reasoning is that if covariates or placebo outcomes that are known to correlate strongly with the outcome of interest are discontinuous at the cutoff, the continuity of the potential outcome functions is unlikely to hold, and thus the validity of the design is called into question.

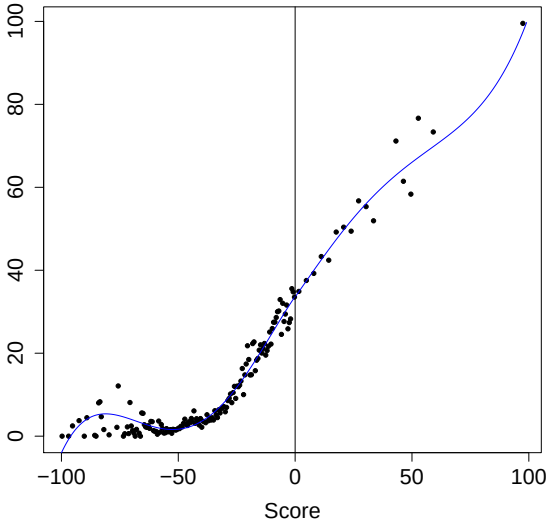
When using the continuity-based approach to RD analysis, this falsification test employs the local polynomial techniques discussed in Section 4 to test whether the predetermined covariates and placebo outcomes are continuous at the cutoff, in other words, to test whether the treatment has an effect on them. We illustrate with the Meyersson application, using the set of predetermined covariates used for covariate adjustment in Section 4.2.4. We start by presenting a graphical analysis, creating an RD plot for every covariate using `rdplot` with the default options (mimicking variance, evenly-spaced bins). The plots are presented in Figure 16. The specific commands are omitted to conserve space, but they are included in the online replication materials.



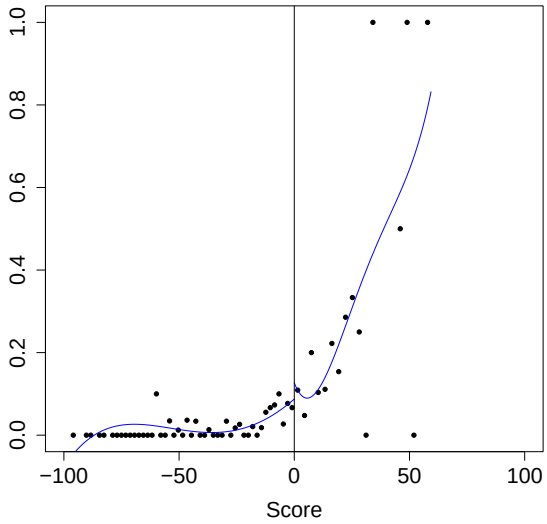
(a) Log Population in 1994



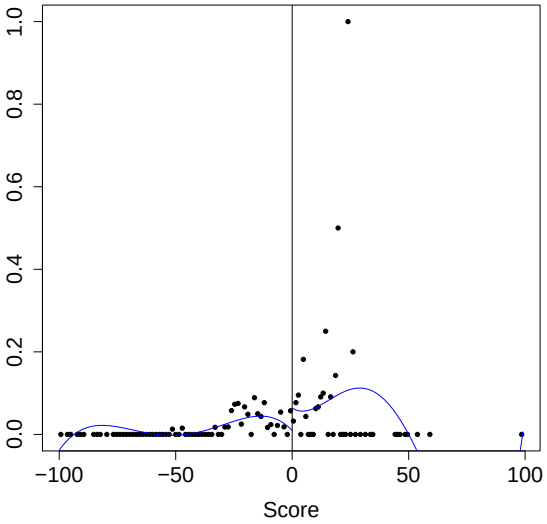
(b) Number of Parties Receiving Votes in 1994



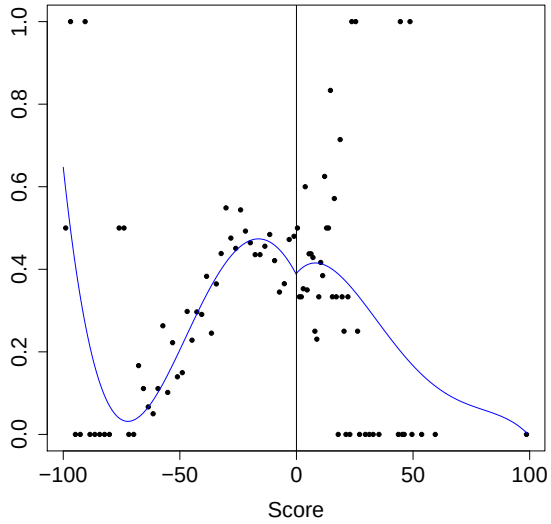
(c) Islamic Vote Percentage in 1994



(d) Islamic Mayor in 1989



(e) Province Center Indicator



(f) District Center Indicator

Figure 16: RD Plots for Predetermined Covariates (Meyersson Application)

The graphical analysis does not reveal obvious discontinuities at the cutoff, but of course a statistical analysis is required before we can reach a formal conclusion. In order to implement the analysis, an optimal bandwidth must be chosen for each covariate. Crucially, these bandwidths will be generally different from the bandwidth used to analyze the original outcome of interest. As shown in the RD plots, each covariate exhibits a different estimated regression function, with different curvature and overall shape. As a result, the optimal bandwidth for local polynomial estimation and inference will be different for every variable, and must be re-estimated accordingly in each case. This implies that the statistical analysis must be conducted separately for each covariate, choosing a different optimal bandwidth for each covariate analyzed.

To implement this formal falsification test, we simply run `rdrobust` using each covariate of interest as the outcome variable. As an example, we analyze the covariate `lpop1994`, the logarithm of the municipality population in 1994. Since this covariate was measured in 1994, it could not have been affected by the treatment, that is, by the party that wins the 1994 election. We estimate a local linear RD effect with triangular kernel weights and common MSE-optimal bandwidth using the default options in `rdrobust`.

R Snippet 29

```
> out = rdrobust(data$lpop1994, X)
> summary(out)
Call: rdrobust

Number of Obs.                  2629
BW type                        mserd
Kernel                    Triangular
VCE method                        NN

Number of Obs.                 2314         315
Eff. Number of Obs.             400         233
Order est. (p)                    1           1
Order bias  (p)                   2           2
BW est. (h)                  13.319      13.319
BW bias (b)                  21.366      21.366
rho (h/b)                     0.623       0.623

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     0.012     0.278     0.045     0.964     [-0.532 , 0.557]
        Robust         -         -     0.001     0.999     [-0.644 , 0.645]
================================================================================
```

Stata Snippet 29

```
. rdrobust lpop1994 X
```

The point estimate is very close to zero and the robust p-value is 0.999, so we find no evidence that, at the cutoff, treated and control municipalities differ systematically in this covariate. In other words, we find no evidence that the population size of the municipalities is discontinuous at the cutoff. In order to provide a complete falsification test, the same estimation and inference procedure should be repeated for all important covariates, that is, for all available covariates that would be expected to be correlated with the treatment in the presence of manipulation. In a convincing RD design, these tests would show that there are no discontinuities in any variable. Table 4 contains the local polynomial estimation and inference results for several predetermined covariates in the Meyersson dataset. All results were obtained employing `rdrobust` with the default specifications, as shown for `lpop1994` above.

Table 4: Formal Continuity-Based Analysis for Covariates

Variable	MSE-Optimal Bandwidth	RD Estimator	Robust Inference p-value	Robust Inference Conf. Int.	Eff. Number Observations
Percentage of men aged 15-20 with high school education	12.055	1.561	0.358	[-1.757, 4.862]	590
Islamic Mayor in 1989	11.782	0.053	0.333	[-0.077, 0.228]	418
Islamic percentage of votes in 1994	13.940	0.603	0.711	[-2.794, 4.095]	668
Number of parties receiving votes 1994	12.166	-0.168	0.668	[-1.357, 0.869]	596
Log population in 1994	13.319	0.012	0.999	[-0.644, 0.645]	633
District center	13.033	-0.067	0.462	[-0.285, 0.130]	624
Province center	11.556	0.029	0.609	[-0.064, 0.109]	574
Sub-metro center	10.360	-0.016	0.572	[-0.114, 0.063]	513
Metro center	13.621	0.008	0.723	[-0.047, 0.068]	642

All point estimates are small and all 95% robust confidence intervals contain zero, with p-values ranging from 0.333 to 0.999. In other words, there is no empirical evidence that these predetermined covariates are discontinuous at the cutoff. Note that the number of observations used in the analysis varies for each covariate; this occurs because the MSE-optimal bandwidth is different for every covariate analyzed. Note also that we employ the default `rdrobust` options for simplicity, but for falsification purposes it may be more appropriate to use the CER-optimal bandwidth because, in this case, we are only interested in testing the null hypothesis of no effect, that is, we are mostly interested in inference and the point estimates are of no particular interest. These two alternative bandwidth choices give a natural trade-off between size and power of the falsification tests: the MSE-optimal bandwidth leads to more powerful hypothesis tests with possibly larger size distortions than tests implemented using the CER-optimal bandwidth. In this application, switching to `bwselect="cerrd"` does not change any of the empirical conclusions (results available in the replication files).

We complement these results with a graphical illustration of the RD effects for every covariate, to provide further evidence that in fact these covariates do not jump discretely at the cutoff. For this, we employ `rdplot` with the same options we used for inference in `rdrobust`: we plot each covariate inside their respective MSE-optimal bandwidth, using a polynomial of order one, and a triangular kernel function to weigh the observations. Below we illustrate the specific command for the `lpop1994` covariate.

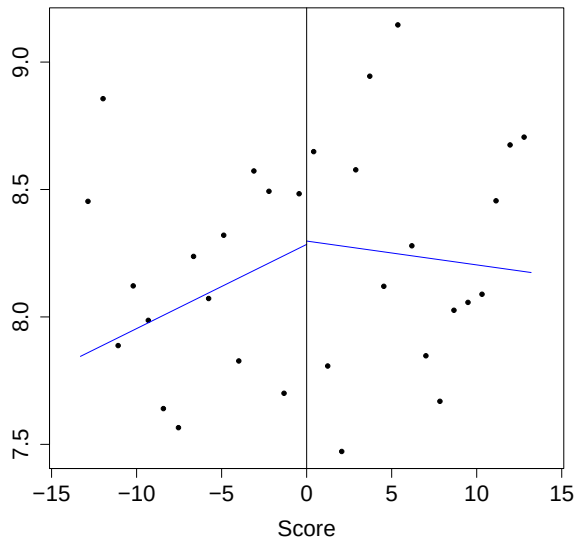
R Snippet 30

```
> bandwidth = rdrobust(data$lpop1994, X)$bws[1, 1]
> xlim = ceiling(bandwidth)
> rdplot(data$lpop1994[abs(X) <= bandwidth], X[abs(X) <= bandwidth],
+ p = 1, kernel = "triangular", x.lim = c(-xlim, xlim), x.label = "Score",
+ y.label = "", title = "", cex.axis = 1.5, cex.lab = 1.5)
```

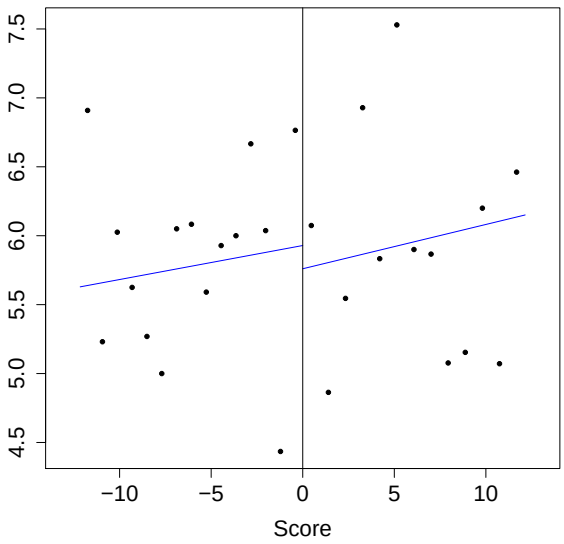
Stata Snippet 30

```
. rdrobust lpop1994 X
. local bandwidth = e(h_l)
. rdplot lpop1994 X if abs(X) <= `bandwidth', h(`bandwidth') p(1) kernel(triangular)
```

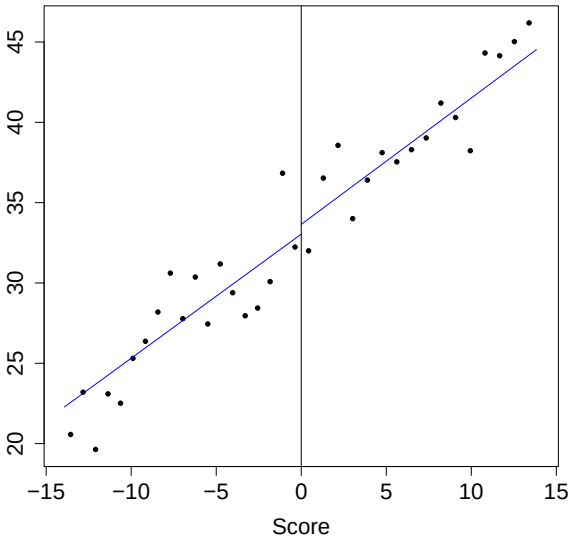
We run the same commands for each covariate. A sample of the resulting plots is presented in Figure 17. Consistent with the formal statistical results, the graphical analysis within the optimal bandwidth shows that the right and left intercepts in the local linear fits are very close to each other in most cases (the variable `Islamic Mayor in 1989` shows a more noticeable jump, but the formal analysis above indicates that this jump is not distinguishable from zero).



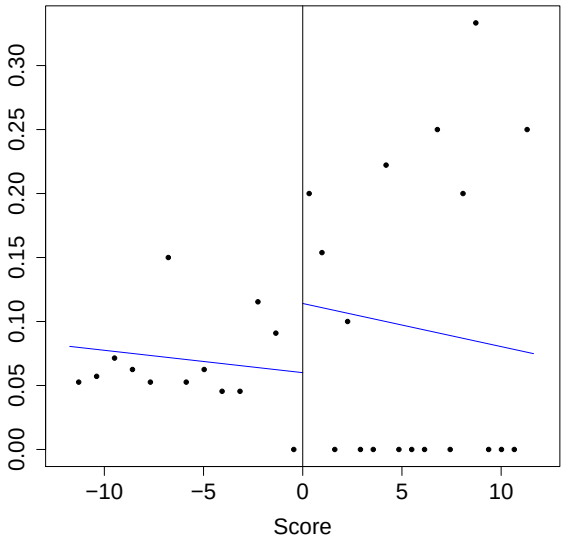
(a) Log Population in 1994



(b) Number of Parties Receiving Votes in 1994



(c) Islamic Vote Percentage in 1994



(d) Islamic Mayor in 1989

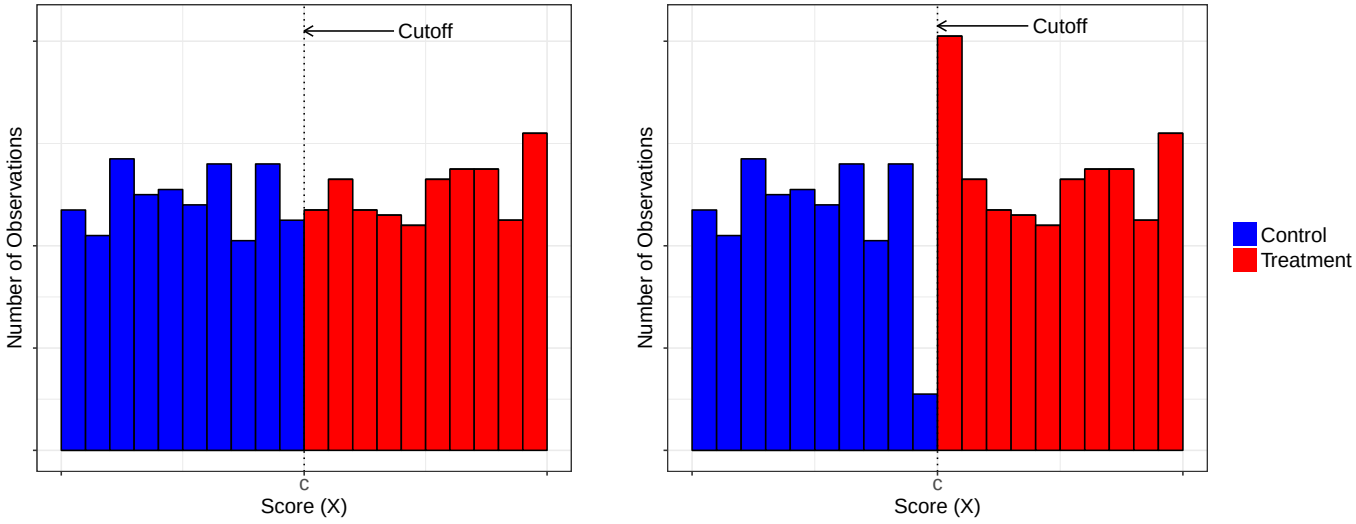
Figure 17: Graphical Illustration of Local Linear RD Effects for Predetermined Covariates (Meyersson data)

The plots and estimated effects for the covariates stand in contrast to the analogous results we reported for the outcome of interest in the previous sections, where, despite some variability, we saw a more noticeable jump at the cutoff. In general, a stark contrast between null effects for the covariates and a large nonzero effect for the outcome can be interpreted as evidence in favor of the validity of the RD design. However, the converse is not true, as it is possible to see a valid RD design where the treatment has no effect on the outcome, and thus where both covariate and outcome results are null.

5.2 Density of Running Variable

The second type of falsification test examines whether, in a local neighborhood near the cutoff, the number of observations below the cutoff is surprisingly different from the number of observations above it. The underlying assumption is that, if units 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. In other words, even if units actively attempt to affect their score, in the absence of precise manipulation, random change would place roughly the same amount of units on either side of the cutoff, leading to a continuous probability density function when the score is continuously distributed. RD applications where there is an abrupt change in the number of observations at the cutoff will tend to be less credible.

Figure 18 shows a histogram of the running variable in two hypothetical RD examples. In the scenario illustrated in Figure 18(a), the number of observations above and below the cutoff is very similar. In contrast, Figure 18(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 suggests that units were able to systematically increase the value of their original score to be assigned to the treatment instead of the control group.



(a) No Sorting (b) Sorting

Figure 18: Histogram of Score

In addition to a graphical illustration of the density of the running variable, researchers should explore the assumption more formally using a statistical test, often called a density test. One possible strategy is to choose a small neighborhood around the cutoff, and perform

a simple Bernoulli test within that neighborhood with a probability of "success" equal to 1/2. This strategy tests whether the number of treated observations in the chosen neighborhood is compatible with what would have been observed if units had been assigned to the treatment group (i.e., to being above the cutoff) with a 50% probability. The test is finite sample exact, under the assumptions imposed.

For example, if we keep only the observations with $X_i \in [-2, 2]$ in the Meyersson application, we find that in this neighborhood there are 47 control observations and 53 treated observations. Using this information and setting a probability of success equal to 1/2, we can perform a binomial test using standard functions in `R` or `Stata`.

R Snippet 31

```
> binom.test(53, 100, 1/2)

        Exact binomial test

data:  53 and 100
number of successes = 53, number of trials = 100, p-value = 0.6173
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4275815 0.6305948
sample estimates:
probability of success
                  0.53
```

Stata Snippet 31

```
. bitesti 100 53 1/2
```

The p-value is 0.6173, so this simple test finds no evidence of "sorting" around the cutoff in this neighborhood: the numbers of treated and control observations are consistent with what would be expected if municipalities were assigned to an Islamic win or loss by the flip of an unbiased coin.

In a continuity-based approach, however, there are often not clear guidelines about how to choose the neighborhood where the binomial test should be conducted. Nevertheless, it is natural to conduct this test for different (nested) neighborhoods around the cutoff. Furthermore, the use of this randomization-based test is also natural in the context of a local randomization approach to RD analysis, which we discuss extensively in the accompanying Element (Cattaneo, Idrobo, and Titiunik, forthcoming).

A complementary approach is to conduct a test of the null hypothesis that the density of the running variable is continuous at the cutoff, which fits naturally into the continuity-based framework adopted in this Element. The implementation of this test requires the estimation of the density of observations near the cutoff, separately for observations above and below the cutoff. We employ here an implementation based on a local polynomial density estimator that does not require pre-binning of the data and leads to size and power improvements relative to other approaches. The null hypothesis is that there is no "manipulation" of the density at the cutoff, formally stated as continuity of the density functions for control and treatment units at the cutoff. Therefore, failing to reject implies that there is no statistical evidence of manipulation at the cutoff, and offers evidence supporting the validity of the RD design.

We implement this density test using the Meyersson data using the `rddensity` command, which is part of the `rddensity` library/package. Its only required argument is the running variable.

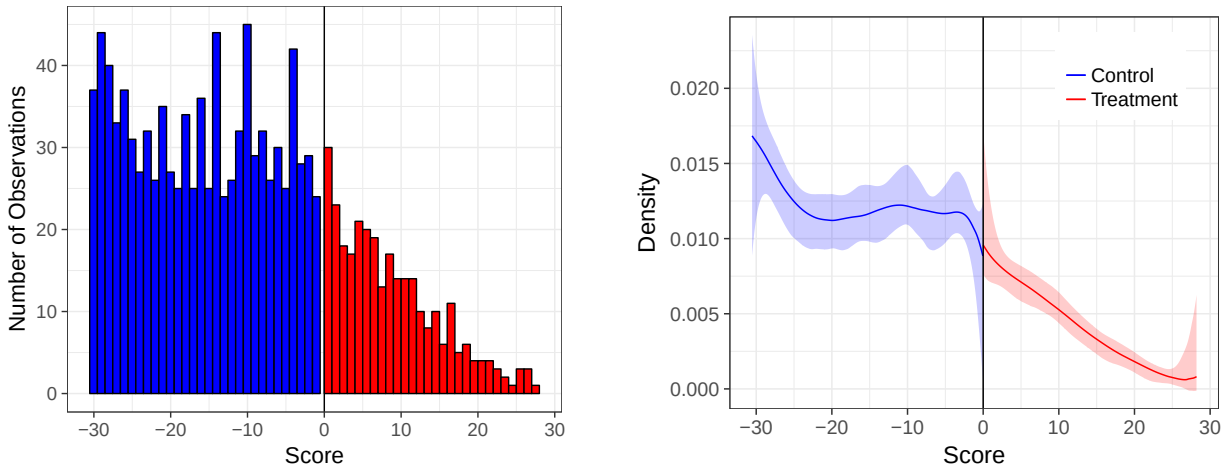
R Snippet 32

```
> out = rddensity(X)
> summary(out)

RD Manipulation Test using local polynomial density estimation.

Number of obs =       2629
Model =               unrestricted
Kernel =              triangular
BW method =           comb
VCE method =          jackknife

Cutoff c = 0          Left of c           Right of c
Number of obs         2314                315
Eff. Number of obs    965                 301
Order est. (p)        2                   2
Order bias (q)        3                   3
BW est. (h)           30.54               28.285

Method                T                   P > |T|
Robust                -1.394              0.1633
```

Stata Snippet 32

```
. rddensity X
```

Figure 19: Histogram and Estimated Density of the Score



(a) Histogram (b) Estimated Density

The value of the statistic is −1.394 and the associated p-value is 0.1633. This means that under the continuity-based approach, we fail to reject the null hypothesis of no difference in the density of treated and control observations at the cutoff. Figure 19 provides a graphical representation of the continuity in density test approach, exhibiting both a histogram of the data and the actual density estimate with shaded 95% confidence intervals. As we can see in 19(b), the density estimates for treated and control groups at the cutoff (the two intercepts in the figure) are very near each other, and the confidence intervals (shaded areas) overlap. This plot is consistent with the results from the formal test.

5.3 Placebo Cutoffs

Another useful falsification analysis examines treatment effects at artificial or placebo cutoff values. To understand the motivation behind this falsification test, recall that the key RD identifying assumption is the continuity (or lack of abrupt changes) of the regression functions for treatment and control units at the cutoff in the absence of the treatment. While such a condition is fundamentally untestable at the cutoff, researchers can investigate empirically whether the estimable regression functions for control and treatment units are continuous at points other than the cutoff. Evidence of continuity away from the cutoff is, of course, neither necessary nor sufficient for continuity at the cutoff, but the presence of discontinuities away from the cutoff can be interpreted as potentially casting doubt on the RD design, at the very least in cases where such discontinuities can not be explained by substantive knowledge of the

specific application. Another related use of this approach is to check whether the smoothness and other conditions needed for RD inference are supported by the data, at least in regions other than at the cutoff point.

This test replaces the true cutoff value by another value at which the treatment status does not really change, and performs estimation and inference using this artificial cutoff point. The expectation is that no significant treatment effect will occur at placebo cutoff values. A graphical implementation of this falsification approach follows directly from the RD plots discussed extensively in Section 3, by simply assessing whether there are jumps in the observed regression functions at points other than the true cutoff. A more formal implementation of this idea conducts statistical estimation and inference for RD treatment effects at artificial cutoff points, using control and treatment units separately. In the continuity-based framework adopted in this Element, we implement this approach using local-polynomial methods within an optimally-chosen bandwidth around the artificial cutoff to estimate treatment effects on the outcome, as we explained in Section 4.

In order to illustrate the procedure with the Meyersson data, we employ `rdrobust` after restricting to the appropriate group and specifying the artificial cutoff. To avoid "contamination" due to real treatment effects, for artificial cutoffs above the real cutoff we use only treated observations, and for artificial cutoffs below the real cutoff we use only control observations. Restricting the observations in this way guarantees that the analysis of each placebo cutoff uses only observations with the same treatment status. Thus, by construction, the treatment effect at each artificial cutoff should be zero.

We conduct estimation and inference at the artificial cutoff $c = 1$ in the Meyersson application, using the option `c = 1` in `rdrobust` and including only treated observations. Our analysis thus compares the educational outcomes of municipalities where Islamic mayors won by a margin of 1% or more, to municipalities where Islamic mayors won by less than 1%. Since there is an Islamic mayor on both sides of the cutoff, we expect to see no discontinuity in the outcome at 1%.

R Snippet 33

```
> out = rdrobust(Y[X >= 0], X[X >= 0], c = 1)
> summary(out)
Call: rdrobust

Number of Obs.                  315
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                  30         285
Eff. Number of Obs.             30          49
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                  2.362       2.362
BW bias (b)                  3.326       3.326
rho (h/b)                    0.710       0.710

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional    -1.131     4.252    -0.266     0.790    [-9.464 , 7.202]
       Robust          -         -     0.270     0.787    [-9.967 , 13.147]
================================================================================
```

Stata Snippet 33

```
. rdrobust Y X if X >= 0, c(1)
```

The robust p-value is 0.787, consistent with the conclusion that the outcome of interest does not jump at the artificial 1% cutoff, and in contrast to the results at the true cutoff reported in Section 4. Table 5 presents the results of similar analyses for other placebo cutoffs ranging from −5% to 5% in increments of 1%. Figure 20 graphically illustrates the main results from this falsification test.

Table 5: Continuity-Based Analysis for Alternative Cutoffs

Alternative Cutoff	MSE-Optimal Bandwidth	RD Estimator	Robust Inference p-value	Robust Inference Conf. Int.	N. of Obs. Left	N. of Obs. Right
−3	3.934	1.688	0.421	[−3.509, 8.397]	135	74
−2	4.642	−2.300	0.991	[−9.414, 9.518]	152	47
−1	4.510	−3.003	0.992	[−11.295, 11.409]	139	24
0	17.239	3.020	0.076	[−0.309, 6.276]	529	266
1	2.362	−1.131	0.787	[−9.967, 13.147]	30	49
2	2.697	−1.973	0.488	[−15.333, 7.313]	53	50
3	2.850	3.766	0.668	[−8.700, 13.569]	68	56

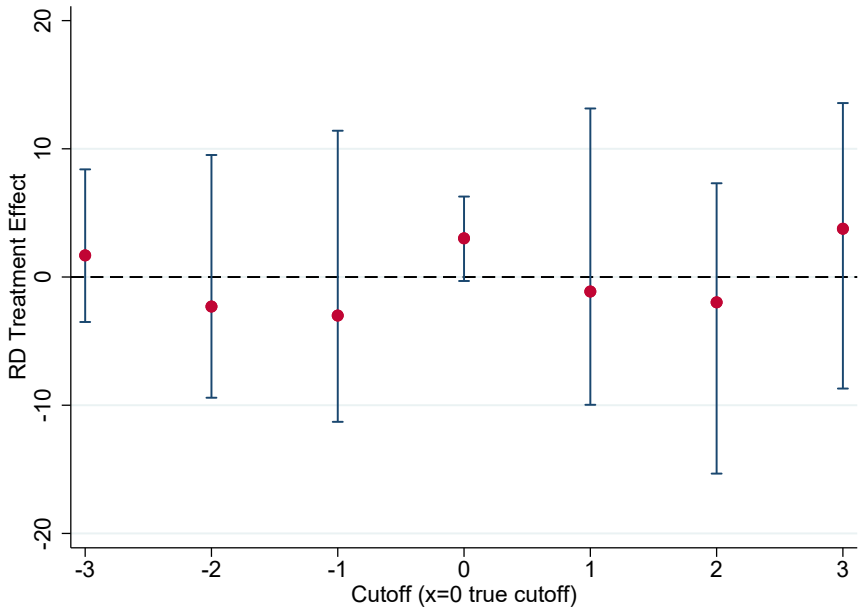


Figure 20: RD Estimation for True and Artificial Cutoffs

The true cutoff of 0 is included in order to have a benchmark to compare—the particular results regarding the true cutoff were discussed at length in Section 4. All other cutoffs are artificial or placebo, in the sense that treatment did not actually change at those points. We find that in all but one of the artificial cutoff points, the RD point estimator is smaller in absolute value than the true RD estimate (3.020), and that all p-values are above 0.4. Therefore, we conclude that the outcome of interest does not jump discontinuously at the artificial cutoffs considered.

5.4 Sensitivity to Observations near the Cutoff

Another falsification approach seeks to investigate how sensitive the results are to the response of units who are located very close to the cutoff. If systematic manipulation of score values has occurred, it is natural to assume that the units closest to the cutoff are those most likely to have engaged in manipulation. The idea behind this approach is to exclude such units and then repeat the estimation and inference analysis using the remaining sample. This idea is sometimes referred to as a "donut hole" approach. Even when manipulation of the score is not suspected, this strategy is also useful to assess the sensitivity of the results to the unavoidable extrapolation involved in local polynomial estimation, as the few observations closest to the cutoff are likely to be the most influential when fitting the local polynomials.

For implementation in the continuity-based approach, we use `rdrobust` after subsetting the data. For example, in the Meyersson application, we consider first the case where units with score $|X_i| < 0.3$ are excluded from the analysis, which requires us to engage in more extrapolation than before. The exclusion of observations implies that a new optimal bandwidth will be selected.

R Snippet 34

```
> out = rdrobust(Y[abs(X) >= 0.3], X[abs(X) >= 0.3])
> summary(out)
Call: rdrobust

Number of Obs.                 2616
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                2307         309
Eff. Number of Obs.            482         248
Order est. (p)                   1           1
Order bias  (p)                  2           2
BW est. (h)                 16.043      16.043
BW bias (b)                 27.520      27.520
rho (h/b)                    0.583       0.583

================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     3.414     1.517     2.251     0.024     [0.441 , 6.387]
       Robust         -         -     1.923     0.055    [-0.067 , 6.965]
================================================================================
```

Stata Snippet 34

```
. rdrobust Y X if abs(X) >= 0.3
```

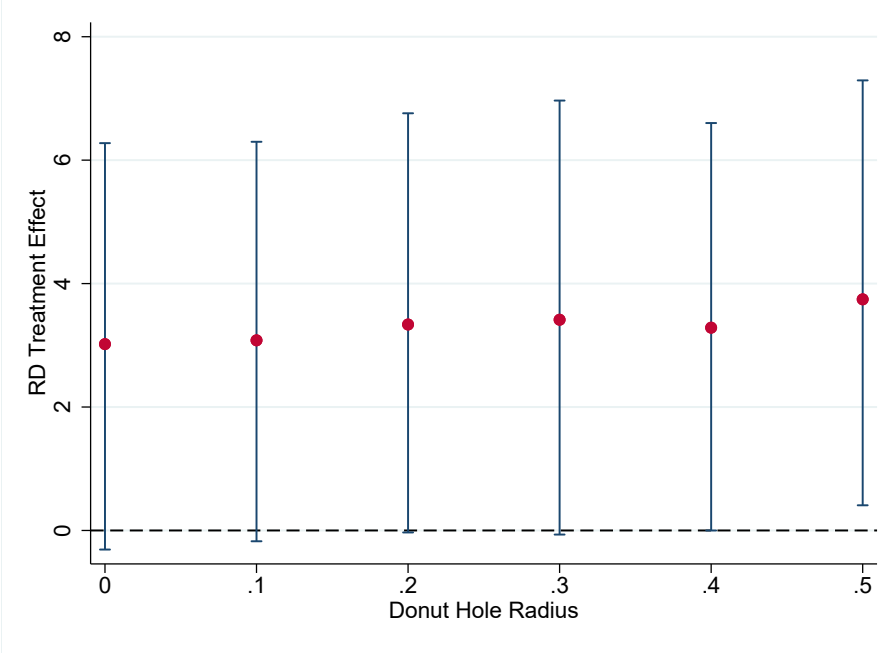
The results show that the conclusions from the analysis are robust to excluding observations with $|X_i| < 0.3$. In the new analysis, we have 2307 total observations to the left of the cutoff, and 309 total observations to the right of it. As expected, these numbers are smaller than those employed in the original analysis (2314 and 315). Note that, although the total number of observations will always decrease when observations closest to the cutoff are excluded, the effective number of observations used in the analysis may increase or decrease, depending on how the bandwidth changes. In this case, the bandwidth changes from 17.239 in the original analysis to 16.043 in the analysis that excludes units with $|X_i| < 0.30$; this results in a loss of 65 effective observations, from 795 (529 + 266) to 730 (482 + 248), which is much larger than the decrease in total observations, which is only 13. The exclusion of these observations changes the point estimate from 3.020 to 3.414, and the robust confidence interval from $[-0.309, 6.276]$ to $[-0.067, 6.965]$. The conclusion of the analysis remains largely unchanged, however, since both the original and the new estimated effect are significant at 10% level.

In practice, it is natural to repeat this exercise a few times to assess the actual sensitivity for different amounts of excluded units. Table 6 illustrates this approach, and Figure 21 depicts the results graphically. In all the cases considered, the conclusions remain unchanged.

Table 6: Continuity-Based Analysis for the Donut-Hole Approach

Donut-Hole Radius	MSE-Optimal Bandwidth	RD Estimator	Robust Inference p-value	Robust Inference Conf. Int.	Number of Observations	Excluded Obs. Left	Excluded Obs. Right
0.00	17.239	3.020	0.076	[-0.309, 6.276]	795	0	0
0.10	17.954	3.081	0.064	[-0.175, 6.298]	815	1	1
0.20	16.621	3.337	0.052	[-0.033, 6.759]	765	5	4
0.30	16.043	3.414	0.055	[-0.067, 6.965]	730	7	6
0.40	17.164	3.286	0.050	[-0.001, 6.601]	774	9	9
0.50	15.422	3.745	0.028	[0.408, 7.292]	697	13	14

Figure 21: RD Estimation for the Donut-Hole Approach



5.5 Sensitivity to Bandwidth Choice

The last falsification method we discuss analyzes the sensitivity of the results to the bandwidth choice. In contrast to the donut hole approach, which investigates sensitivity as units from the center of the neighborhood around the cutoff are removed, the method we discuss now investigates sensitivity as units are added or removed at the end points of the neighborhood. The implementation of this method is also straightforward, as it requires employing local polynomial methods with different bandwidth choices. However, the interpretation of the results must be done with care. As we discussed throughout this Element, choosing the bandwidth is one of the most consequential decisions in RD analysis, because the bandwidth may affect the results and conclusions.

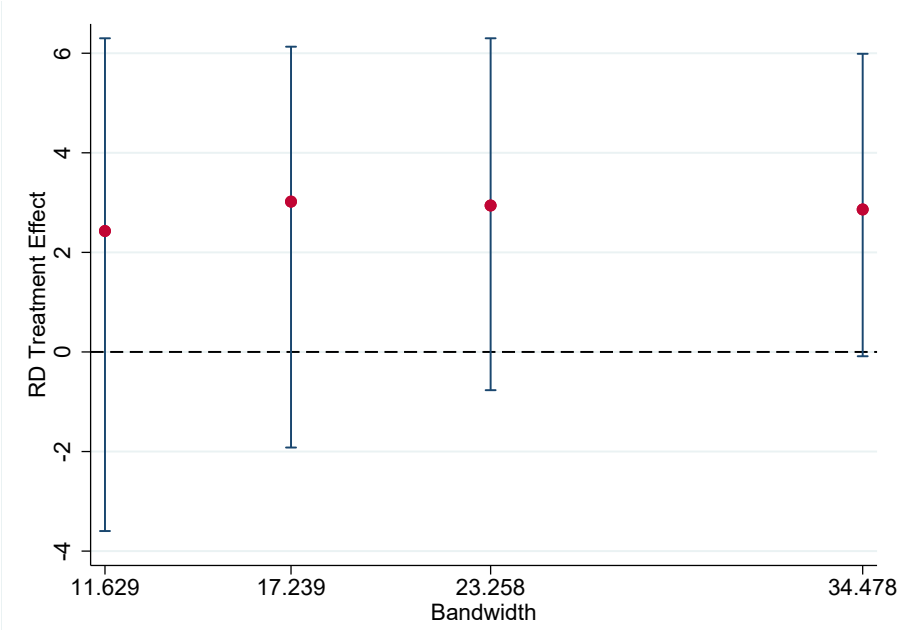
In the continuity-based approach, this falsification test is implemented by changing the bandwidth used for local polynomial estimation. It is well understood how the bandwidth will affect the results: as the bandwidth increases, the bias of the local polynomial estimator increases and its variance decreases. Thus, it is natural to expect that, as we increase the bandwidth, the confidence intervals will decrease in length but will also be displaced (because of the bias).

The considerations above suggest that when the goal is to interpret point estimators, investigating the sensitivity to bandwidth choices is only useful over small ranges around

the MSE-optimal bandwidth; otherwise, the results will be mechanically determined by the statistical properties of the estimation and inference methods. In other words, bandwidths much larger than the MSE-optimal bandwidth will lead to estimated RD effects that have too much bias, and bandwidths much smaller than the MSE-optimal choice will lead to RD effects with too much variance. In both cases, point estimation will be unreliable, and so will be the conclusions from the falsification test. Similarly, if the emphasis is on optimal inference, the sensitivity of the results should only be explored for bandwidth values near the CER-optimal choice.

We illustrate this sensitivity approach with the Meyersson data for four bandwidth choices close to the MSE-optimal and CER-optimal choices: (i) the CER-optimal choice $h_{\texttt{CER}} = 11.629$, (ii) the MSE-optimal choice $h_{\texttt{MSE}} = 17.239$, (iii) $2 \cdot h_{\texttt{CER}} = 23.258$, and (iv) $2 \cdot h_{\texttt{MSE}} = 34.478$. Figure 22 shows the local polynomial RD point estimators and robust 95% confidence intervals for each bandwidth. The code is omitted, but is included in the replication files.

Figure 22: Sensitivity to Bandwidth in the Continuity-Based Approach



The results based on the CER-optimal choice $h_{\texttt{CER}} = 11.629$ are consistent with the results based on the MSE-optimal choice $h_{\texttt{MSE}} = 17.239$ in that they both lead to a similar point estimate, but the CER-optimal choice results in a longer confidence interval according to which the effect cannot be distinguished from zero at conventional levels. The two largest bandwidths, $2 \cdot h_{\texttt{CER}} = 23.258$ and $2 \cdot h_{\texttt{MSE}} = 34.478$, lead to results that are broadly consistent with the empirical findings obtained with the MSE-optimal choice.

5.6 Further Reading

The density test to detect RD manipulation was first proposed by McCrary (2008). Cattaneo, Jansson, and Ma (2019) develop the local polynomial density estimator implemented in `rddensity`; see also Cattaneo, Jansson, and Ma (2018) for details on this statistical package and further numerical evidence. Frandsen (2017) develops a related manipulation test for cases where the score is discrete. The importance of falsification tests and the use of placebo outcomes is generally discussed in the analysis of experiments literature (e.g., Imbens and Rubin, 2015; Rosenbaum, 2002, 2010). Lee (2008) applies and extends these ideas to the context of RD designs, and Canay and Kamat (2018) develop a permutation inference approach in the same context. Ganong and Jäger (2018) develop a permutation inference approach based on the idea of placebo RD cutoffs for the Kink RD designs, Regression Kink designs, and related settings. Finally, falsification testing based on donut hole specifications is discussed in Bajari, Hong, Park, and Town (2011) and Barreca, Lindo, and Waddell (2016), among others.

6 Final Remarks

We have discussed foundational aspects of identification, estimation, inference, and falsification in the Sharp RD design, when the parameter of interest is the average treatment effect at the cutoff. Because our goal in this Element was to discuss the conceptual foundations of RD methodology, we focused on the simplest possible case where (i) there is a single running variable, (ii) there is a single cutoff, (iii) compliance with treatment assignment is perfect, (iv) the running variable is continuous and hence has no mass points, (v) the object of interest is the average treatment effect at the cutoff, and (vi) results are based on continuity and smoothness assumptions. This canonical setup is the most standard and commonly encountered in empirical work dealing with RD designs.

In the accompanying Element (Cattaneo, Idrobo, and Titiunik, forthcoming), we discuss several departures from the canonical Sharp RD design setting. The first topic we consider is an alternative interpretation of the RD design based on the idea of local random assignment. In contrast to the continuity-based approach adopted in this Element, the local randomization approach assumes that there is a window around the cutoff where the treatment can be assumed to have been as-if randomly assigned, and the analysis proceeds by adopting the usual tools from the analysis of experiments. This approach is also well suited to analyze RD designs where the running variable is discrete with relatively few mass points, a situation that occurs often in practice and we also discuss in detail. Additional topics covered in the accompanying Element include the Fuzzy RD design, where compliance with treatment is imperfect, RD settings with multiple running variables, which have as an important special case the geographic RD design where treatment assignment depends on the spatial distance to the border between geographic regions, and RD setups where treatment assignment depends on multiple cutoffs instead of only one.

We hope that the discussion in this Element, together with the additional methods presented in the accompanying Element, will provide a useful and practical template to guide applied researchers in analyzing and interpreting RD designs in a principled, rigorous, and transparent way.

Bibliography

Abadie, A., and M. D. Cattaneo (2018): "Econometric Methods for Program Evaluation," *Annual Review of Economics*, 10, 465–503.

Angrist, J. D., and M. Rokkanen (2015): "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutoff," *Journal of the American Statistical Association*, 110(512), 1331–1344.

Arai, Y., and H. Ichimura (2018): "Simultaneous Selection of Optimal Bandwidths for the Sharp Regression Discontinuity Estimator," *Quantitative Economics*, 9(1), 441–482.

Bajari, P., H. Hong, M. Park, and R. Town (2011): "Regression Discontinuity Designs with an Endogenous Forcing Variable and an Application to Contracting in Health Care," NBER Working Paper No. 17643.

Barreca, A. I., J. M. Lindo, and G. R. Waddell (2016): "Heaping-Induced Bias in Regression-Discontinuity Designs," *Economic Inquiry*, 54(1), 268–293.

Bartalotti, O., and Q. Brummet (2017): "Regression Discontinuity Designs with Clustered Data," in *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, ed. by M. D. Cattaneo, and J. C. Escanciano, pp. 383–420. Emerald Group Publishing.

Bartalotti, O., G. Calhoun, and Y. He (2017): "Bootstrap Confidence Intervals for Sharp Regression Discontinuity Designs," in *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, ed. by M. D. Cattaneo, and J. C. Escanciano, pp. 421–453. Emerald Group Publishing.

Bertanha, M., and G. W. Imbens (2019): "External Validity in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, forthcoming.

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2018): "On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference," *Journal of the American Statistical Association*, 113(522), 767–779.

——— (2019a): "Coverage Error Optimal Confidence Intervals for Local Polynomial Regression," arXiv:1808.01398.

——— (2019b): "Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression Discontinuity Designs," arXiv:1809.00236.

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik (2017): "rdrobust: Software for Regression Discontinuity Designs," *Stata Journal*, 17(2), 372–404.

——— (2019): "Regression Discontinuity Designs Using Covariates," *Review of Economics and Statistics*, 101(3), 442–451.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014a): "Robust Data-Driven Inference in the Regression-Discontinuity Design," *Stata Journal*, 14(4), 909–946.

——— (2014b): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82(6), 2295–2326.

——— (2015a): "Optimal Data-Driven Regression Discontinuity Plots," *Journal of the American Statistical Association*, 110(512), 1753–1769.

——— (2015b): "rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs," *R Journal*, 7(1), 38–51.

Canay, I. A., and V. Kamat (2018): "Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design," *Review of Economic Studies*, 85(3), 1577–1608.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2019): "On Binscatter," arXiv:1902.09608.

Cattaneo, M. D., and J. C. Escanciano (2017): *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*. Emerald Group Publishing.

Cattaneo, M. D., and M. H. Farrell (2013): "Optimal convergence rates, Bahadur representation, and asymptotic normality of partitioning estimators," *Journal of Econometrics*, 174(2), 127–143.

Cattaneo, M. D., M. H. Farrell, and Y. Feng (2019): "Large Sample Properties of Partitioning-Based Series Estimators," *Annals of Statistics*, forthcoming.

Cattaneo, M. D., B. Frandsen, and R. 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, M. D., N. Idrobo, and R. Titiunik (in press): *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Cambridge Elements: Quantitative and Computational Methods for Social Science, Cambridge University Press.

Cattaneo, M. D., M. Jansson, and X. Ma (2018): "Manipulation Testing based on Density Discontinuity," *Stata Journal*, 18(1), 234–261.

——— (2019): "Simple Local Polynomial Density Estimators," *Journal of the American Statistical Association*, forthcoming.

Cattaneo, M. D., L. Keele, R. Titiunik, and G. Vazquez-Bare (2016): "Interpreting Regression Discontinuity Designs with Multiple Cutoffs," *Journal of Politics*, 78(4), 1229–1248.

——— (2019): "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs," arXiv:1808.04416.

Cattaneo, M. D., and R. Titiunik (2019): "Regression Discontinuity Designs: A Review," manuscript in preparation, Princeton University.

Cattaneo, M. D., R. Titiunik, and G. Vazquez-Bare (2016): "Inference in Regression Discontinuity Designs under Local Randomization," *Stata Journal*, 16(2), 331–367.

——— (2017): "Comparing Inference Approaches for RD Designs: A Reexamination of the Effect of Head Start on Child Mortality," *Journal of Policy Analysis and Management*, 36(3), 643–681.

——— (2018): "Power Calculations for Regression Discontinuity Designs," *Stata Journal*, 19(1), 210–245.

——— (2019): "The Regression Discontinuity Design," in *Handbook of Research Methods in Political Science and International Relations*, ed. by L. Curini, and R. J. Franzese. Sage Publications.

Cattaneo, M. D., and G. Vazquez-Bare (2016): "The Choice of Neighborhood in Regression Discontinuity Designs," *Observational Studies*, 2, 134–146.

Cook, T. D. (2008): ""Waiting for Life to Arrive": A History of the Regression-Discontinuity Design in Psychology, Statistics and Economics," *Journal of Econometrics*, 142(2), 636–654.

Dong, Y. (2019): "Regression Discontinuity Designs with Sample Selection," *Journal of Business & Economic Statistics*, 37(1), 171–186.

Dong, Y., Y.-Y. Lee, and M. Gou (2019): "Regression Discontinuity Designs with a Continuous Treatment," SSRN working paper No. 3167541.

Dong, Y., and A. Lewbel (2015): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," *Review of Economics and Statistics*, 97(5), 1081–1092.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications*, vol. 66 of *Monographs on Statistics and Applied Probability*. CRC Press.

Frandsen, B. (2017): "Party Bias in Union Representation Elections: Testing for Manipulation in the Regression Discontinuity Design When the Running Variable is Discrete," in *Regression Discontinuity Designs: Theory and Applications (Advances in Econometrics, volume 38)*, ed. by M. D. Cattaneo, and J. C. Escanciano, pp. 281–315. Emerald Group Publishing.

Galiani, S., P. Gertler, and E. Schargrodsky (2005): "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy*, 113(1), 83–120.

Ganong, P., and S. Jäger (2018): "A Permutation Test for the Regression Kink Design," *Journal of the American Statistical Association*, 113(522), 494–504.

Gelman, A., and G. W. Imbens (2019): "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 37(3), 447–456.

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

Hyytinen, A., J. Meriläinen, T. Saarimaa, O. Toivanen, and J. Tukiainen (2018): "When Does Regression Discontinuity Design Work? Evidence from Random Election Outcomes," *Quantitative Economics*, 9(2), 1019–1051.

Imbens, G., and D. B. Rubin (2015): *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

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

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

Klašnja, M., and R. Titiunik (2017): "The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability," *American Political Science Review*, 111(1), 129–148.

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

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

Ludwig, J., and D. L. Miller (2007): "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics*, 122(1), 159–208.

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

Meyersson, E. (2014): "Islamic Rule and the Empowerment of the Poor and Pious," *Econometrica*, 82(1), 229–269.

Pettersson-Lidbom, P. (2008): "Do Parties Matter for Economic Outcomes? A Regression-Discontinuity Approach," *Journal of the European Economic Association*, 6(5), 1037–1056.

Rosenbaum, P. R. (2002): *Observational Studies*. Springer, New York.

——— (2010): *Design of Observational Studies*. Springer, New York.

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

Wing, C., and T. 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.

Xu, K.-L. (2017): "Regression Discontinuity with Categorical Outcomes," *Journal of Econometrics*, 201(1), 1–18.