Randomization Inference in the Regression Discontinuity Design: Re-examining the empirical evidence on incumbency advantage in the U.S. House

Matias D. Cattaneo† Brigham Frandsen† Rocío Titiunik§

This draft: February, 2012
First draft: July, 2010

Abstract

We propose a randomization-inference framework to conduct exact finite-sample inference in the regression discontinuity (RD) design, following a recent strand of the literature that advocates interpreting this design as a local randomized experiment. In the RD design, units are assigned a treatment based on whether their value of an observed covariate or “score” is above or below a fixed cutoff. Our RD randomization-inference framework is motivated by the observation that local randomization is most likely to hold very near this cutoff, where only a few data points are available and hence standard large-sample procedures may be suspect. We apply this framework to re-examine the use of the RD design to study the incumbency advantage in the U.S. House, and present new evidence supporting a positive (conditional) effect of incumbency on electoral outcomes. Our methodological approach involves three steps. The first step is to select a window around the cutoff where a randomization-type condition holds. We develop data-driven, randomization-based procedures for selecting this window. The next step is to hypothesize a randomization mechanism assigning units within the window to either side of the cutoff. This choice necessarily depends on the specific application, but common examples include unrestricted randomization and a random allocation rule where the number of treated units is predetermined. The final step is to apply established randomization inference tools, using the hypothesized assignment mechanism, to conduct exact finite-sample inference for units with score inside the selected window. These tools include hypothesis tests, confidence intervals, point estimates, and inference on quantiles, leading to alternative finite-sample inference methods that complement existing procedures for the analysis of RD designs.

Keywords: Regression discontinuity, randomization inference, exact inference, as-if randomization, local experiments

*For valuable comments and suggestions, we thank Kosuke Imai, Luke Keele, and participants at the 2010 Political Methodology Meeting in the University of Iowa and at the Political Methodology Seminar in Princeton University.
†Department of Economics, University of Michigan
‡Robert Wood Johnson Scholar in Health Policy Research, Harvard University
§Department of Political Science, University of Michigan
1 Introduction

Empirical researchers across the social and behavioral sciences are often interested in estimating the effect of a treatment on an outcome. In observational studies, where controlled experimentation is not available, applied work relies on quasi-experimental strategies carefully tailored to eliminate the effect of potential confounders that would otherwise affect the validity of the analysis. Originally proposed by Thistlethwaite and Campbell (1960), the regression discontinuity (RD) design has recently become one of the most widely used quasi-experimental strategies. In this design, units receive treatment based on whether their value of an observed covariate or “score” is above or below a fixed cutoff or threshold. The key feature of the design is that the probability of receiving treatment jumps discontinuously at the cutoff, inducing variation in treatment assignment that is assumed to be unrelated with potential confounders. Recent reviews, including comprehensive lists of empirical examples, are given in Imbens and Lemieux (2008), Lee and Lemieux (2010) and Cook (2008), among others.

The traditional approach to conduct statistical inference in the RD design is to rely on flexible extrapolation using observations around the known cutoff, usually done by means of nonparametric curve estimation techniques such as local polynomial regression. This approach follows the work of Hahn, Todd, and van der Klaauw (2001), who showed that in a constant treatment effect model the key identifying assumption is that the conditional expectation of a potential outcome is continuous at the threshold. The idea is that, since nothing changes abruptly at the threshold other than the probability of treatment assignment, any differences in the limits of the conditional expectation of the outcome variable from either side of the threshold are attributed to the (average) causal effect of the treatment. Therefore, modern RD analysis employs local nonparametric curve estimation at a boundary point, with local linear regression being the preferred choice in most cases. See, among others, Porter (2003), McCrary (2008) and Imbens and Kalyanaraman (2012) for related theoretical results.

This traditional approach to identification and inference, however, bears little relationship with the widespread interpretation of RD designs as local randomized experiments, where in a neighborhood of the threshold treatment status is considered as good as randomly assigned (Lee 2008, Lee
Lee (2008) first argued that if individuals are unable to precisely manipulate their score (even if they can influence it to some degree), then variation in treatment near the threshold approximates a randomized experiment. This idea has been expanded in Lee and Lemieux (2010) and Dinardo and Lee (2011), where RD designs are described as the “close cousins” of randomized experiments. Moreover, the RD design has been found to closely replicate results from randomized experiments when both designs are available, further bolstering this “as-good-as randomized” interpretation. (See, e.g., Black, Galdo, and Smith 2007, Buddelmeyer and Skoufas 2003, Green, Leong, Kern, Gerber, and Larimer 2009).

Characterizing the RD design as a local randomized experiment not only has an intuitive appeal, but also leads to an alternative way of conducting statistical inference. In this paper, we develop a methodological framework for analyzing RD designs as local randomized experiments. Employing the extensive work of Rosenbaum (2002, 2010), we propose a randomization-inference framework to conduct exact finite-sample inference in the RD design. This idea is motivated by the observation that local randomization is most likely to hold very near the cutoff, where only a few data points are available and hence standard large-sample procedures may be suspect.\(^1\) Our approach complements existing procedures for the analysis of RD designs, providing alternative inference methods and a robustness check on the large-sample approximations usually invoked. We employ this framework to re-examine the use of the RD design to study the incumbency advantage in the U.S. House, and present new evidence supporting a positive (conditional) effect of incumbency on electoral outcomes.

To develop our methodological framework, we first make precise a set of conditions under which RD designs are equivalent to local randomized experiments within a randomization inference framework. These conditions are strictly stronger than the usual continuity assumptions imposed in the RD literature, but similar in spirit to those imposed in Hahn, Todd, and van der Klaauw (2001, Theorem 2) for identification of heterogeneous treatment effects. The key assumption is that there exists a neighborhood around the cutoff where a randomization-type condition holds. Employing these conditions, we then discuss how randomization inference tools may be used to

\(^1\)Randomization inference has proved useful in several treatment effect contexts. See, for example, Imbens and Rosenbaum (2005), Ho and Imai (2006), Hansen and Bowers (2009), and Barrios, Diamond, Imbens, and Kolesar (2010).
conduct exact finite-sample inference in the RD context, and we also propose different methods for implementation in applications.

Our resulting empirical approach is summarized as follows. The first step is choosing a neighborhood or window around the cutoff where treatment status is as-if randomly assigned. The size of this window is a critical choice, and we develop data-driven, randomization-based procedures for choosing its size. The next step is to hypothesize a randomization mechanism assigning units within the window to either side of the threshold. This choice necessarily depends on the specific application, but common examples include an unrestricted randomization mechanism that assign units to treatment by independent coin flips, a random allocation rule where the number of treated units is predetermined, and a group-randomization rule where clusters of units are jointly assigned to treatment. The final step is to apply established randomization inference tools (Rosenbaum 2002, Rosenbaum 2010) using the hypothesized assignment mechanism, to conduct exact finite-sample inference for units with score inside the selected window. These tools include hypothesis tests, confidence intervals, point estimates, and inference on quantiles.

To illustrate our methodological framework, we re-examine the RD-based empirical evidence on incumbency advantage in the U.S. House. In a two-party system, the incumbency status of political parties changes abruptly when the vote share reaches 50%. As first noted by Lee (2008), under the assumption that observed and unobserved factors that affect future vote shares change smoothly at the 50% cutoff, this discontinuity in incumbency status can be used to estimate a measure of incumbency advantage. This design compares U.S. House districts where the Democratic party barely won a given election to districts where the Democratic party barely lost, and calculates the difference in vote shares between these groups in the following election. This difference is often interpreted as a measure of the party incumbency advantage. In a recent contribution, Caughey and Sekhon (2011) challenge the identifying assumption in this RD design, providing evidence of considerable sorting around the cutoff. Using our randomization inference approach, we obtain novel results that support this evidence of sorting around the cutoff and find that, conditioning on whether the Democratic party won the previous election, this evidence of sorting is no longer present. Motivated by this finding, we estimate the incumbency advantage using our randomization inference framework in both the full sample (where our assumptions are likely to be violated) and
the two subsamples defined by whether the Democratic party won or lost in the previous election. As we discuss in further detail below, our method is particularly well-suited to estimating the incumbency advantage in these subsamples because the resulting sample sizes are small.

[TO BE COMPLETED AT THE VERY END AFTER ALL DATA RUNS ARE DONE] Our substantive findings include evidence of sorting and new results...

The rest of the paper is organized as follows. Section 2 introduces the problem of estimating the incumbency advantage in the U.S. House using RD methods, and motivates our framework using this example. Section 3 sets up our statistical framework, formally states the baseline assumptions required to apply randomization inference procedures to the RD design, and describes these procedures briefly. Section 4 discusses methods for choosing the window, while Section 5 applies our framework to re-examine the use of the RD design to estimate the incumbency advantage in the U.S. House. Section 6 summarizes and concludes. Extensions of our methodology to the fuzzy RD design are provided in Appendix 7. Additional empirical results are provided in a supplemental online appendix.

2 Example: Incumbency advantage in the U.S. House

Political scientists have long studied the question of whether the incumbent status of legislators translates into an electoral or *incumbency* advantage.\(^2\) This advantage is believed to stem from a variety of factors, including access to franking privileges, name recognition, the ability to secure federal funds for their districts, casework, and the ability to deter high-quality challengers. The hypothesis is concerned with the causal effect of incumbency on subsequent electoral performance. Providing a correct estimate of the incumbency advantage, however, is complicated by the fact that high-quality politicians tend to obtain higher vote shares than their low-quality counterparts, making them more likely to both become incumbents and obtain high vote shares in future elections. Any empirical strategy that ignores this potential selection problem might seriously overestimate the size of the incumbency advantage, mistakenly attributing the good electoral outcomes of incumbents to their incumbency status instead of to their inherent higher ability as politicians.

\(^2\)The incumbency advantage literature is vast. See, among many others, Erikson (1971), Gelman and King (1990), Cox and Katz (1996), and Ansolabehere and Jr. (2002).
Lee (2008) proposed to use the RD design to overcome this selection problem, taking advantage of the fact that the incumbency status of parties in a two-party system changes abruptly at the 50% cutoff. In districts where a party obtains 50% of the vote or more, this party becomes the incumbent, but in districts where the party’s vote share falls short of 50%, the party loses the election—and the opposing party becomes the incumbent. The idea is to exploit this discontinuity in the incumbency status of parties to estimate the incumbency advantage. The RD design compares U.S. House districts where the Democratic Party barely won a given election, referred to as election $t$, to districts where it barely lost, and computes the difference in the party’s vote shares between both groups in the following election, at $t + 1$.

The design is illustrated in Figure 1(a) for U.S. House elections between 1942 and 2008. The x-axis is the Democratic margin of victory in election $t$, defined as the Democratic vote share minus the vote share of the strongest opponent and ranging from -1 to 1. The y-axis is the outcome of interest, the vote share obtained by the Democratic party in election $t + 1$. The dots are binned means and the solid line is a 4th order polynomial fit on the full data. As expected, the relationship between the Democratic party’s vote share at $t$ and $t + 1$ is strongly positive. This suggests that incumbency advantage estimates that do not take into account the factors that affect election outcomes at both $t$ and $t + 1$ may lead to incorrect conclusions. The figure also shows a clear jump right at the cutoff: the Democratic party obtains much higher vote shares at $t + 1$ in districts where it barely won at $t$ than in districts where it barely lost at $t$. The difference is about ten percentage points.

A crucial assumption is required for this jump to be a valid estimate of the causal effect of incumbency at the cutoff: other than incumbency status, all observed and unobserved factors that affect vote shares at $t + 1$ change smoothly at the threshold. If, for example, bare-winner districts had systematically stronger party organizations than bare-loser districts, the RD estimate would be estimating the effect of incumbency plus stronger party networks, and would therefore overestimate

---

3 The RD design focuses on the incumbency advantage of parties. For a discussion of how the RD design can be used to recover incumbency advantage of legislators, see Erikson and Titiunik (2011).

4 The margin of victory is defined in this way because this is not a pure two-party system: in a number of districts small parties or write-in candidates obtain a small fraction of the vote, making the sum of Democratic and Republican vote shares slightly less than 1. Also, in the majority of cases, the strongest opponent is the Republican party, but there are exceptions: sometimes the Democratic party is unopposed, and in few cases the strongest opponent is a third party.
the incumbency advantage. But if local party organization only changed smoothly as a function of previous vote share around the cutoff, the RD effect could be entirely attributed to the change in incumbent status. This assumption cannot be tested directly because it involves statements about unobserved factors, but it may be justified when subjects cannot manipulate their scores precisely enough to fully determine their treatment status (Lee (2008)). For example, political parties can affect their vote shares with get-out-the-vote campaigns, TV ads, and town hall meetings; but if they lack precise control over the final total number of votes, there will still be an element of chance to which party ultimately wins close races. Thus, close enough to the threshold, winning or losing the election can be thought to be as-if randomly assigned. An important consequence of interpreting RDs as local experiments is that, in a neighborhood of the cutoff, the data can be analyzed as one would analyze a randomized experiment. In particular, close enough to the cutoff, the outcome and the score should be unrelated at either side of the cutoff, and pre-treatment characteristics on both sides of the cutoff should be similar to each other.

Figure 1: RD Design for different windows

**Democratic Vote Share at $t + 1$**

![Graphs showing RD design for different windows](image)

(a) Districts in [-1, 1] (b) Districts in [-0.30, 0.30] (c) Districts in [-0.03, 0.03]

Figure 1 shows that, at least at first glance, these data may be consistent with a local randomization condition. Figure 1(a) uses the full sample, Figure 1(b) employs the subsample of districts with margin of victory within 30 percentage points, and Figure 1(c) uses the subsample of districts with margin of victory within 3 percentage points. The last two figures are constructed analogously to Figure 1(a) – showing binned means and a 4th order polynomial fit. Although for the two largest
windows around the cutoff the outcome and the score are strongly associated on either side of the cutoff, Figure 1(c) shows that this association vanishes when considering districts with close elections: except for the 10 percentage point jump that occurs right at the cutoff, the plot of vote share against margin of victory is approximately a horizontal line.

This phenomenon is consistent with what would be expected from an experiment that randomly assigned the margin of victory in the [-0.03, 0.03] window. If treatment were truly randomly assigned in this window, pre-treatment covariates should also be balanced. However, Caughey and Sekhon (2011) recently showed that substantively important covariates are severely imbalanced in small neighborhoods around the cutoff, casting doubts on the validity of the RD design for this empirical application. In the upcoming sections, we use the interpretation of RD designs as local experiments as a starting point, and assume that there is a neighborhood around the cutoff in which a randomization-type condition actually holds. Using a randomization inference framework, we then explore the consequences of this assumption to analyze both the relationship between the score and the outcome, as well as the similarity of pre-treatment characteristics in treatment and control groups on both sides of the cutoff. In Section 5, we employ this framework to revisit the lack of balance present in this application, and show that conditioning on one of the most important covariates (Democratic victory at $t-1$) leads to an alternative design that appears to be consistent with the implications of our randomization-type condition. We also employ this (conditional) empirical strategy to obtain new estimates of the incumbency advantage in the U.S. House.

3 Randomization Inference in RD

We consider a setting with $n$ units, indexed $i = 1, 2, \ldots n$, where the scalar $R_i$ is the score observed for unit $i$, with the $n$-vector $R$ collecting the observations. In our application, $R_i$ is the Democratic margin of victory at $t$ and $i$ indexes congressional districts. We denote unit $i$’s potential outcome with $y_i(r)$, where $r$ is a given value of the vector of scores. In the randomization inference framework that we adopt, the potential outcome functions $y_i(r)$ are considered fixed characteristics of the finite population of $n$ units, and the observed vector of scores $R$ is a random variable. Thus, the
observed outcome for unit \( i \) is \( Y_i \equiv y_i(R) \), and is likewise a random variable with observations collected in the \( n \)-vector \( Y \). The essential feature of the RD design is embodied in a treatment variable \( Z_i = 1 (R_i \geq r_0) \), which is determined by the position of the score relative to the cutoff or threshold value \( r_0 \). The \( n \)-vector of treatment status indicators is denoted \( Z \), with \( Z_i = 1 \) if unit \( i \) receives treatment and \( Z_i = 0 \) otherwise. We focus on the so-called sharp RD design, where all units comply with their assigned treatment, but we extend our methodology to the so-called fuzzy design, where treatment status is not completely determined by the score, in the appendix.

Our approach begins by specifying conditions within a neighborhood of the threshold that allow us to analyze the RD design as a randomized experiment. Specifically, we focus on an interval or window \( W_0 = [\underline{r}, \overline{r}] \) on the support of the score, containing the threshold value \( r_0 \), where the assumptions described below hold. We denote the subvector of \( R \) corresponding to units with \( R_i \) inside this window as \( R_{W_0} \), and likewise for other vectors. In addition, we define \( F_{R_i | R_i \in W_0}(r) \) to be the conditional distribution function (cdf) of the score \( R_i \) given \( R_i \in W_0 \), for each unit \( i \).

**Assumption A1: Local Randomized Experiment.** There exists a neighborhood \( W_0 = [\underline{r}, \overline{r}] \) with \( \underline{r} < r_0 < \overline{r} \) such that for all \( i \) with \( R_i \in W_0 \),

a. \( F_{R_i | R_i \in W_0}(r) = F(r) \).

b. \( y_i(r) = y_i(z_{W_0}) \) for all \( r \).

The first part of A1 requires that the distributions of the scores are the same for all units inside \( W_0 \), implying that the scores can be considered “as good as randomly assigned” in this window. This is a strong assumption, and would be violated, for example, if the score were affected by the potential outcomes even near the threshold – but may be relaxed, for example, by explicitly modeling the relationship between \( R_i \) and potential outcomes. The second part of A1 requires that potential outcomes within the window depend on the score only through treatment indicators within the window. This implicitly makes two restrictions. First, it prevents potential outcomes of units inside \( W_0 \) from being affected by the scores of units outside (i.e., \( y_i(r) = y_i(r_{W_0}) \)). Second, for units in \( W_0 \), it requires that potential outcomes depend on the score only through the treatment indicators but not the particular value of the scores (i.e, \( y_i(r_{W_0}) = y_i(z_{W_0}) \)). This part of the assumption is
plausible in many settings where, for example, $R_i$ is primarily an input into a mechanical formula allocating assignment to the treatment $Z_i$.

The conditions in A1 are stronger than those typically required for identification and inference in the classical RD literature when treatment effects are assumed constant (Hahn, Todd, and van der Klaauw (2001, Theorem 1)). Instead of only assuming continuity of the relevant population functions at $r_0$ (e.g., conditional expectations, distribution functions), our assumption implies that, in the window $W_0$, these functions are not only continuous, but also completely unaffected by the score. As discussed above, this assumption is motivated by Lee (2008), who argued that if subjects have imprecise control over their scores, the discontinuity in the probability of treatment assignment generates conditions near the threshold $r_0$ that resemble those of a randomized experiment.

Assumption A1 has two main implications for our approach. First, it means that near the threshold we can ignore the score values for purposes of statistical inference and just focus on the treatment indicators $Z_{W_0}$. Second, since the distribution of $Z_{W_0}$ does not depend on potential outcomes, comparisons of observed outcomes across the threshold have a causal interpretation. In most settings, A1 is plausible only within a narrow window of the threshold, leaving only a small number of units for analysis. Thus, the problems of estimation and inference using this assumption in the context of RD are complicated by small-sample concerns.

Following Rosenbaum (2002, 2010), we propose to use exact randomization inference methods to overcome this potential small-sample problem. In the remainder of this section we assume A1 holds and take as given the window $W_0$, but we discuss empirical methods for choosing this window in Section 4.

Hypothesizing the randomization mechanism

The first task in applying randomization inference to the RD design is to choose a randomization mechanism for $Z_{W_0}$ that is assumed to describe the data generating process that places units on either side of the threshold. A natural starting place for a setting in which $Z_i$ is an individual-level variable (as opposed to a group-level characteristic) assumes $Z_i$ is a Bernoulli random variable with parameter $\pi$. In this case the probability distribution of $Z_{W_0}$ is given by $\Pr(Z_{W_0} = z) = \pi^z(1-\pi)^{(1-z)/1}$, for all vectors $z$ in $\Omega_{W_0}$, which in this case consists of the $2^{n_{W_0}}$ possible vectors
of zeros and ones, where \( n_{W_0} \) is the number of units in \( W_0 \) and \( 1 \) is a conformable vector of ones. This randomization distribution is fully determined up to the value \( \pi \), which is typically unknown in the context of RD applications. A natural choice for \( \pi \) would be \( \hat{\pi} = \frac{Z_{W_0}' 1}{n_{W_0}} \), the fraction of units within the window with scores exceeding the threshold, which corresponds to the maximum likelihood estimate of \( \pi \).

While the simplicity of this Bernoulli mechanism is attractive, a practical disadvantage is that it results in a positive probability of all units in the window being assigned to the same group. An alternative mechanism that avoids this problem, and is also likely to apply in settings where \( Z_i \) is an individual-level variable, is a random allocation rule or “fixed margins randomization” in which the number of units within the window assigned to treatment is fixed at \( m_{W_0} \). Under this mechanism, \( \Omega_{W_0} \) consists of the \( \binom{n_{W_0}}{m_{W_0}} \) possible \( n_{W_0} \)-vectors with \( m_{W_0} \) ones and \( n_{W_0} - m_{W_0} \) zeros. The probability distribution is therefore given by \( \Pr(Z_{W_0} = z) = \binom{n_{W_0}}{m_{W_0}}^{-1} \), for all \( z \in \Omega_{W_0} \).

Other more complicated settings include those where \( Z_i \) is a group-level variable or where additional variables are known to affect the probability of treatment. In such cases, mechanisms approximating a block-randomized or stratified design will be more appropriate, and are also straightforward to construct.

**Test of no effect**

Having chosen an appropriate randomization mechanism, we can test the sharp null hypothesis of no treatment effect under Assumption A1. No treatment effect means observed outcomes are fixed regardless of the realization of \( Z_{W_0} \). Under this null hypothesis, potential outcomes are not a function of treatment status inside \( W_0 \); that is, \( y_i(z) = y_i \) for all \( i \) within the window and for all \( z \in \Omega_{W_0} \), where \( y_i \) is a fixed scalar. The distribution of any test statistic \( T(Z_{W_0}, y_{W_0}) \) is known, since it depends only on the known distribution of \( Z_{W_0} \), and \( y_{W_0} \) is a fixed vector of observed responses. The test thus consists of computing a significance level for the observed value of the test statistic. The one-sided significance level is simply the sum of the probabilities of assignment vectors \( z \) leading to values of \( T(z, y_{W_0}) \) at least as large as the observed value \( \tilde{T} \), that is, \( \Pr(T(Z_{W_0}, y_{W_0}) \geq \tilde{T}) = \sum_{z \in \Omega_{W_0}} 1(T(z, y_{W_0}) \geq \tilde{T}) \cdot \Pr(Z_{W_0} = z) \), where \( \Pr(Z_{W_0} = z) \) is the chosen randomization mechanism.
Any test statistic may be used, including difference-in-means, the Kolmogorov-Smirnov test statistic, and difference-in-quantiles. While in typical cases the significance level of the test may be approximated when a large number of units is available, randomization-based inference remains valid (given A1) even for a small number of units. This feature is particularly important in the RD design where the number of units within \( W_0 \) is likely to be small.

Confidence intervals and point estimates

While the test of no treatment effect is often an important starting place, and appealing for the minimal assumptions it relies on, in most applications we would like to construct confidence intervals and point estimates of treatment effects. This requires additional assumptions. The next assumption we introduce is that of no interference between units.

Assumption A2: Local Stable Unit Treatment Value Assumption. For all \( i \) with \( R_i \in W_0 \), if \( z_i = \tilde{z}_i \) then \( y_i(z_{W_0}) = y_i(\tilde{z}_{W_0}) \).

This assumption means that unit \( i \)'s potential outcome depends only on \( z_i \), which, together with A1, allows us to write potential outcomes simply as \( y_i(z_i) \) for units in \( W_0 \).

Assumptions A1-A2 enable us to characterize the effects of treatment through inference on the distribution or quantiles of potential outcomes \( y_i(z_i) \). The goal is to construct a confidence interval \([a(q), b(q)]\) that covers with at least some specified probability the \( q \)-quantile of \( y_i(1) \), denoted \( Q^1(q) \), which is simply the \([q \times n_{W_0}]\)-th order statistic of \( y_i(1) \) for units within the window, and a similar confidence interval for \( Q^0(q) \). The confidence interval for \( Q^1(q) \) consists of the observed treated values \( x \) above the threshold (but in the window) such that the hypothesis \( H_0 : Q^1(q) = x \) is not rejected by a test of at most some specified size. The test statistic is \( J(x) = Z'_{W_0}1(Y_{W_0} \leq x) \), the number of units above the threshold whose outcomes are less than or equal to \( x \), and has distribution \( \Pr(J(x) = j) = \binom{[q \times n_{W_0}] - 1}{m_{W_0} - j} \binom{n_{W_0}}{m_{W_0} - 1} \) under a fixed margins randomization mechanism. Inference on the quantile treatment effect, or \( Q^1(q) - Q^0(q) \), can be based on confidence regions for \( Q^1(q) \) and \( Q^0(q) \).

Point estimates and potentially shorter confidence intervals for the treatment effect can be obtained at the cost of a parametric model for the treatment effect. A simple (albeit restrictive)
model that is commonly used is the constant treatment effect model below.

**Assumption A3: Local Constant Treatment Effect Model.** For all \( i \) with \( R_i \in W_0 \), \( y_i(1) = y_i(0) + \tau \), for some \( \tau \in \mathbb{R} \).

Under assumptions A1-A3, and hypothesizing a value \( \tau = \tau_0 \) for the treatment effect, the adjusted responses, \( Y_i - \tau_0 Z_i = y_i(0) \) are constant under alternative realizations of \( Z_{W_0} \). Thus, under this model, a test of the hypothesis \( \tau = \tau_0 \) proceeds exactly as the test of the sharp null discussed above, except that now the adjusted responses are used in place of the raw responses. The test statistic is therefore \( T(Z_{W_0}, Y_{W_0} - \tau_0 Z_{W_0}) \), and the significance level is computed as before. Confidence intervals for the treatment effect can be found by finding all values \( \tau_0 \) such that the test \( \tau = \tau_0 \) is not rejected, and Hodges-Lehman-type point estimates can also be constructed finding the value of \( \tau_0 \) such that the observed test statistic \( T(Z_{W_0}, Y_{W_0} - \tau_0 Z_{W_0}) \) equals its expectation under the null hypothesis.

### 4 Choosing the Window

If there exists a window \( W_0 = [r, \bar{r}] \) where our randomization-type condition A1 holds and this window is known, applying randomization inference procedures to the RD design is straightforward. In practice, however, this window will be unknown and must be chosen by the researcher. This constitutes the main methodological challenge of applying a randomization-inference approach to RD designs, and is analogous to the problem of bandwidth selection in conventional nonparametric RD approaches (Imbens and Lemieux 2008, Imbens and Kalyanaraman 2012).

In this section we discuss two main methods to select the window \( W_0 \) where A1 holds. The motivation behind each procedure is the same, although the implementation may be different in each case. The idea is to exploit the fact that Assumption A1 holds for the window \( W_0 \), but does not hold in general for a larger window containing \( W_0 \). This observation implies that for a window \( W \) larger than \( W_0 \), some observations will satisfy A1 (those units with \( R_i \in W_0 \)) while others will not (those units with \( R_i \in W - W_0 \)), leading to testable implications that can be exploited for window selection purposes. Each method implements this idea by considering a sequence of tests,
one for each window candidate, beginning with a large window and then sequentially shrinking it until the test fails to reject a violation of Assumption A1.

The first method considers the relationship between observed outcomes and the score for units inside and outside \( W_0 \), while the second method assumes the existence of pre-treatment covariates (for which the treatment effect is zero by construction) and performs “balance tests” to analyze whether the treatment has a significant effect. For these procedures to work, additional assumptions must hold for units outside \( W_0 \), since without these additional assumptions it is impossible to identify \( W_0 \) in general from the observed data. We explicitly describe a set of these additional assumptions in the context of the first method proposed, and provide an intuitive discussion of why these additional conditions are reasonable in empirical applications. For the second method, similar assumptions can be proposed, although we do not explicitly discuss them for brevity.

4.1 Methods based on outcomes

Under Assumption A1, inside the window \( W_0 \) the score \( R_i \) affects the units’ potential outcomes only through \( Z_i \), that is, only through placement above or below the cutoff. This implies that, conditional on \( Z_i \), the observed outcomes \( Y_i \) and the scores \( R_i \) are trivially independent variables for all \( i \) with \( R_i \in W_0 \). (When \( Z_i = z \) the random variable \( Y_i \) is degenerate and the distribution of \( R_i|R_i \in W_0 \) is not a function of \( i \) inside \( W_0 \).) For units with \( R_i \notin W_0 \), however, this need not be the case.

This observation motivates our first window selection procedure. The idea is to select a window \( W \) within which \( Y_W \) and \( R_W \) are not associated, conditionally on \( Z_W = z_W \). This leads to a valid window selector provided certain assumptions hold for units outside the window (\( R_i \notin W_0 \)). The procedure will work empirically if there is observable association between \( Y_i \) and \( R_i \) for units outside \( W_0 \), conditionally on \( Z_i = z \).

We formalize this requirement in the following assumption, which is stronger than needed, but justifies this window selection procedure in an intuitive way, as further discussed below. Since we need to work conditionally on \( Z_i = z \), we only discuss the assumption and resulting procedure for the treated units (\( Z_i = 1 \)) to save space; an analogous discussion applies for units with \( Z_i = 0 \). Define \( W^+ = [\bar{r}, \rho] \) for a \( \rho > \bar{r} \), and recall that \( r_0 \in W_0 = [\underline{r}, \bar{r}] \).
Assumption A4: Association Outside $W_0$. For all $i$ with $R_i \in W^+$ and for all $r \in W^+$,

a. $r_i = \tilde{r}_i$ implies $y_i(r) = y_i(\tilde{r})$.

b. At least one of the following holds:

b1. $y_i(r)$ is strictly increasing.

b2. $F_{R_i|R_i \in W^+}(r) = F(r; y_i(r), \rho)$ and, for all $j \neq k$,

$$y_j(r) > y_k(r) \Rightarrow F(r; y_j(r), \rho) < F(r; y_k(r), \rho).$$

Assumption A4a extends SUTVA to those units outside $W_0$, allowing us to write unit $i$’s potential outcome simply as $y_i(r)$. Assumption A4b is a key assumption leading to a valid window selector, which requires a form of non-random selection among units outside $W_0$, leading to an observable association between the vectors $Y_{W^+}$ and $R_{W^+}$, and therefore between the vectors $Y_W$ and $R_W$ for any window $W$ containing $W_0$.

The first part of A4b explicitly requires that the function $y_i(r)$ not be a constant function on the full support of $R_i$, leading to a (positive) association between the observed outcome and score for those treated units with $R_i \notin W_0$. This assumption is naturally justified whenever the score has a direct effect on the potential outcomes, as one might suspect is the case in our incumbency advantage application. As shown in Figures 1(a), 1(b) and 1(c), except for a narrow window around the cutoff, higher values of the Democratic margin of victory at $t$ are associated with higher Democratic vote shares at $t + 1$, suggesting that A4.b1 holds. This association may occur either because of a direct effect of $r$ on $y_i(r)$ (high vote shares at $t$ lead to increased campaign contributions which in turn lead to higher vote shares at $t + 1$) or because $r$ might be a marker for underlying factors that affect $y_i(r)$ (more skilled politicians obtain higher vote shares at $t$ and $t + 1$). In other empirical applications, however, units may sort based on observed or unobserved heterogeneity even when potential outcomes are constant for all values of $r$. The second part of Assumption A4b allows for this possibility, permitting units with high $R_i$ to have high observed outcome, even when the potential outcome satisfies $y_i(r) = y_i$ for all $i$.

Assumptions A1, A2 and A4 justify a simple procedure to find $W_0$. This procedure finds the widest window for which the observed outcomes and scores are not associated inside this window,
but associated outside of it. The chosen window is the widest window such that an appropriate testing procedure for association between scores and outcomes, conditional on placement above or below the threshold, does not reject the null hypothesis. Such a procedure can be implemented empirically in different ways. A simple approach is to begin by considering all the treated observations (i.e., choosing the largest possible upper end for $W_0$), conduct a test of association between $Y_i$ and $R_i$ for these observations and, if the null hypothesis is rejected, continue by decreasing the upper end until the resulting test fails to reject the null hypothesis of no association. This procedure can be implemented as follows.

**Procedure 1: One-sided Window Selection.** Select a distribution-free testing procedure for association of two random vectors, and denote the corresponding test statistic by $T(Y, R)$.

Let $i = 1, \ldots, n_1$ index the observations with $Z_i = 1$, and $R_{[j]}$ be the $j$-th order statistic of $R$ in this subsample.

Step 1: Define $W(j) = [r_0, R_{[j]}]$, and set $j = n_1$. Choose a minimum value for $j$, denoted $j_{\text{min}}$, which sets the minimum number of observations in $W(j)$.

Step 2: Test for association between $Y$ and $R$ for units in $W(j)$ using $T(Y_{W(j)}, R_{W(j)})$.

Step 3: If the null hypothesis is rejected, decrease $j$. If $j \geq j_{\text{min}}$ go back to Step 2, else stop and conclude that an upper end for $W_0$ cannot be selected.

If the null hypothesis is not rejected, keep $R_{[j]}$ as the upper end of the selected window.

This procedure selects the upper end of $W_0$, and can be applied symmetrically to observations with $Z_i = 0$ to select the lower end of the window. Several practical issues need to be considered to implement this procedure. First, a significance level $\alpha$ for the tests needs to be selected; as a practical issue, this value should be chosen to be higher than conventional levels to ensure that the null hypothesis of no association fails to be rejected only when the chance of its being false is considerably low. This will likely lead to an underestimation of the (absolute value of) the ends of the window, and thus result in a conservative choice for $W_0$. Second, the procedure requires multiple testing, a phenomenon that leads to rejecting the null hypothesis too often and is often interpreted as leading to a test that is too liberal. In this case, however, rejecting too often leads
to a more conservative procedure, since it will cause the procedure to shrink the window even further, making it more likely to select a window that is contained in $W_0$. Third, because of power considerations, in applications it may be desirable to further shrink the selected window to avoid cases where there are very few observations outside $W_0$ and the test fails to detect the association between outcomes and scores. Finally, the procedure also requires selecting the minimum number of observations $j_{\text{min}}$ required for the testing procedure to be valid.

This procedure splits the window selection problem into two different problems, one for control units and the other for treated units. This is necessary to remove the association between scores and outcomes when the treatment effect is non-zero, but it reduces the effective sample size and hence the power of the test employed. A refinement of this method, which employs the full sample, assumes a constant treatment effect model for all observations, and instead of conditioning on $Z_i$ to test for association, conducts tests of association jointly for treated and control observations using adjusted observed outcomes. The idea is to exploit the fact that under a constant treatment effect model it is possible to adjust the observed responses so that the lack of association between outcomes and running variables holds (only) for all units within the true window $W_0$ if and only if the hypothesized treatment effect equals the true treatment effect. To conserve space, we provide further details of this refined window selector in the supplemental online appendix.

4.2 Method based on pre-treatment covariates

In most empirical applications of the RD design, researchers have access to pre-treatment covariates in addition to the outcome and score variables. Whenever these covariates are available, and under appropriate assumptions, they can also be employed to select a window where Assumption A1 holds.

One possible approach is to apply the same methodology outlined above, but replacing potential outcomes by pre-treatment covariates, which we denote by $x_i(r)$. In this approach the idea is to treat $x_i(r)$ as potential outcomes, although in this case the treatment effect is zero by construction. Under appropriate assumptions such as those introduced above, the procedures described in the previous section will lead to a valid window selection procedure. This idea requires the observed pre-treatment covariates to be independent of the observed scores within the true window $W_0$, but
Another approach to exploit the existence of pre-determined covariates for window selection purposes is to directly carry out a randomization test of no effect, which is akin to tests of covariate balance commonly done by researchers in classical RD settings to provide empirical evidence of continuity at the cutoff. This alternative method is based on the idea that for units with $R_i \in W_0$ the treatment assignment vector $Z_{W_0}$ has no effect on a covariate vector $x_{W_0}$ by construction, since the latter vector is fixed prior to the treatment (i.e., predetermined). Under Assumptions A1-A2 the size of the test of no effect is known, and therefore we can control the probability with which we accept a window where the assumptions hold. In addition, under appropriate assumptions such as those discussed above, this procedure will also be able to detect the true window $W_0$. This procedure could be justified, for example, when outside $W_0$, (i) there is explicit sorting into treatment based on $x_i(r)$, or (ii) selection into treatment is a function of potential outcomes and potential outcomes in turn depend non-trivially on $x_i(r)$. A wide range of observational settings display this kind of non-trivial selection on observable characteristics.

**Procedure 2: Window Selection with Pre-determined Covariates.** Select a test statistic of interest, denoted $T(X, R)$, where $X$ is the $n$-vector containing the observed pre-treatment covariates $X_i = x_i(R)$. Let $R_{[j]}$ be the $j$-th order statistic of $R$ in the sample of all observations indexed by $i = 1, \ldots, n$.

Step 1: Define $W(j_0, j_1) = [R_{[j_0]}, R_{[j_1]}]$, and set $j_0 = 1, j_1 = n$. Choose minimum values $j_{0,min}$ and $j_{1,min}$ satisfying $j_{0,min} < r_0 < j_{1,min}$, which set the minimum number of observations required in $W(j_{0,min}, j_{1,min})$.

Step 2: Conduct a test of no effect using $T(X_{W(j_1,j_2)}, R_{W(j_0,j_1)})$.

Step 3: If the null hypothesis is rejected, increase $j_0$ and decrease $j_1$. If $j_0 < j_{0,min}$ and $j_{1,min} < j_1$ go back to Step 2, else stop and conclude that lower and upper ends for $W_0$ cannot be selected.

If the null hypothesis is not rejected, keep $R_{[j_0]}$ and $R_{[j_1]}$ as the ends of the selected window.
As above, implementing this procedure requires choosing a sequence of candidate windows, and a rejection rule. We suggest symmetric windows around $r_0$, so that choosing the window reduces to choosing its length. A rejection rule consists of a significance level $\alpha$, such that when the test gives a p-value of less than $\alpha$, we reject that randomization holds within the window. Our application below illustrates how these procedures work in practice.

5 Revisiting RD Estimates of Incumbency Advantage

Note for Brigham: this section will be updated when the final runs are ready (they are running on a cluster). As a consequence, the section is preliminary and outdated.

We now use our to re-analyze the RD based evidence on the incumbency advantage in the U.S. House of Representatives. We first focus on choosing the windows where randomization plausibly holds, and then provide estimates within these windows. Taken together, the window selection methods discussed above show that local randomization seems to fail in the overall dataset, a conclusion that is consistent with the evidence and arguments presented in Caughey and Sekhon (2011).

Figure 2 shows the results of the window selector based on the association between outcome and score. A Kendall test of association between the Democratic vote share at $t+1$ (the outcome) and the Democratic margin of victory at $t$ (the score) is used separately to the right and left of the cutoff. For a given positive value $r_1$ of the score, the figure shows the p-value corresponding to the Kendall test of association between the score and the outcome for observations whose score is above zero and below $r_1$ —and above $r_1$ and below zero when $r_1$ is negative. Using a cutoff p-value of 0.2 leads to a clear choice for the right limit of $W_0$, since once the p-values increase above 0.2, they do not fall below this cutoff for smaller windows. As we get closer to the cutoff from the left, p-values also start to increase dramatically although when we get very close to the cutoff p-values do fall below 0.2 again. Nonetheless, the figure shows very clearly that the association between outcome and score tends to disappear for values of the margin of victory approximately between -0.04 and 0.04.

But when we consider window choices based on the covariate balance selection method, we
arrive to a very different conclusion. In Figure 3, we choose $W_0$ using two pre-treatment covariates: Democratic vote share at $t-1$, and Democratic victory at $t-1$ – an indicator equal to 1 if the Democratic Party won election $t-1$. In the lower part of Figure 3(a) and Figure 3(b), for a given positive value $r_1$ of the score, each figure shows the randomization-inference based p-value (left y-axis) corresponding to a balance test between treated observations with score between zero and $r_1$, and control observations with score above $-r_1$ and below zero. All balance tests in these and subsequent figures are based on the Kolmogorov-Smirnov test-statistic, and assume fixed margins to compute randomization-inference p-values.\footnote{Figure 3 plots the p-values for the entire range of the score, $[-1,1]$, for easy analogy with Figure 2, but note that both plots in Figure 3 are exactly symmetric around the cutoff: for any value $r_1$ of the score, $r_1$ and $-r_1$ are associated with the same p-value, so that each plot’s right half is a mirror image of its left half.} The upper part of each figure plots the outcome of interest in each case (right y-axis) against the Democratic margin of victory at $t$. The dots are
means in bins of 0.05 percentage points and the straight line is a 4th order polynomial fit, as in Figure 1(a) in Section 2.

The contrast between Figure 2 and Figure 3 is striking. Figure 3(a) shows that, even among districts with extremely close elections at \( t \), there is considerable imbalance on the vote share obtained by the Democratic Party at \( t - 1 \) between districts where the Democratic party barely won and barely lost at \( t \). Similarly, Figure 3(b) shows that there is severe imbalance on whether the Democratic party won the previous election, at \( t - 1 \), between barely-winner and barely-loser districts, and this imbalance does not improve even for the smallest window around the cutoff, which has just 5 observations on either side. This suggests that there is sorting around the cutoff, as argued by Caughey and Sekhon (2011): the Democratic party wins close elections at \( t \) disproportionately more often in districts where it also won at \( t - 1 \), that is, in districts where it already was the incumbent party.

Figure 3: Covariate balance window selector for all districts

![Graph showing covariate balance window selector for all districts](attachment:image.png)

(a) Democratic Vote at \( t - 1 \)
(b) Democratic Victory at \( t - 1 \)

To address this possible selection problem, we analyze two sets of districts separately: districts where the Democratic Party won election \( t - 1 \), and districts where the Democratic Party lost this election. Figure 4 shows what happens when we apply our balance-based window selection
procedure to both subsets. Both figures in Figure 4 are based on covariate balance on Democratic vote share at \( t - 1 \), and are therefore analogous to Figure 3(a).

Figure 4: Covariate balance window selector separately by Democratic victory at \( t-1 \)

(a) Districts where Democratic Party won at \( t - 1 \)  
(b) Districts where Democratic Party Lost at \( t - 1 \)

The massive imbalance on Democratic vote share at \( t - 1 \) that we saw in Figure 3(a) largely dissappears when we considers both subsets separately. As shown in Figure 4(a), when we consider only districts where the Democratic party won the \( t - 1 \) election, for any window contained in the range \([-0.029, 0.029]\), p-values are always above 0.2, with the only exception of the window before last, where the p-value is 0.19. Similarly, Figure 4(b) shows that balance on Democratic vote share at \( t - 1 \) also improves vastly when barely-loser districts are considered separately. For windows contained in the range \([0.0031, 0.0031]\), all p-values are above 0.2, with no exceptions.

Figure 5 shows that the same phenomenon is observed for other important pre-treatment covariates. The figure presents plots analogous to those in Figure ?? for four covariates: campaign spending of the Democratic candidate, Congressional Quarterly (CQ) race predictions\(^6\), the Democratic incumbent’s ideological position as measured by her DW-NOMINATE score\(^7\), and the number

---

\(^6\)This variable is equal to 1, 0, or -1 if CQ favors the Democratic party, neither party, or the Republican party, respectively.

\(^7\)We use the first dimension of the DW-NOMINATE score, which is commonly used in the U.S. House as a measure of the legislator’s position on the liberal-conservative continuum. See Poole and Rosenthal (1997)
of previous terms served by the current Democratic candidate.
Figure 5: Covariate balance window selector for additional covariates

- Full Sample IncDWNOM1
- Democratic victory t−1 Subsample IncDWNOM1
- Democratic loss t−1 Subsample IncDWNOM1

- Full Sample PrvTrmsD
- Democratic victory t−1 Subsample PrvTrmsD
- Democratic loss t−1 Subsample PrvTrmsD

- Full Sample CQRating3
- Democratic victory t−1 Subsample CQRating3
- Democratic loss t−1 Subsample CQRating3

- Full Sample DSnpdPct
- Democratic victory t−1 Subsample DSnpdPct
- Democratic loss t−1 Subsample DSnpdPct
The evidence indicates that local randomization assumption is unlikely to hold in the entire dataset, but may be plausible once we condition on Democratic party victory at \( t - 1 \). In Table 1, we estimate the incumbency advantage using our method, and contrast the results to those obtained with two classical approaches: a 4th-order parametric fit as in Lee (2008), and a non-parametric local linear regression as suggested by Imbens and Lemieux (2008) with triangular kernel and optimal bandwidth selected with the method in Imbens and Kalyanaraman (2012). The first column presents results for the full sample, the second presents results for the subsets of districts where the Democratic party wins election \( t - 1 \), and the third for the subset of districts where the Democratic party loses election \( t - 1 \). In the full sample, our randomization-based window selectors indicate that there is no window for which a local randomization assumption is plausible. Nonetheless, for comparison with classical methods, we present randomization-inference results for the ad-hoc window \([-0.03, 0.03]\). For the subsets based on Democratic victory at \( t - 1 \), we perform randomization inference in the windows selected by our method and illustrated in Figure 4. As mentioned above, these windows are \([-0.029, 0.029]\) and \([0.0031, 0.0031]\) for districts with Democratic victory and loss at \( t \), respectively.

6 Summary and Conclusion

Note for Brigham: we will work on the conclusion at the very end...

Motivated by the analogy between regression discontinuity and randomized experiments, we described a set of conditions allowing the RD design to be analyzed as a randomized experiment and proposed applying the randomization inference framework to the RD design. While for settings with a large number of units randomization inference coincides with traditional techniques, our approach has potential to be useful when the number of units close to the discontinuity threshold is small and traditional methods break down. Thus our results can be viewed as an alternative to the traditional large-sample approximations. Indeed, one way to gauge the accuracy of traditional large-sample approximations would be to compare traditional results with those obtained using our approach. The randomization inference approach to inference also motivated our methodology for choosing a window around the RD threshold, complementing existing bandwidth selection algorithms.
The empirical example based on incumbency advantage in U.S. Congressional elections illustrated the methodology and showed our approach compares favorably to traditional methods of RD analysis.

7 Appendix: Fuzzy RD

In the sharp RD design, treatment status is equal to \( Z_i = 1 (R_i \geq r_0) \), an indicator for exceeding the threshold \( r_0 \), but in the fuzzy design, treatment status \( D_i \) (with observations collected in \( n \)-vector \( D \)), is not completely determined by placement relative to \( r_0 \), so \( D_i \) may differ from \( Z_i \).

Our framework applies directly to the fuzzy RD design and continues to make a randomization-type assumption for \( Z_{W_0} \), but it allows \( D \) to be nonrandom even very close to the threshold. Since treatment status is no longer determined by the score, we generalize our potential outcomes framework as follows. We let \( d_i (r) \) be unit \( i \)'s potential treatment status when the vector of scores is \( R = r \). Similarly, we let \( y_i (r, d) \) be unit \( i \)'s potential outcome when the vector of scores is \( R = r \) and the treatment status vector is \( D = d \). Observed treatment status and outcomes are \( D_i = d_i (R) \) and \( Y_i = y_i (R, D) \).

Assumption A1′: Local Randomized Experiment. There exists a neighborhood \( W_0 = [r, \bar{r}] \) with \( r < r_0 < \bar{r} \) such that for all \( i \) with \( R_i \in W_0 \),

a. \( F_{R_i|R_i \in W_0} (r) = F(r) \).

b. \( d_i (r) = d_i (z_{W_0}) \) and \( y_i (r, d) = y_i (z_{W_0}, d_{W_0}) \) for all \( r, d \).

Assumption A1′ means the same thing as its counterpart in the sharp RD setup, namely, sufficiently close to the threshold the running variable is ignorable and we can treat \( Z_{W_0} \) as randomly assigned. Just as before, this assumption permits testing the null hypothesis of no effect, which is performed exactly as described in Section 3. The interpretation of the test differs, however: it can only be considered a test of no effect of treatment among those units whose potential treatment status \( d_i (z_{W_0}) \) varies with \( z_{W_0} \).

Constructing confidence intervals and point estimates in the fuzzy design requires generalizing assumption A2 and introducing two additional assumptions: a local exclusion restriction which
restricts \( y_i(z, d) \) and a local monotonicity assumption which restricts \( d_i(z) \).

**Assumption A2′: Local Stable Unit Treatment Value Assumption (LSUTVA).** For all \( i \) with \( R_i \in W_0 \),

a. If \( z_i = \tilde{z}_i \), then \( d_i(z_{W_0}) = d_i(\tilde{z}_{W_0}) \).

b. If \( z_i = \tilde{z}_i \) and \( d_i = \tilde{d}_i \), then \( y_i(z_{W_0}, d_{W_0}) = y_i(\tilde{z}_{W_0}, \tilde{d}_{W_0}) \).

Assumption A2′ means unit \( i \)'s treatment status and outcome depends only upon unit \( i \)'s own placement relative to the threshold and treatment status.

**Assumption A6: Local Exclusion Restriction.** For all \( i \) with \( R_i \in W_0 \), \( y_i(z, d) = y_i(\tilde{z}, d) \) for all \((z, \tilde{z})\) and for all \( d \).

Assumption A6 means potential responses depend on placement with respect to the threshold only through its effect on treatment status; there is no direct effect of placement with respect to the threshold. Under assumptions A1′-A2′ and A4, we can write potential responses within the window in terms of \( d_i \) only: \( y_i(z, d) = y_i(d_i) \).

**Assumption A7: Local Monotonicity.** \( d_i(1) \geq d_i(0) \) for all \( i \) with \( R_i \in W_0 \) and \( d_i(1) > d_i(0) \) for some \( i \) with \( R_i \in W_0 \).

Assumption A7 means placement with respect to the threshold may affect treatment status for some units and not others, but where it does have an effect, it is only in one direction. In other words, crossing the threshold never induces a unit not to take the treatment. This monotonicity assumption allows us to define a subset of the \( n_{W_0} \) units who would take the treatment if assigned above the threshold \( (Z_i = 1) \) but would not if assigned below the threshold \( (Z_i = 0) \). This group, usually called “compliers,” is the group for whom inference on the effects of treatment \( D_i \) is possible without stronger assumptions.

Assumptions A1′-A2′ and A6-A7 allow inference on the quantiles of potential outcomes among compliers within the window, as in Section 3 (see Frandsen 2010). Under the constant treatment effect model in A3, estimation and inference proceeds exactly as before. Under A1′-A2′, A3, and A6 (A7 is not necessary given A3) the hypothesis \( \tau = \tau_0 \) is tested as before, but now defining the adjusted responses as \( Y_{W_0} - \tau_0 D_{W_0} \).
References


Table 1: Incumbency advantage in the U.S. House using an RD design

<table>
<thead>
<tr>
<th></th>
<th>Full sample</th>
<th>Dem won $t$</th>
<th>Dem lost $t$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standard Approaches</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A. Parametric model: 4th-order global polynomial</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td>0.102</td>
<td>0.095</td>
<td>0.114</td>
</tr>
<tr>
<td>p-value</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>CI</td>
<td>[0.084, 0.121]</td>
<td>[0.062, 0.127]</td>
<td>[0.09, 0.138]</td>
</tr>
<tr>
<td>Sample size treated</td>
<td>8003</td>
<td>4596</td>
<td>3407</td>
</tr>
<tr>
<td>Sample size control</td>
<td>8003</td>
<td>4596</td>
<td>3407</td>
</tr>
<tr>
<td>B. Non-parametric model: local linear regression</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Estimate</td>
<td>0.093</td>
<td>0.095</td>
<td>0.098</td>
</tr>
<tr>
<td>p-value</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>CI</td>
<td>[0.080, 0.106]</td>
<td>[0.073, 0.116]</td>
<td>[0.079, 0.117]</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>0.347</td>
<td>0.404</td>
<td>0.196</td>
</tr>
<tr>
<td>Sample size treated</td>
<td>8003</td>
<td>4596</td>
<td>3407</td>
</tr>
<tr>
<td>Sample size control</td>
<td>8003</td>
<td>4596</td>
<td>3407</td>
</tr>
<tr>
<td><strong>Randomization-based Approach</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hodges-Lehmann Estimate</td>
<td>0.084</td>
<td>0.074</td>
<td>0.070</td>
</tr>
<tr>
<td>Difference-in-means</td>
<td>0.100</td>
<td>0.094</td>
<td>0.073</td>
</tr>
<tr>
<td>Sharp-null p-value</td>
<td>0.000</td>
<td>0.000</td>
<td>0.009</td>
</tr>
<tr>
<td>CI for constant effect</td>
<td>[0.080, 0.121]</td>
<td>[0.059, 0.129]</td>
<td>[0.015, 0.132]</td>
</tr>
<tr>
<td>CI for 25th quantile effect</td>
<td>[0.055, 0.105]</td>
<td>[0.034, 0.100]</td>
<td>[-0.061, 0.154]</td>
</tr>
<tr>
<td>CI for 50th quantile effect</td>
<td>[0.067, 0.092]</td>
<td>[0.045, 0.093]</td>
<td>[-0.028, 0.164]</td>
</tr>
<tr>
<td>CI for 75th quantile effect</td>
<td>[0.060, 0.110]</td>
<td>[0.041, 0.107]</td>
<td>[-0.022, 0.169]</td>
</tr>
<tr>
<td>Window</td>
<td>[-0.030, 0.030]</td>
<td>[-0.029, 0.029]</td>
<td>[0.003, 0.003]</td>
</tr>
<tr>
<td>Sample size treated</td>
<td>210</td>
<td>98</td>
<td>13</td>
</tr>
<tr>
<td>Sample size control</td>
<td>240</td>
<td>86</td>
<td>20</td>
</tr>
</tbody>
</table>

Note: Outcome variable is Democratic vote share at $t + 1$. Parametric model is 4th order polynomial fit using all observations, and non-parametric model is local linear regression with triangular kernel using mean-square-error optimal bandwidth choice (Imbens and Kalyanaraman 2012). In randomization-based approach, p-value is based on difference-in-means test-statistic, and confidence intervals are obtained inverting the test under a constant treatment effect model.
Supplemental Appendix (intended for online publication only)

7.1 Window Selector Under Constant Treatment Effect Model

In this section, we briefly describe the window selection procedure based on a constant model of treatment effects. The procedure is based on the assumption below.

**Assumption A5: Constant Treatment Effect Model.** \( y_i(r_i, 1) = y_i(r_i, 0) + Z_i\tau_0 \) for all \( i \).

We define the adjusted responses \( \tilde{Y}_i(\tau) = Y_i - Z_i\tau \), which lead to \( \tilde{Y}_i(\tau_0) = y_i(r_i, 0) \) for all units by Assumption A5, and to \( \tilde{Y}_i(\tau_0) = y_i(0) \) for units with scores inside \( W_0 \) by Assumption A1. Therefore, assuming A1–A5 and a corresponding analogue of A4 for the control group, the adjusted responses \( \tilde{Y}_i(\tau) \) are constant and hence trivially independent of \( Z_i \) and hence of the score \( R_i \) in the true window \( W_0 \) if and only if \( \tau = \tau_0 \), which implies that in any window strictly containing \( W_0 \), the running variable \( R_i \) and \( \tilde{Y}_i(\tau) \) will be associated for any \( \tau \), even \( \tau_0 \).

This observation motivates a procedure to select \( W_0 \), which also obtains an estimate of \( \tau_0 \) as a by-product. The procedure is to find the value of \( \tau \) and the largest possible window such that adjusted responses and the running variables are not associated. This method can be implemented as follows.

**Procedure 2: Window Selection with Constant Treatment Effect.** Select a testing procedure for association of two random variables, and denote the corresponding test statistic by \( T(Y, R) \). Let \( R_{[j]} \) be the \( j \)-th order statistic of \( R \) in the sample of all observations indexed by \( i = 1, \ldots, n \).

Step 1: Construct a grid of possible values for the treatment effect \( \{\tau_k : k = 1, \ldots, K\} \), which is assumed to include \( \tau_0 \), and set \( k = 1 \). Define \( W(j_0, j_1) = [R_{[j_0]}, R_{[j_1]}] \), and set \( j_0 = 1 \), \( j_1 = n \). Choose minimum values \( j_{0,\min} \) and \( j_{1,\min} \) satisfying \( j_{0,\min} < j_0 < j_{1,\min} \), which set the minimum number of observations required in \( W(j_{0,\min}, j_{1,\min}) \).

Step 2: Construct the adjusted outcomes \( \tilde{Y}_i(\tau_k) \).
Step 3: Conduct a test of association between $\tilde{Y}_i(\tau_k)$ and $R$ for units in $W(j_1, j_2)$ using

$$T(\tilde{Y}_{W(j_0,j_1)}(\tau_k), R_{W(j_0,j_1)})$$

Step 4: If the null hypothesis is rejected, increase $j_0$ and decrease $j_1$. If $j_0 < j_{0,min}$ and $j_{1,min} < j_1$ go back to Step 3, else stop and conclude that lower and upper ends for $W_0$ cannot be selected for this value of $\tau_k$, increase $k$, and go back to Step 2.

If the null hypothesis is not rejected, keep $R_{[j_0]}$ and $R_{[j_1]}$ as the ends of the selected window (and keep $\tau_k$ as the estimated treatment effect).

The resulting procedure picks the largest the window for which the association between adjusted outcomes and scores is not rejected for a given $\tau$. Since the only value of $\tau$ for which the null of no association holds is $\tau_0$ and the only window where this null holds is $W_0$, the procedure can be used to estimate $W_0$ and $\tau_0$ jointly. In addition to the practical issues mentioned in Section 4, the implementation of this method requires researcher to establish a procedure to “search” for different windows. A simplification we propose is to consider only windows symmetric around $r_0$, which reduces the choice of the window to the choice of its length, $\ell$. With this simplification, the windows considered by the algorithm are perfectly nested. The implementation then requires to define a grid of possible values for $\tau$ and $\ell$, select the maximum length $\ell^*(\tau)$ where the null hypothesis of no association is not not rejected for every $\tau$, and then select the $\tau^*$ with the largest $\ell^*(\tau)$.

### 7.2 Additional Empirical Evidence for Covariate-Balance Window Selector

This section presents graphs analogous to those in Figures 3(a) and 3(b) in the paper for a long list of pre-treatment covariates. For a given positive value $r_1$ of the score, each figure shows the randomization-inference based p-value corresponding to a balance test between treated observations with score between zero and $r_1$, and control observations with score above $-r_1$ and below zero. For non-binary variables, each row in each figure plots p-values from randomization-inference balance tests based, respectively, on (i) difference-in-means test statistic and fixed margins randomization, (ii) difference-in-means test statistic and binomial randomization, (iii) Kolmogorov-Smirnov test-
statistic test statistic and fixed margins randomization, and (iv) Kolmogorov-Smirnov test-statistic test statistic and binomial randomization. As mentioned in the main body of the paper, all figures plot the p-values for the entire range of the score, $[-1, 1]$, for easy analogy with Figure 2 in the paper, but all plots are exactly symmetric around the cutoff: for any value $r_1$ of the score, $r_1$ and $-r_1$ are associated with the same p-value, so that each plot’s right half is a mirror image of its left half.
Figure 6: Window selector based on Incumbent’s DW Nominate Score (1st dimension)
Figure 7: Window selector based on number of Democratic candidate’s previous terms in office
Figure 8: Window selector based on number of Republican candidate’s previous terms in office
Figure 9: Window selector based on Republican experience adv
Figure 10: Window selector based on Democratic experience adv
Figure 11: Window selector based on partisan swing
Figure 12: Window selector based on CQ Rating -1,0,1
Figure 13: Window selector based on Democratic spending percentage
Figure 14: Window selector based on Democratic donations percentage
Figure 15: Window selector based on Democratic secretary of state

Full Sample
Diff means – Fixed margins

Democratic victory t−1 Subsample
Diff means – Fixed margins

Democratic loss t−1 Subsample
Diff means – Fixed margins

Full Sample
Diff means – Bernoulli

Democratic victory t−1 Subsample
Diff means – Bernoulli

Democratic loss t−1 Subsample
Diff means – Bernoulli

P-value

Cutoff

Democratic margin of victory t

Democratic margin of victory t

Democratic margin of victory t

Democratic margin of victory t
Figure 16: Window selector based on Democratic governor

Full Sample
Diff means – Fixed margins

Democratic victory t–1 Subsample
Diff means – Fixed margins

Democratic loss t–1 Subsample
Diff means – Fixed margins

Full Sample
Diff means – Bernoulli

Democratic victory t–1 Subsample
Diff means – Bernoulli

Democratic loss t–1 Subsample
Diff means – Bernoulli
Figure 17: Window selector based on Democratic presidential percentage margin
Figure 18: Window selector based on open seat
Figure 19: Window selector based on voter turnout percentage
Figure 20: Window selector based on government workers percentage
Figure 21: Window selector based on urban population percentage
Figure 22: Window selector based on black population percentage
Figure 23: Window selector based on foreign population percentage