Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate*

Matias D. Cattaneo† Brigham Frandsen‡ Rocío Titiunik§

February 28, 2014

Abstract

In the Regression Discontinuity (RD) design, units are assigned a treatment based on whether their value of an observed covariate is above or below a fixed cutoff. Under the assumption that the distribution of potential confounders changes continuously around the cutoff, the discontinuous jump in the probability of treatment assignment can be used to identify the treatment effect. Although a recent strand of the RD literature advocates interpreting this design as a local randomized experiment, the standard approach to estimation and inference is based solely on continuity assumptions that do not justify this interpretation. In this article, we provide precise conditions in a randomization-inference context under which this interpretation is directly justified, and develop exact finite-sample inference procedures based on them. Our randomization-inference framework is motivated by the observation that only a few observations might be available close enough to the threshold where local randomization is plausible, and hence standard large-sample procedures may be suspect. Our proposed methodology is intended as a complement and a robustness check to standard RD inference approaches. We illustrate our framework with a study of two measures of party-level advantage in U.S. Senate elections, where the number of close races is small and our framework is well suited for the empirical analysis.

Keywords: Regression discontinuity, randomization inference, exact inference, as-if randomization, local experiments, incumbency advantage, U.S. Senate

*We thank the associate editor, Kosuke Imai, three anonymous referees, Peter Aronow, Jake Bowers, Devin Caughey, Andrew Feher, Don Green, Luke Keele, Jasjeet Sekhon, and participants at the 2010 Political Methodology Meeting in the University of Iowa and at the 2012 Political Methodology Seminar in Princeton University for valuable comments and suggestions. Previous versions of this manuscript circulated under the titles "Randomization Inference in the Regression Discontinuity Design" and "Randomization Inference in the Regression Discontinuity Design to Study the Incumbency Advantage in the U.S. Senate" (first draft: July, 2010). The first and third authors gratefully acknowledge financial support from the National Science Foundation.

†Department of Economics, University of Michigan.

‡Department of Economics, Brigham Young University.

§Department of Political Science, University of Michigan. Corresponding author: 5700 Haven Hall, 505 South State St, Ann Arbor, MI 48109; titiunik@umich.edu.

1 Introduction

Inference on the causal effects of a treatment is one of the basic aims of empirical research. 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 compromise 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. The key feature of the design is that the probability of receiving the treatment conditional on the score jumps discontinuously at the cutoff, inducing variation in treatment assignment that is assumed to be unrelated to potential confounders. Recent reviews, including comprehensive lists of empirical examples, are given in Cook (2008), Imbens and Lemieux (2008) and Lee and Lemieux (2010).

The traditional inference approach in the RD design relies on flexible extrapolation (usually nonparametric curve estimation techniques such as local polynomial regression) using observations near the known cutoff. This approach follows the work of Hahn, Todd, and van der Klaauw (2001), who showed that, when placement relative to the cutoff completely determines treatment assignment, the key identifying assumption is that the conditional expectation of a potential outcome is continuous at the threshold. Intuitively, since nothing changes abruptly at the threshold other than the probability of receiving treatment, any jump in the conditional expectation of the outcome variable at the threshold is attributed to the effects of the treatment. Modern RD analysis employs local nonparametric curve estimation at either side of the threshold to estimate RD treatment effects, with local-linear regression being the preferred choice in most cases. See Porter (2003), Imbens and Kalyanaraman (2012) and Calonico, Cattaneo, and Titiunik (2014b) for related theoretical results and further discussion.

Although not strictly justified by the standard framework, RD designs are routinely interpreted as local randomized experiments, where in a neighborhood of the threshold treatment status is considered *as good as* randomly assigned. Lee (2008) first argued that if individuals are unable to precisely manipulate or affect 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 replicate results from randomized experiments when both designs are available, further bolstering heuristically this "as-good-as randomized" interpretation (Black, Galdo, and Smith, 2007, Buddelmeyer and Skoufias, 2003, Cook, Shadish, and Wong, 2008, Green, Leong, Kern, Gerber, and Larimer, 2009).

Motivated by this common interpretation, we develop a methodological framework for analyzing RD

designs as local randomized experiments employing a randomization inference setup.[1] Characterizing the RD design in this way not only has intuitive appeal, but also leads to an alternative way of conducting statistical inference. Building on Rosenbaum (2002, 2010), we propose a randomization-inference framework to conduct exact finite-sample inference in the RD design that is most appropriate when the sample size in a small window around the cutoff –where local randomization is most plausible– is small. Small sample sizes are a common phenomenon in the analysis of RD designs, since the estimation of the treatment effect at the cutoff typically requires that observations far from the cutoff be given zero or little weight; this may constrain researchers' ability to make inferences based on large-sample approximations. In order to increase the sample size, researchers often include observations far from the cutoff and engage in extrapolation. However, incorrect parametric extrapolation invalidates standard inferential approaches because point estimators, standard errors and test statistics will be biased. In such cases, if a local randomization assumption is plausible, our approach offers a valid alternative that minimizes extrapolation by relying only on the few closest observations to the cutoff. More generally, even when there is no reason to doubt the validity of standard procedures for the analysis of RD designs, our methodological framework offers a complement and a robustness check to these more conventional procedures by providing a framework that requires minimal extrapolation and allows for exact finite-sample inference.

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 et al. (2001, Theorem 2) for identification of heterogeneous treatment effects. The key assumption is that, for the given sample, there exists a neighborhood around the cutoff where a randomization-type condition holds. More generally, this assumption may be interpreted as an approximation device to the conventional continuity conditions that allows us to proceed as if only the few closest observations near the cutoff are randomly assigned. As we discuss in our empirical application, the plausibility of this assumption will necessarily be context-specific, requiring substantive justification and empirical support. Employing these conditions, we then discuss how randomization inference tools may be used to conduct exact finite-sample inference in the RD context, and we also propose different methods for implementation in applications.

Our resulting empirical approach consists of two steps. The first step is choosing a neighborhood or window around the cutoff where treatment status is assumed to be as-if randomly assigned. The size of this window is a critical choice, and it should generally be as small as possible, although what constitutes

[1] Randomization inference has been employed in other treatment effect contexts, including weak instrumental variables (Imbens and Rosenbaum, 2005), natural experiments (Ho and Imai, 2006), spatial statistics (Barrios, Diamond, Imbens, and Kolesar, 2012) and experiments with noncompliance and cluster assignment (Hansen and Bowers, 2009).

a small window will depend on each particular application. We develop a data-driven, randomization-based window selection procedure based on "balance tests" of pre-treatment covariates and illustrate how this data-driven approach for window selection performs in our empirical illustration. The second step is to apply established randomization inference tools, given a hypothesized treatment assignment mechanism, to construct hypothesis tests, confidence intervals, and point estimates. The choice of the assignment mechanism will also depend on the specific application, but common examples include an unrestricted randomization mechanism that assigns 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.

Our approach is analogous to the conventional nonparametric approach, but makes a different tradeoff: our randomization assumption constitutes an approximation that is likely valid within a smaller neighborhood of the threshold than the local linear approach, but allows for exact finite-sample inference in a setting where large-sample approximations may be poor. The choices involved in implementation, however, are parallel. In order to implement standard local-polynomial RD estimation, researchers need to choose (i) a bandwidth and (ii) a kernel and polynomial order, while for our approach researchers need to choose (i) the size of the window around the cutoff where randomization is plausible and (ii) a randomization mechanism and test statistic. As is well known in the nonparametrics literature, bandwidth selection is difficult and estimation results can be highly sensitive to their choice (Calonico, Cattaneo, and Titiunik, 2014b). In our approach, selecting the window is also crucial, and researchers should pay special attention to how it is chosen. On the other hand, selecting a kernel and polynomial order is relatively easier, as is choosing a randomization mechanism and test statistic in our approach.

We illustrate our methodological framework with a study of party-level advantages in U.S. Senate elections, comparing future Democratic vote shares in states where the Democratic party barely won an election to states where it barely lost. Our analysis considers two different outcomes. First, we estimate the effect of the Democratic party barely winning an election for a Senate seat on its vote share in the following election for that seat. We call this the incumbent-party advantage, and find that it is large and positive – similar to results obtained by standard methods. Second, we estimate the effect of the Democratic party barely winning an election for a Senate seat on its vote share in the next Senate election in the state, which is for the state's other Senate seat. Our approach reveals potentially significant heterogeneity that is masked by standard methods. The effect appears to be near zero in a larger window, consistent with existing evidence using standard methods. In a smaller window, however, there is evidence of a negative effect, suggesting that when a party disputes a seat after having very narrowly won the other seat in the state's delegation it suffers electoral losses. This phenomenon, which we call the opposite-party advantage, is consistent with

several hypotheses in the political science literature, including the hypothesis that voters have a preference for balancing the partisanship of their Senate delegation (see Section 5). Our findings complement the results in Butler and Butler (2006), who found zero effects when they studied balancing and related hypotheses in the Senate using a standard RD design. We show that these null results obtained with standard RD methods are sensitive to the choice of bandwidth or window, and may mask substantial heterogeneity in effects.

In sum, our paper makes several contributions. Conceptually, ours is the first to precisely define a set of conditions under which RD designs may be analyzed as randomized experiments. Methodologically, the randomization-based methods we propose for choosing the RD window are new. Empirically, we show suggestive evidence for an opposite-party advantage following very narrow senate races, a new finding relative to the prior literature.

The rest of the paper is organized as follows. Section 2 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 3 discusses data-driven methods to select the window around the cutoff where the randomization assumption may be plausible. Section 4 introduces the classical notion of incumbency advantage in the political science literature and discusses its differences with RD-based measures, and also provides details about our research design and hypotheses in the context of the U.S. Senate. Section 5 presents the result of our empirical analysis, and Section 6 discusses several extensions and applications of our methodology. Section 7 concludes.

2 Randomization Inference in RD

2.1 Empirical Motivation

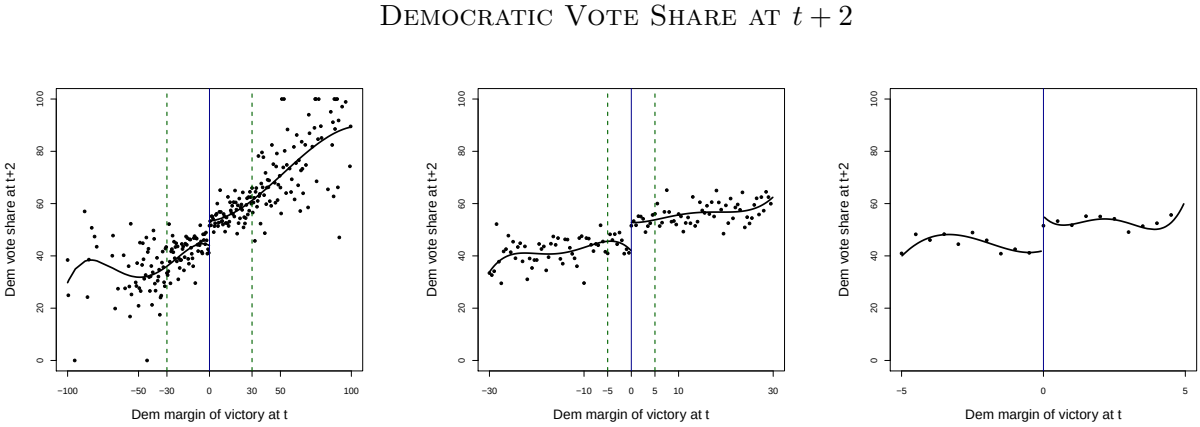
Before presenting the formal statistical framework for randomization-inference in RD designs, we provide some motivation using data from our empirical illustration on U.S. Senate elections. We describe the data and estimands briefly here; details are provided in Sections 4.1 and 5.

We focus on the effect of the Democratic party barely winning an election for a Senate seat at time t on the Democratic vote share in the following election for the same Senate seat, which occurs two elections later, at $t + 2$ (see Section 4.1 for details). Although the U.S. is largely a two-party system where the Democratic and Republican parties compete for elected offices, in some Senate races there are also third-party or write-in candidates. To accommodate situations where more than two parties contest the election, our running variable 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 -100 to 100 percentage points. This is the score that determines treatment in this RD design: when the margin of victory exceeds zero, the Democratic party

wins the election, otherwise it loses. We adopt the Democratic vote share as our outcome of interest; this is without loss of generality, since in the overwhelming majority of races a Democratic defeat is a Republican victory. Our data are all (state-level) Senate elections returns between 1914 and 2010.

We illustrate the estimation of this RD effect in Figure 1(a). The x-axis is the Democratic margin of victory in election t. The y-axis is the outcome of interest, which in this case is the vote share obtained by the Democratic party in election $t+2$. 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 $t+2$ is strongly positive, showing that the higher the vote share at election t for a given Senate seat, the higher the vote share in the following election at $t+2$ for the same seat. The figure also shows a clear jump at the cutoff: the Democratic party obtains discretely higher vote shares at $t+2$ in states where it barely won at t than in states where it barely lost at t. The difference is about ten percentage points.

Figure 1: RD Design in U.S. Senate Elections, 1914-2010 – Polynomial fit for different windows



(a) Elections in [-100, 100] (b) Elections in [-30, 30] (c) Elections in [-5, 5]

Intuitively, this jump will be a valid estimate of the incumbent-party advantage at the cutoff under the assumption that if the Democratic party had counterfactually not enjoyed incumbent status regardless of the previous election's outcome, its average vote share conditional on the previous election's margin of victory would have varied continuously at the threshold. This continuity assumption, however, is not directly testable because naturally we do not observe the $t+2$ vote share the Democratic party would have obtained if it had barely won election t but nevertheless had not been the incumbent party in election $t+2$. Lee (2008) introduced the stronger idea that RD designs resemble local experiments, arguing that this interpretation is justified whenever scores cannot be manipulated precisely. 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. According to this interpretation, winning or losing the election may be considered as-if randomly assigned for states with margins of victory near the threshold. An important consequence of interpreting RD designs 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, potential outcomes 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.

Senate races in our data show little association between outcomes and scores when the prior election was close. Figure 1 illustrates the relationship between the period $t+2$ vote shares and the same-seat prior election (period t) margin of victory in our application. Figure 1(a) uses the full sample, Figure 1(b) employs the subsample of races with t margin of victory within 30 percentage points, and Figure 1(c) uses the subsample of races with t margin of victory within 5 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 mostly vanishes when considering states with elections decided by 5 percentage points or less: 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, as we would expect if victory had been randomly assigned among the electoral races in this window.[2]

In the next subsection, we take the interpretation of RD designs as local experiments as a starting point, and assume that there is a neighborhood around the cutoff where a randomization-type condition holds. We then apply a randomization-inference framework that allows us to both estimate effects and validate the RD design via testable implications of the local randomization assumption we propose.

2.2 Statistical Framework

We employ a randomization inference framework Rosenbaum (2002, 2010). The framework permits exact tests of the presence of treatment effects under minimal assumptions, but allows further assumptions to be incorporated for constructing confidence intervals and point estimates. This section lays out the basic framework, defines the minimal baseline assumptions for inference, and then introduces additional assumptions for confidence intervals and point estimates.

Consider a setting with n units, indexed $i = 1, 2, \cdots, n$, where the scalar R_i is the score observed for unit i, with the n-vector $\mathbf{R}$ collecting the observations. In our application, R_i is the Democratic margin of victory

[2]This window is chosen here purely for motivation. The lack of association between score and outcome is only one implication of local as-if randomization, but is not enough to support this assumption. From a substantive consideration, a $[-5, 5]$ window is too large to justify such an assumption. Indeed, in this window, important pre-determined covariates are significantly different between treatment and control groups. We address this issue directly in Section 3.

(at election t) for state i. We denote unit i's "potential outcome" with $y_i(\mathbf{r})$, where $\mathbf{r}$ is a given value of the vector of scores. The outcome $y_i(\mathbf{r})$ is called a potential outcome because it denotes the outcomes that unit i would exhibit under each possible value of the score vector $\mathbf{r}$.[3] In the randomization inference framework, the potential outcome functions $y_i(\mathbf{r})$ are considered fixed characteristics of the finite population of n units, and the observed vector of scores $\mathbf{R}$ is random.[4] Thus, the observed outcome for unit i is $Y_i \equiv y_i(\mathbf{R})$, and is likewise a random variable with observations collected in the n-vector $\mathbf{Y}$. The essential feature of the RD design is embodied in a treatment variable $Z_i = \mathbb{1}(R_i \geq r_0)$, 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 $\mathbf{Z}$, 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 Section 6.1.

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 $\mathbf{R}$ corresponding to units with R_i inside this window as $\mathbf{R}_{W_0}$, and likewise for other vectors. In addition, we define $F_{R_i|R_i \in W_0}(r)$ to be the conditional distribution function of the score R_i given $R_i \in W_0$, for each unit i. Our main condition casts the RD design as a local randomized experiment.

Assumption 1: Local Randomization. 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) $F_{R_i|R_i \in W_0}(r) = F(r)$, and

(b) $y_i(\mathbf{r}) = y_i(\mathbf{z}_{W_0})$ for all $\mathbf{r}$.

The first part of Assumption 1 says that the distribution of the score is 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 if, for example, the score were affected by the potential outcomes even near the threshold —but may be relaxed, for instance, by explicitly modeling the relationship between R_i and potential outcomes. The second part of this assumption 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

[3] See Holland (1986) for a thorough discussion of the potential outcomes framework.

[4] Note that in this framework, the potential outcomes are fixed and thus the n units are not seen as a sample from a larger population. This could be interpreted as a standard inference approach that conditions on the sampled observations. In Section 6.5, we discuss further connections between our randomization-inference approach and the standard approach to inference in RD.

scores of units outside (i.e., $y_i(\mathbf{r}) = y_i(\mathbf{r}_{W_0})$). 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, $y_i(\mathbf{r}_{W_0}) = y_i(\mathbf{z}_{W_0})$). 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. In our party advantages application, this assumption implies that, in a small window around the cutoff, a party's margin of victory does not affect its vote share in the next election except through winning the previous election.

The conditions in Assumption 1 are stronger than those typically required for identification and inference in the classical RD literature. 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 constant as a function of the score.[5] But Assumption 1 can also be viewed as an approximation to the standard continuity conditions in much the same way the nonparametric large-sample approach approximates potential outcomes as locally linear. This connection is made precise in Section 6.5. Assumption 1 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 focus on the treatment indicators $\mathbf{Z}_{W_0}$. Second, since the distribution of $\mathbf{Z}_{W_0}$ does not depend on potential outcomes, comparisons of observed outcomes across the threshold have a causal interpretation.

In most settings, Assumption 1 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 using exact randomization inference methods to overcome this potential small-sample problem. In the remainder of this section we assume Assumption 1 holds and take as given the window W_0, but we discuss explicitly empirical methods for choosing this window in Section 3.

Hypothesizing the randomization mechanism

The first task in applying randomization inference to the RD design is to choose a randomization mechanism for $\mathbf{Z}_{W_0}$ 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 π. In this case the probability distribution of $\mathbf{Z}_{W_0}$ is given by $\Pr(\mathbf{Z}_{W_0} = \mathbf{z}) = \pi^{\mathbf{z}'\mathbf{1}}(1-\pi)^{(\mathbf{1}-\mathbf{z})'\mathbf{1}}$, for all vectors $\mathbf{z}$ in Ω_{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 $\mathbf{1}$ is a conformable vector of ones. This randomization distribution is fully determined up to the

[5] This assumption could be relaxed to $F_{R_i|R_i\in W_0}(r) = F_i(r)$, allowing each unit to have different probabilities of treatment assignment. However, in order to conduct exact-finite sample inference based on this weaker assumption, further parametric or semiparametric assumptions are needed. See footnote 6 for further discussion on this point.

value π, which is typically unknown in the context of RD applications. A natural choice for π would be $\hat{\pi} = \mathbf{Z}'_{W_0}\mathbf{1}/n_{W_0}$, the fraction of units within the window with scores exceeding the threshold.[6]

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, Ω_{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 $\Pr(\mathbf{Z}_{W_0} = \mathbf{z}) = \binom{n_{W_0}}{m_{W_0}}^{-1}$, for all $\mathbf{z} \in \Omega_{W_0}$.

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.

Test of no effect

Having chosen an appropriate randomization mechanism, we can test the sharp null hypothesis of no treatment effect under Assumption 1. No treatment effect means observed outcomes are fixed regardless of the realization of $\mathbf{Z}_{W_0}$. Under this null hypothesis, potential outcomes are not a function of treatment status inside W_0; that is, $y_i(\mathbf{z}) = y_i$ for all i within the window and for all $\mathbf{z} \in \Omega_{W_0}$, where y_i is a fixed scalar. The distribution of any test statistic $T(\mathbf{Z}_{W_0}, \mathbf{y}_{W_0})$ is known, since it depends only on the known distribution of $\mathbf{Z}_{W_0}$, and $\mathbf{y}_{W_0}$ 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 $\mathbf{z}$ leading to values of $T(\mathbf{z}, \mathbf{y}_{W_0})$ at least as large as the observed value $\tilde{T}$, that is, $\Pr(T(\mathbf{Z}_{W_0}, \mathbf{y}_{W_0}) \geq \tilde{T}) = \sum_{\mathbf{z} \in \mathbf{\Omega}_{W_0}} \mathbb{1}(T(\mathbf{z}, \mathbf{y}_{W_0}) \geq \tilde{T}) \cdot \Pr(\mathbf{Z}_{W_0} = \mathbf{z})$, where $\Pr(\mathbf{Z}_{W_0} = \mathbf{z})$ follows the assumed 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 Assumption 1) 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.

[6] Under the generalization discussed in footnote 5, the parameter π in the Bernoulli randomization mechanism becomes π_i (different probabilities for different units), which could be modeled, for instance, as $\pi_i = \pi(r_i)$ for a parametric choice of the function $\pi(\cdot)$.

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

This assumption means that unit i's potential outcome depends only on z_i, which, together with Assumption 1, allows us to write potential outcomes simply as $y_i(0)$ and $y_i(1)$ for units in W_0. Assumptions 1–2 enable us to characterize the effects of treatment through inference on the distribution or quantiles of the population of n_{W_0} potential outcomes in W_0, $\{y_i(z) : R_i \in W_0\}$, as in Rosenbaum (2002). The goal is to construct a confidence interval $[a(q), b(q)]$ that covers with at least some specified probability the q-quantile of $\{y_i(1) : R_i \in W_0\}$, denoted $Q^1(q)$, which is simply the $\lceil q \times n_{W_0} \rceil$-th order statistic of $\{y_i(1) : 1 \leq i \leq n_{W_0}\}$ for units within the window W_0, 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 $J(x) = \mathbf{Z}'_{W_0}\mathbf{1}(\mathbf{Y}_{W_0} \leq x)$, the number of units above the threshold whose outcomes are less than or equal to x, and has distribution $\Pr(J(x) = j) = \binom{\lceil q \times n_{W_0} \rceil - 1}{j-1}\binom{n_{W_0} - \lceil q \times n_{W_0} \rceil}{m_{W_0} - j} / \binom{n_{W_0}-1}{m_{W_0}-1}$ under a fixed margins randomization mechanism where m_{W_0} denotes the number of treated units inside W_0. Inference on the quantile treatment effect $Q^1(q) - Q^0(q)$ 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 described below.

Assumption 3: 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 1–3, 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 $\mathbf{Z}_{W_0}$. 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 $T(\mathbf{Z}_{W_0}, \mathbf{Y}_{W_0} - \tau_0 \mathbf{Z}_{W_0})$, and the significance level is computed as before. Confidence intervals for the treatment effect can be found by finding all values τ_0 such that the test $\tau = \tau_0$ is not rejected, and Hodges-Lehmann-type point estimates

can also be constructed finding the value of τ_0 such that the observed test statistic $T(\mathbf{Z}_{W_0}, \mathbf{Y}_{W_0} - \tau_0 \mathbf{Z}_{W_0})$ equals its expectation under the null hypothesis.

We discuss this constant and additive treatment effect model because it allows us to illustrate how confidence intervals can be easily derived by inverting hypothesis tests about a treatment effect parameter. But there is nothing in the randomization inference framework that we have adopted that necessitates Assumption 3. This assumption can be easily generalized to allow for non-constant treatment effects, such as tobit or attributable effects (see Rosenbaum 2010, ch.2). Indeed, the technique of constructing adjusted potential outcomes and inverting hypothesis tests of the sharp null hypothesis is general and allows for arbitrarily heterogeneous models of treatment effects. Furthermore, the confidence intervals for quantile treatment effects described above do not require a parametric treatment effect model.

3 Window selection

If there exists a window $W_0 = [\underline{r}, \overline{r}]$ where our randomization-type condition Assumption 1 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 is 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 (see, e.g., Imbens and Kalyanaraman, 2012, Calonico et al., 2014b).

We propose a method to select W_0 based on covariates. These could be either *predetermined* covariates (determined before treatment is assigned and thus, by construction, unaffected by it) or *placebo* covariates (determined after treatment is assigned but nonetheless expected to be unaffected by treatment given prior theoretical knowledge about how the treatment operates). In most empirical applications of the RD design, researchers have access to predetermined covariates and use them to assess the plausibility of the RD assumptions and/or to reduce sampling variability. A typical strategy to validate the design is to test whether there is a treatment effect at the discontinuity for these covariates and interpret the absence of such effect as evidence supporting the credibility of the design.

Our window selector procedure is inspired by this common empirical practice. In particular, we assume that there exists a covariate for each unit, denoted $x_i(\mathbf{r})$, that is unrelated to the score inside W_0 but related to it outside of W_0. This implies that for a window $W \supset W_0$, the score and covariate will be associated for units with $R_i \in W - W_0$ but not for units with $R_i \in W_0$. This means that if the sharp null hypothesis is rejected in a given window, that window is strictly larger than W_0, which leads naturally to a procedure for selecting W_0: perform a sequence of "balance" tests for the covariates, one for each window candidate, beginning with the largest window and then sequentially shrinking it until the test fails to reject balance.

The first step to formalize this approach is to assume that the treatment effect on the covariate x is zero inside the window where Assumption 1 holds. We collect the covariates in $\mathbf{X} = (X_1, X_2, \cdots, X_n)'$ where, as before, $X_i = x_i(\mathbf{R})$.

Assumption 4: Zero Treatment Effect for Covariate. For all i with $R_i \in W_0$: the covariate $x_i(\mathbf{r})$ satisfies $x_i(\mathbf{r}) = x_i(\mathbf{z}_{W_0}) = x_i$ for all $\mathbf{r}$.

Assumption 4 states that the sharp null hypothesis holds for X_i in W_0. This assumption simply states what is known to be true when the available covariate is determined before treatment: treatment could not have possibly affected the covariates and therefore its effect is zero by construction. Note that if X_i is a predetermined covariate, the sharp null holds everywhere, not only in W_0. However, we require the weaker condition that it hold only in W_0 to include placebo covariates.

The second necessary step to justify our procedure for selecting W_0 based on covariate balance is to require that the covariate and the score be correlated outside of W_0. 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. Define $\widetilde{W} = [\underline{\rho}, \underline{r}) \cup (\overline{r}, \overline{\rho}]$ for a pair $(\underline{\rho}, \overline{\rho})$ satisfying $\underline{\rho} < \underline{r} < \overline{r} < \overline{\rho}$, and recall that $r_0 \in W_0 = [\underline{r}, \overline{r}]$.

Assumption 5: Association Outside W_0 between Covariate and Score. For all i with $R_i \in \widetilde{W}$ and for all $r \in \widetilde{W}$:

(a) $F_{R_i|R_i \in \widetilde{W}}(r) = F(r; x_i(r))$, and

(b) For all $j \neq k$,

$$x_j > x_k \Rightarrow F(r; x_j) < F(r; x_k) \qquad \text{or} \qquad x_j > x_k \Rightarrow F(r; x_j) > F(r; x_k).$$

Assumption 5 is key to obtain a valid window selector, since it requires a form of non-random selection among units outside W_0 that leads to an observable association between the covariate and the score for those units with $R_i \notin W_0$, i.e., between the vectors $\mathbf{X}_{\widetilde{W}}$ and $\mathbf{R}_{\widetilde{W}}$. In other words, under Assumption 5 the vectors $\mathbf{X}_W$ and $\mathbf{R}_W$ will be associated for any window W such that $W \supset W_0$. Since x is predetermined or placebo, this association cannot arise because of a direct effect of r on x. Instead, it may be that x affects r (higher campaign contributions at $t-1$ lead to higher margin of victory at t) or that some observed or unobserved factor affects both x and r (more able politicians are both more likely to raise high contributions and win by high margins). In other words, Assumption 5 leads to units with high R_i having high (or low) X_i, even when X_i is constant for all values of r.

Assumptions 4 and 5 justify a simple procedure to find W_0. This procedure finds the widest window for which the covariates and scores are not associated inside this window, but are associated outside of it. We

base our procedure on randomization-based tests of the sharp null hypothesis of no effect for each available covariate x. Given Assumption 4 above, for units with $R_i \in W_0$, the treatment assignment vector $\mathbf{Z}_{W_0}$ has no effect on the covariate vector $\mathbf{X}_{W_0}$. Under this assumption, 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 Assumption 5 (or a similar assumption), this procedure will be able to detect the true window W_0. Such a procedure can be implemented in different ways. A simple approach is to begin by considering all observations (i.e., choosing the largest possible window W_0), test the sharp null of no effect of Z_i on X_i for these observations and, if the null hypothesis is rejected, continue by decreasing the size of the window until the resulting test fails to reject the null hypothesis.

The procedure depends crucially on sequential testing in *nested* windows: if the sharp null hypothesis is rejected for a given window, then this hypothesis will also be rejected in any window that contains it (with a test of sufficiently high power). Thus, the procedure searches windows of different sizes until it finds the largest possible window such that the sharp null hypothesis cannot be rejected for any window contained in it. This procedure can be implemented as follows.

Window Selection Procedure Based on Predetermined Covariates. Select a test statistic of interest, denoted $T(\mathbf{X}, \mathbf{R})$. Let $R_{(j)}$ be the j-th order statistic of $\mathbf{R}$ 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 $T(\mathbf{X}_{W(j_1,j_2)}, \mathbf{R}_{W(j_0,j_1)})$.

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.

An important feature of this approach is that, unlike conventional hypothesis testing, we are particularly concerned about the possibility of failing to reject the null hypothesis when it is false (Type II error). Usually, researchers are concerned about controlling Type I error to avoid rejecting the null hypothesis too often when it is true, and thus prefer testing procedures that are not too "liberal". In our context, however, rejecting the null hypothesis is used as evidence that the local randomization assumption 1 does not hold, and our ultimate goal is to learn whether the data supports the existence of a neighborhood around the cutoff where our null hypothesis fails to be rejected. In this sense, the roles of Type I and Type II error are interchanged

in our context.[7] This has important implications for the practical implementation of our approach, which we discuss further below.

3.1 Implementation

Implementing the procedure proposed above requires three choices: (i) a test statistic, (ii) the minimum sample sizes ($j_{0,min}$, $j_{1,min}$), and (iii) a testing procedure and associated significance level α. We discuss here how these choices affect our window selector, and give guidelines for researchers who wish to use this procedure in their empirical applications.

(i) Choice of test statistic. This choice is important because different test statistics will have power against different alternative hypotheses and, as discussed above, we prefer tests with low type II error. In our procedure, the sharp null hypothesis of no treatment effect could employ different test statistics such as difference-in-means, Wilcoxon rank sum or Kolmogorov-Smirnov, because the null randomization distribution of any of them is known. Lehmann (2006) and Rosenbaum (2002, 2010) provide a discussion and comparison of alternative test statistics. In our application, we employ the difference-in-means test statistic.

(ii) Choice of minimum sample size. The main goal of setting a minimum sample size is to prevent the procedures from having too few observations when conducting the hypothesis test in the smallest possible window. These constants should be large enough so that the test statistic employed has "good" power properties to detect departures from the null hypothesis. We recommend setting $j_{0,min}$ and $j_{1,min}$ so that a minimum of roughly 10 observations are included at either side of the threshold. One way of justifying this choice is by considering a two-sample standard normal shift model with a true treatment effect of one standard deviation and 10 observations in each group, in which case a randomization-based test of the sharp null hypothesis of no treatment effect using the difference-in-means statistic has power of roughly 80 percent with significance level of 0.15 (and 60 percent with significance level of 0.05).[8] Setting $j_{0,min}$ and $j_{1,min}$ at higher values will increase the power to detect departures from Assumption 1 and will lead to a more conservative choice of W_0 (assuming the chosen window based on those higher values is feasible, that is, has positive length).

(iii) Choice of testing procedure and α. First, our procedure performs hypothesis tests in a sequence of nested windows and thus involves multiple hypothesis testing (see Efron (2010) for a recent review). This implies that, even when the null hypothesis is true, it will be rejected several times (e.g., if the hypotheses are independent, they will be rejected roughly as many times as the significance level times the number of windows considered). For the family-wise error rate, multiple testing implies that our window selector will

[7] An alternative is to address this issue directly by changing the null hypothesis to be the existence of a treatment effect. This could be implemented with sensitivity analysis (Rosenbaum, 2002) or equivalence tests (Wellek, 2010).

[8] Power calculations based on simulations are not reported here, but available upon request.

reject more windows than it should, because the associated p-values will be too small. But since we are more concerned about failing to reject a false null hypothesis (type II error) than we are about rejecting a true one (type I error), this implies that our procedure will be more conservative, selecting a smaller window than the true window (if any) where the local randomization assumption is likely to hold. For this reason, we recommend that researchers do not adjust p-values for multiple testing.[9] Secondly, we must choose a significance level α to test whether the local randomization assumption is rejected in each window. As our focus is on type II error, this value should be chosen to be higher than conventional levels for a conservative choice for W_0. Based on the power calculations discussed above, a reasonable choice is to adopt $\alpha = 0.15$; higher values will lead to a more conservative choice of W_0 if a feasible window satisfies the stricter requirement. Nonetheless, researchers should report all p-values graphically so that others can judge how varying α would alter the size of the chosen window. Finally, when the sharp null is tested for multiple covariates in every candidate window, the results of multiple tests must be aggregated in a single p-value. To be as conservative as possible, we choose the minimum p-value across all tests in every window.[10]

We illustrate how our methodological framework works in practice with a study of party advantages in U.S. Senate elections. The following section discusses the relationship between the RD estimand of the incumbent-party advantage and the classical notions of incumbency advantage that have been extensively studied in the political science literature, presents our research design, and defines our estimands of interest. We present window selector and estimation results in Section 5.

4 Regression Discontinuity and the Party Incumbency Advantage

Political scientists have long studied the question of whether the incumbent status of previously elected legislators translates into an electoral or *incumbency* advantage. This advantage is believed to stem from a variety of factors, including access to franking privileges, name recognition, the ability to perform case-work and cultivate a personal vote, the ability to deter high-quality challengers, the implementation of pro-incumbent redistricting plans, and the easy availability of the incumbency cue amidst declining party attachments. Although the literature is vast,[11] it has focused overwhelmingly on the incumbency advantage of members of the U.S. House of Representatives. With few exceptions, incumbency advantage scholars have

[9] An alternative approach is to select a false discovery rate among all windows such that the non-discovery rate, an analog of type II error in multiple testing contexts, is low enough (Craiu and Sun, 2008).

[10] Other procedures that maximize covariate balance have focused on the minimum p-value across many covariates. See, for example, Diamond and Sekhon (2013). An alternative, but less conservative, procedure would be to use an omnibus test.

[11] See Ansolabehere and Snyder (2002), Erikson (1971), Gelman and King (1990), Erikson and Titiunik (2013) and references therein.

paid little attention to other offices.[12]

Estimating the incumbency advantage is complicated by several factors. One is that high-quality politicians tend to obtain higher vote shares than their low-quality counterparts, making them more likely both to become incumbents in the first place and to obtain high vote shares in future elections. Another is that incumbents tend to retire strategically when they anticipate a poor performance in the upcoming election, making "open seats" (races where no incumbent is running) a dubious baseline for comparison. Any empirical strategy that ignores these methodological issues will likely overestimate the size of the incumbency advantage. Political scientists have been aware of the methodological difficulties since at least the 1970s (see Erikson, 1971), and Gelman and King (1990) clarified many of the inferential challenges to interpreting this advantage as a causal effect by defining the incumbency advantage estimand in terms of counterfactuals. In particular, Gelman and King (1990) defined the incumbency advantage as the difference between two potential outcomes: the vote share the incumbent legislator receives in her district when she runs against a major party opposition, minus the vote share the incumbent party receives in the same district when the legislator does not run and all major parties compete for the open seat.

In recent years, scholars have also begun to explore the use of natural experiments and quasi-experimental research designs to study the incumbency advantage, hoping to overcome some of the inferential obstacles while avoiding assumptions about retirement decisions that in some applications may be too strong. In this vein, Lee (2008) proposed using a regression discontinuity design based on the discontinuous relationship between the incumbency status of a party in a given election and its vote share in the previous election: assuming a two-party system for simplicity, a party enjoys incumbency status when it obtains 50% of the vote or more in the previous election, but loses incumbency status to the opposing party when its share in the previous election falls short of 50%. In this RD design, the score is the vote share obtained by a party at election t, the cutoff is 50%, and the treatment (incumbent status) is assigned deterministically based on whether vote share at t exceeds the cutoff. The outcome of interest is the party's vote share in the following election, at $t+1$. Thus, in its original formulation, the RD design compares districts where the party barely won election t to districts where the party barely lost election t, and computes the difference in the vote share obtained by the party in the following election, at $t+1$. This difference is the boost in the party's vote share obtained by barely-winning relative to barely-losing, and although it contains information about the advantages of incumbent parties, it is not a measure of the incumbency advantage according to Gelman

[12] In particular, the incumbency advantage literature has largely ignored the U.S. Senate. A number of studies focused on related issues, such as a comparison of reelection rates between Senate and House incumbents (e.g., Abramowitz, 1980, Collier and Munger, 1994) and the role of challengers in senatorial incumbent success (e.g., Krasno, 1994, Lublin, 1994), but none of these articles analyzed the measures of incumbency advantage that have been so extensively employed in the analysis of House incumbents. A few articles do consider some of these measures (e.g., Ansolabehere and Snyder, 2002, Ansolabehere, Hansen, Hirano, and Snyder, 2007), analyzing the Senate along with many other offices.

and King's (1990) classical definition. The differences become clearest when both estimands are defined in terms of potential outcomes. While the Gelman and King (1990) estimand is the difference between the vote received by the incumbent legislator and the vote received by the legislator's party in the same district in an open seat (i.e., if the legislator retires), the RD estimand is the vote received by a party when the party wins the previous election minus the vote received by the same party when it loses the previous election, *regardless* of whether the individual legislator who won the previous election runs for reelection. The RD estimand is therefore better conceived as an estimand of the party-level incumbency advantage, which is conceptually distinct from the classical notion of the incumbency advantage.[13]

4.1 RD Design in U.S. Senate Elections: Two Estimands of Party Advantage

Our application of the RD design to U.S. Senate elections focuses on two specific estimands that capture local advantages and disadvantages at the party level. The first estimand, which we call the incumbent-party advantage, focuses on the effect of the Democratic party winning a Senate seat on its vote share in the following election for that seat. This is analogous to the original RD design proposed by Lee (2008) to estimate incumbent-party advantages in the U.S. House described above. This RD-based estimand of the incumbent-party advantage has not been previously explored in the context of U.S. Senate elections.

The other estimand, which we call the opposite-party advantage following Alesina, Fiorina, and Rosenthal (1991), is unrelated to the traditional concept of the incumbency advantage, and reveals the disadvantages faced by the party that tries to win the second seat in a state's Senate delegation. In contrast to the incumbent-party advantage, which has not received much scholarly attention, the concept of the opposite-party advantage has been more thoroughly studied. In particular, establishing whether the opposite-party advantage exists has been of central importance to theories of split-party Senate delegations, and there are different explanations of why it may arise. For example, Alesina et al. (1991) argue that it results from voters' preferences for policy balancing and the staggered structured of Senate terms. The more extreme the position of the sitting senator, the more voters will prefer to vote for a senator of the opposing party in the other Senate seat to balance the position of the sitting senator and achieve a more moderate average position in the state delegation as a whole. Others such as Jung, Kenny, and Lott (1994), claim that this advantage arises from the fact that the party of the sitting senator finds it more difficult to mobilize its core supporters than the opposite party because voters obtain relatively less utility from their preferred party winning the second rather than the first seat. And Segura and Nicholson (1995) argue that the opposite party advantage does not exist and split-party delegations simply result from the political characteristics of each race.

[13] See also Caughey and Sekhon (2011, p. 402) for a discussion about the connection between a global polynomial RD estimator and the Gelman and King (1990) estimator, and Erikson and Titiunik (2013) for a discussion of the relationship between the RD estimand of the incumbency advantage and the personal incumbency advantage.

Empirical tests of the opposite party advantage have been inconclusive. The evidence most relevant to our study was presented by Butler and Butler (2006), who studied Senate elections in the 1946-2004 period with an RD design that focused on the effect of winning at t on the vote share at $t+1$. Using standard inference approaches, they found a null effect and concluded that consecutive Senate elections are independent of each other. As we show in Section 5, we find similar results using our randomization-based inference methods within our data-driven window around the threshold. The null result is sensitive to the choice of window, however, and may be masking heterogeneous effects closer to the threshold: focusing on the closest elections reveals suggestive evidence of an opposite party advantage, suggesting that voters possibly incorporate the partisanship of the sitting senator when voting for the other Senate seat.

Both estimands, formally defined in terms of potential outcomes below, are derived from applying an RD design to the staggered structure of Senate elections, which we now describe briefly. Term length in the U.S. Senate is six years and there are only 100 seats. These Senate seats are divided into three classes of roughly equal size (Class I, Class II and Class III), and every two years only the seats in one class are up for election. As a result, the terms are staggered: in every general election, which occurs every two years, only one third of Senate seats are up for election.[14] Each state elects two senators in different classes to serve a six-year term in popular statewide elections. Since its two senators belong to different classes, each state has Senate elections separated by alternating 2-year and 4-year intervals. Moreover, in any pair of consecutive elections, each election is for a *different* senate seat – that is, for a seat in a different class. For example, Florida's two senators belong to Class I and III. The senator in Class I was elected in 2000 for six years, and was up for reelection in 2006, while the senator in Class III was elected in 2004 for six years and was up for reelection in 2010. Thus, Florida had Senate elections in 2000 (Class I senator), 2004 (Class III senator), 2006 (Class I senator), and 2010 (Class III senator).

We apply the RD design in the Senate analogously to its previous applications in the House, comparing states where the Democratic party barely won election t to states where the Democratic party barely lost. But in the Senate, the staggered structure of terms adds a layer of variability that allows us to both study party advantages and validate our design in more depth than would be possible in a non-staggered legislature such as the House. Using t, $t+1$ and $t+2$ to denote three successive elections, the staggered structure of the Senate implies that the incumbent elected at t, if he or shee decides to run for reelection, will be on the ballot at $t+2$, but not at $t+1$, when the Senate election will be for the *other* seat in the state. As summarized in Table 1, this staggered structure leads to two different research designs analyzing two separate effects.

The first design (Design I), focuses on the effect of party P's barely winning at t on its vote share at

[14] Senators in Class I were first elected by popular vote in 1916 and every six years thereafter; Senators in Class II were first elected by popular vote in 1918 and every six years thereafter; and Senators in Class III were first elected by popular vote in 1914 and every six years thereafter.

Table 1: Three consecutive Senate elections in a hypothetical state

Election	Seat A	Seat B	Design and Outcomes
t	Election held. Candidate C from party P wins	No election held	-
$t+1$	No election held	Election held. (Candidate C is not a contestant in this race)	Design II: Effect of P winning Seat A at t on P's vote share for Seat B at $t+1$ (**Opposite-party advantage**)
$t+2$	Election held. Candidate C may or may not be P's candidate	No election held	Design I: Effect of P winning Seat A at t on P's vote share for Seat A at $t+2$ (**Incumbent-party advantage**)

$t+2$, the second election after election t, and defines the first RD estimand we study. As illustrated in the third row of Table 1, in Design I elections t and $t+2$ are for the *same* Senate seat, and this incumbent-party effect captures the added vote share received by the Democratic party due to having won (barely) the seat's previous election. The second research design (Design II), illustrated in the second row of Table 1, allows us to analyze the effect of party P's barely winning election t on the vote share it receives in election $t+1$ for the state's other seat, when the incumbent candidate elected at t is, by construction, not contesting the election. Thus, Design II defines the second RD estimand, the opposite-party advantage, which will be negative when the party of the sitting senator (elected at t) is at a disadvantage relative to the opposing party in the election for the other seat (which occurs at $t+1$).

We can define the two estimands defined by Designs I and II formally using the notation introduced in Section 2. We define the treatment indicator as $Z_{it} = \mathbb{1}(R_{it} \geq r_0)$ and the potential outcomes in elections $t+2$ and $t+1$, respectively, as $y_{it+2}(Z_{it})$ and $y_{it+1}(Z_{it})$.[15] Thus, the incumbent-party advantage for an individual state i is defined as $\tau_i^{IP} = y_{it+2}(1) - y_{it+2}(0)$ and the opposite-party advantage as $\tau_i^{OP} = y_{it+1}(1) - y_{it+1}(0)$. Our randomization inference approach to RD offers hypothesis testing and point-type estimators (e.g. Hodges-Lehmann) of these parameters for the units in the window W_0 where local randomization holds.

5 Results: RD-based Party Advantages in the U.S. Senate

We use our randomization-based framework to analyze both the incumbent-party advantage and the opposite-party advantage in the U.S. Senate, in the period 1914–2010. After describing our data sources, we use our window selector to choose the window where we invoke our local randomization assumption. We then

[15] Since our running variable is the Democratic victory at election t and our outcomes of interest occur later in elections $t+1$ and $t+2$, we add a subscript t to R_i and Z_i to clarify that they are determined before the outcomes.

estimate the RD-based party advantages within this window.

5.1 Data

We analyze U.S. Senate elections between 1914 and 2010. This is the longest possible period to study popular U.S. Senate elections, as before 1914 the U.S. Constitution mandated that Senate members be elected indirectly by state legislatures. The Seventeenth Amendment, ratified in 1913, established the direct election of senators by the state's citizens. Although by 1913 many states relied on indirect mechanisms to allow voters to choose their senators, the first year senatorial elections were held by direct popular vote was 1914 (Senate Historical Office, 2012).

We combine several data sources. We collected election returns for the period 1914-1990 from The Interuniversity Consortium for Political and Social Research (ICPSR) Study 7757, and for the period 1990-2010 from the CQ Voting and Elections Collection. We obtained population estimates at the state level from the U.S. Census Bureau. We also used ICPSR Study 3371 and data from the Senate Historical Office to establish whether each individual senator served the full six years of his or her term, and exclude all elections in which a subsequent vacancy occurs. We exclude vacancy cases because, in most states, when a Senate seat is left vacant the governor can appoint a replacement to serve the remaining time in the term or until special elections are held, and in most states appointed senators need not be of the same party as the incumbents they replace, leaving the "treatment assignment" of the previous election undefined.[16]

5.2 Selecting the Window

We selected our window using the method based on predetermined covariates presented in Section 3. The largest window we considered was $[-100, 100]$, covering the entire support of our running variable. Based on power considerations discussed above, the minimum window we considered was $[-0.5, 0.5]$, because within this window there are 9 and 14 outcome observations to the left and right of the cutoff, respectively, and we wanted to set $j_{0,min}$ and $j_{1,min}$ to be approximately equal to 10. Using our notation in Section 3, this means we set $[R_{(j_{0,min})}, R_{(j_{1,min})}] = [-0.50, 0.50]$ and $[R_{(1)}, R_{(n)}] = [-100, 100]$. We analyzed all symmetric windows around the cutoff between $[-0.5, 0.5]$ and $[-100, 100]$ in increments of 0.125 percentage points. That is, we analyzed the sequence of windows $[-100, 100]$, $[-99.875, 99.875]$, $[-99.750, 99.750]$, ..., $[-0.75, 0.75]$, $[-0.625, 0.625]$, $[-0.50, 0.50]$. In every window, we performed randomization-based tests of the sharp null hypothesis of no treatment effect for each of eight predetermined covariates: state-level Democratic percentage of the vote in the past presidential election, state population, Democratic percentage of the vote in the $t-1$ Senate election, Democratic percentage of the vote in the $t-2$ Senate election, indicator for

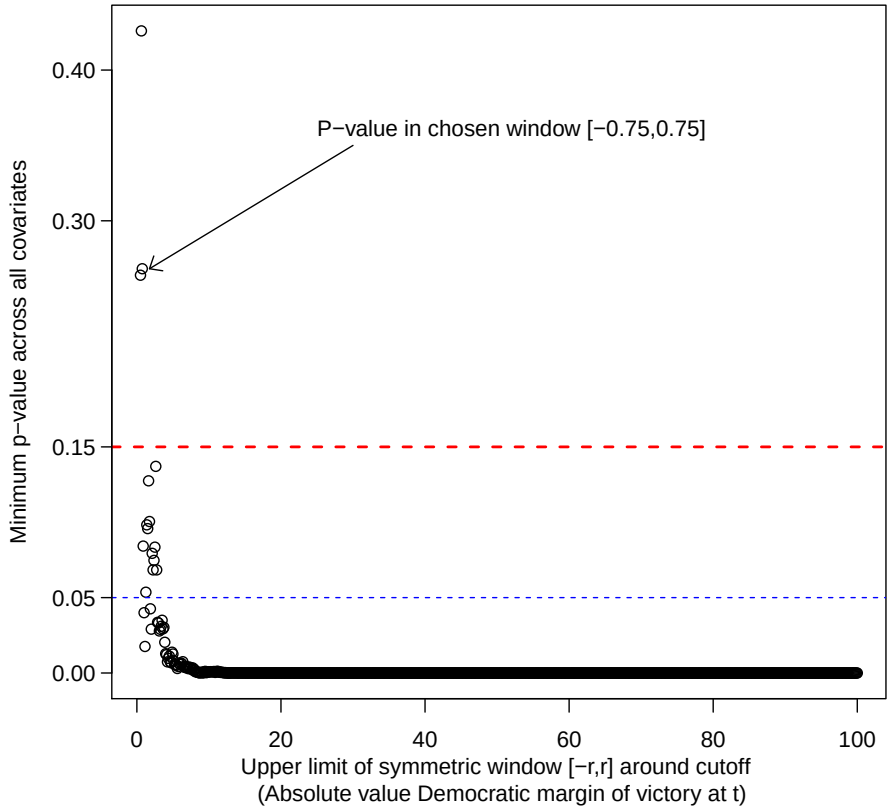
[16] Dropping these observations is equivalent to the routine practice of dropping redistricting years in RD party incumbency analysis of the U.S. House, where incumbency is undefined after redistricting plans are implemented.

Democratic victory in the $t-1$ Senate election, indicator for Democratic victory in the $t-2$ Senate election, indicator for open Senate seat at t, indicator for midterm (non-presidential) election at t and indicator for whether the president of the U.S. at t is Democratic. As discussed above, we set $\alpha = 0.15$, and use the difference-in-means as the test-statistic in our randomization-based tests. These tests (and similar tests for the outcomes presented below) are based on 10,000 simulations of the randomization distribution of $\mathbf{Z_{W_o}}$ assuming a fixed-margins assignment mechanism. For each window, we chose the minimum p-value across these eight covariates.

Figure 2 summarizes graphically the results of our window selector. For every symmetric window considered (x-axis), we plot the minimum p-value found in that window (y-axis), that is, the minimum of the eight p-values associated with each randomization-based test of the sharp null hypothesis that we performed for each of our eight covariates. The x-axis is the absolute value of our running variable, the Democratic margin of victory at election t, which is equivalent to the upper limit of each window considered (since we only consider symmetric windows) and ranges from 0 to 100. For example, the point 20 on the x-axis corresponds to the $[-20, 20]$ window. The figure also shows the conventional significance level of 0.05 and the significance level of 0.15 that we use for implementation. There are a few notable patterns in this figure. First, for most of the windows considered, the minimum p-value is indistinguishable from zero, which means that there is strong evidence against Assumption 1 in most of the support of our running variable. Second, the minimum p-value is above the conventional 5% significance level in very few windows (15 out of the total 797 windows considered). Third, the decrease in p-values is roughly monotonic and very rapid, suggesting that Assumption 1 is implausible except very close to the cutoff. Using $\alpha = 0.15$, our chosen window is $[-0.75, 0.75]$, the third smallest window we considered, since this is the largest window where the minimum p-value exceeds 15% in that window and all windows contained in it. Table 2 shows the minimum p-values for the first five consecutive windows we considered, and also for the windows $[-1.5, 1.5]$, $[-2, 2]$, $[-10, 10]$ and $[-20, 20]$. The minimum p-value in our chosen window is 0.2682, and the minimum p-value in the next largest window, $[-0.875, 0.875]$, is 0.0842. P-values decrease rapidly after that and, with some exceptions such as around window $[-1.50, 1.50]$, do so monotonically. Note also that had we set $\alpha = 0.10$, our chosen window would have still been $[-0.75, 0.75]$. And if we had set $\alpha = 0.05$, our chosen window would have been $[-0.875, 0.875]$, barely larger than our final choice, which shows the steep decline of the minimum p-value as we include observations further from the cutoff.

Our window selection procedure suggests that Assumption 1 is plausible in the window $[-0.75, 0.75]$. Further inspection and analysis of the 38 observations in this window (23 treated and 15 control) shows that these observations are not associated in any predictable way. These electoral races are not concentrated in a particular year or geographic area: these 38 races are spread across 24 different years with no more

Figure 2: Window Selector Based on Predetermined Covariates



than 3 occurring in the same year, and 26 different states with at most 4 occurring in the same state. This empirical finding further supports the idea that these observations might be treated as-if randomly assigned. Moreover, an important implication of this finding is that there is no observable clustering structure in the sample inside the window $[-0.75, 0.75]$, which in turn implies that standard randomization inference techniques are directly applicable. Finally, we also performed standard density tests for sorting and found no evidence of any systematic discrepancy between control and treatment units.[17] Thus, below we proceed to make inferences about the treatment effects of interest under Assumption 1 in this window.

5.3 Inference Within The Selected Window

We now show that results obtained by conventional methods are robust to our randomization-based approach in both Design I and Design II. Randomization-based results within the window imply a sizable advantage when a party's same seat is up for election (Design I) that is very similar to results based on conventional methods. Randomization results on outcomes when the state's other seat is up for re-election (Design II)

[17]The p-value of the McCrary test is 0.39; the null hypothesis of this test is that there is no discontinuity in the density of the running variable around the cutoff (see McCrary (2008) for details). In addition, we cannot reject that our treated and control groups were generated from 38 trials of a Bernoulli experiment with probability of success equal to 0.5 (p-value 0.2559).

Table 2: Window selector based on pretreatment covariates: Randomization-based p-values from balance tests for different windows

Window	Mininum p-value	Covariate with minimum p-value
$[-0.500, 0.500]$	0.2639	Dem Senate Vote t-2
$[-0.625, 0.625]$	0.4260	Open Seat t
$[-0.750, 0.750]$	0.2682	Open Seat t
$[-0.875, 0.875]$	0.0842	Open Seat t
$[-1.000, 1.000]$	0.0400	Open Seat t
$[-1.500, 1.500]$	0.0958	Midterm t
$[-2.000, 2.000]$	0.0291	Midterm t
$[-10.00, 10.00]$	0.0008	Open Seat t
$[-20.00, 20.00]$	0.0000	Dem Senate Vote t-1

show a null effect, also in accordance with conventional methods. However, as we discuss below, the null opposite advantage results from Design I are sensitive to our window choice, and a significant opposite-party advantage appears in the smallest window contained within our chosen window.

Table 3 contrasts the party advantage estimates and tests obtained using our randomization-based framework, reported in column (3), to those obtained from two classical approaches: a 4th-order parametric fit as in Lee (2008) reported in column (1), and a non-parametric local-linear regression with a triangular kernel as suggested by Imbens and Lemieux (2008), using mean-square-error optimal bandwidth (Imbens and Kalyanaraman, 2012), reported in column (2). For both approaches, we show conventional confidence intervals; for the local linear regression results, we also show the robust confidence intervals developed by Calonico et al. (2014b).[18] Panel A presents results for Design I on the incumbent-party advantage, in which the outcome is the Democratic vote share in election $t + 2$. Panel B presents results for Design II on the opposite-party advantage, in which the outcome is the Democratic vote share in election $t + 1$. Our randomization based results are calculated in the window $[-0.75, 0.75]$ chosen above. Note that, as mentioned above, there is no need for clustering in our window, nor is clustering empirically possible.

The point estimates in the first row of Panel A show an estimated incumbent-party effect of around 7 to 9 percentage points for standard RD methods and 9 percentage points for the randomization-based approach. These estimates are highly significant (p-values for all three approaches fall well below conventional levels) and point to a substantial advantage to the incumbent party when the party's seat is up for re-election. In other words, our randomization-based approach shows that the results obtained with standard methods are remarkably robust: a local or global approximation that uses hundreds of observations far away from the cutoff yields an incumbent-party advantage that is roughly equivalent to the one estimated with the 38 races

[18] Conventional local polynomial results are estimated with the package `rdrobust` developed by Calonico, Cattaneo, and Titiunik (2014a,c).

decided by three quarters of a percentage point or less. This robustness is illustrated in the top panel of Figure 3. Figure 3(a) displays the fit of the Democratic Vote Share at $t+2$ from a local linear regression on either side of the Imbens-Kalyanaraman bandwidth, and shows a clear jump at the cutoff of roughly 7.3 percentage points (dots are binned means). Figure 3(b) on the right displays the mean of the Democratic Vote Share at $t+2$ on either side of our chosen $[-0.75, 0.75]$ window (dots are individual data points), and shows a similar (slightly larger) positive jump at the cutoff.

Table 3: Incumbent- and opposite-party advantage in the U.S. Senate using an RD design

	Conventional Approaches		Randomization-based approach
	Parametric	Non-parametric	
	(1)	(2)	(3)
	A. Design I (outcome = Dem Vote Share at t+2)		
Point estimate	9.41	7.32	9.32
p-value	0.0000	0.0000	0.0006
95% CI	[6.16 , 12.65]	[4.47 , 10.17]	[4.56 , 14.84]
95% CI robust	-	[4.11 , 12.54]	-
.25-QTE 95% CI	-	-	[-2.00 , 21.12]
.75-QTE 95% CI	-	-	[3.68 , 18.94]
Bandwidth/Window	-	19.12	[-0.75 , 0.75]
Sample size treated	702	340	22
Sample size control	595	381	15
	B. Design II (outcome = Dem Vote Share at t+1)		
Point estimate	0.64	0.35	-0.79
p-value	0.79	0.83	0.63
95% CI	[-3.16 , 4.44]	[-2.79 , 3.49]	[-8.11 , 5.05]
95% CI robust	-	[-7.05 , 4.19]	-
.25-QTE 95% CI	-	-	[-8.75 , 9.96]
.75-QTE 95% CI	-	-	[-11.15 , 11.31]
Bandwidth/Window	-	23.33	[-0.75 , 0.75]
Sample size treated	731	397	23
Sample size control	610	429	15

Notes: Results based on U.S. Senate elections from 1914 to 2010. Point estimate is from Hodges-Lehmann estimator. Treatment effect confidence intervals are obtained by inverting a test of a constant treatment effect model, using the difference-in-means test statistic and assuming a fixed margins randomization mechanism. P-values are randomization-based and correspond to a test of the sharp null hypothesis of no treatment effect, assuming a fixed margins randomization mechanism. CI denotes 95% confidence intervals (CI). The quantities '.25-QTE CI' and '.75-QTE CI' denote the 95% CI for the 25th-quantile and 75th-quantile treatment effects, respectively, and are constructed as described in the text. For the conventional approaches, results are estimated with the `R` and `Stata` software `rdrobust` developed by Calonico et al. (2014a,c).

In our data-driven window, estimates of the opposite-party party advantage also appear robust to the method of estimation employed. In Panel B, estimates on Democratic Vote Share at $t+1$ based on conventional methods show very small, statistically insignificant effects of around 0.64 to 0.35 in columns (1)

and (2). These standard methods of inference for RD are therefore unable to reject the hypothesis of a null effect, and would suggest that, contrary to the balancing and constituency-based theories discussed above, there is no opposite-party advantage in the U.S. Senate. Our randomization-based approach, presented in column (3) of Panel B, arrives at a similar conclusion, finding a negative but insignificant point estimate and a 95% confidence interval for a constant treatment effect that ranges roughly between -8 and 5. Similarly, the 95% confidence intervals for the 25th and 75th quantile treatment effects are roughly centered around zero and are consistent with a null opposite party advantage.

These results are illustrated in the bottom row of Figure 3, where Figures 3(c) and 3(d) are analogous to Figures 3(a) and 3(b), respectively. The effect of winning an election by 0.75% appears roughly equivalent to the effect estimated by standard methods. In our randomization-based window, the mean of the control group is slightly larger than the mean of the treatment group, but as shown in Table 3 we find no evidence that this constitutes a statistically significant opposite party advantage.

Taken together, our results provide interesting evidence about party-level advantages in the U.S. Senate. First, our results show that there is a strong and robust incumbent-party effect, with the party that barely wins a Senate seat at t receiving on average seven to nine additional percentage points in the following election for that seat. Second, our randomization-based approach confirms the previous finding of Butler and Butler (2006), according to which there is no opposite-party advantage in the U.S. Senate. As we show below, however, and in contrast to the incumbent-party advantage results, the opposite-party advantage result is sensitive to our window choice, and becomes large and significant as predicted by theory inside a smaller window.

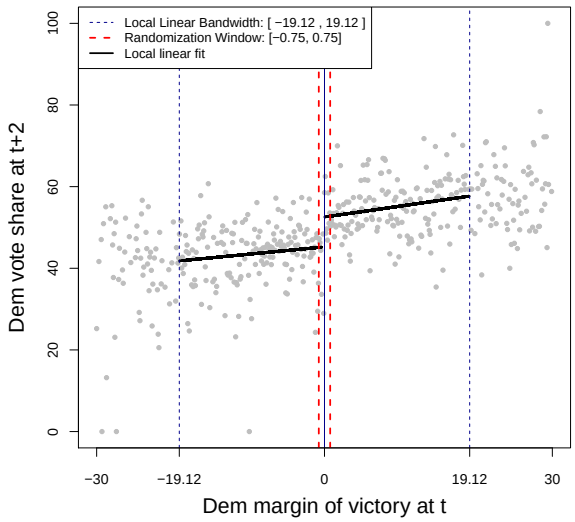
5.4 Sensitivity of Results to Window Choice and Test Statistics

Given different practical choices that had be to made to choose our window, a natural question is whether our empirical findings are sensitive to these choices. In this section, we study the sensitivity of our results to two choices: the window size and the test-statistic used to conduct our tests. We study this question here, and encourage this practice as routine when using our methodological framework.

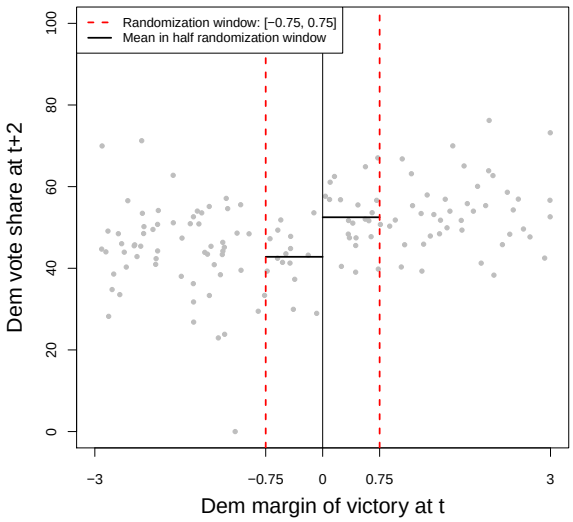
First, we replicate the randomization-based analysis presented above for different windows, both larger and smaller than our chosen $[-0.75, 0.75]$ window. We consider one smaller window, $[-0.5, 0.5]$, and two larger windows, $[-1.0, 1.0]$ and $[-2.0, 2.0]$. We note that, given the results in Table 2, we do not believe that Assumption 1 is plausible in windows larger than $[-0.75, 0.75]$ and we therefore would not interpret a change in results in larger windows as evidence against our chosen window. Nonetheless, it is valuable to know if our findings would continue to hold even under departures of Assumption 1 in larger windows. This observation, however, does not apply when considering smaller windows contained in $[-0.75, 0.75]$, since if Assumption

Figure 3: RD Design in U.S. Senate Elections, 1914-2010 – Standard local-linear approach vs. Randomization-based approach

DESIGN I: DEMOCRATIC VOTE SHARE AT $t+2$

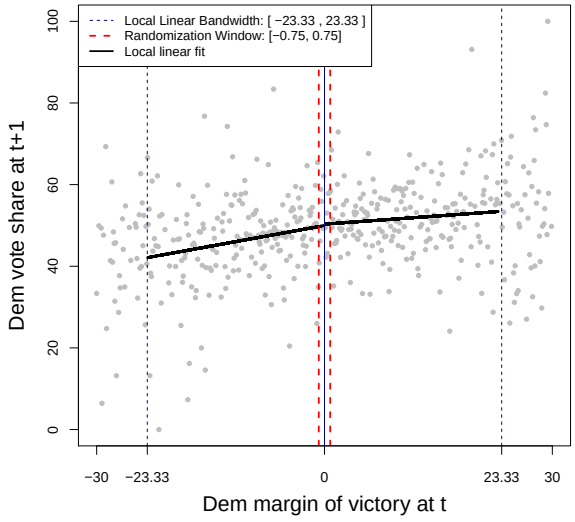


(a) Local Linear Estimation

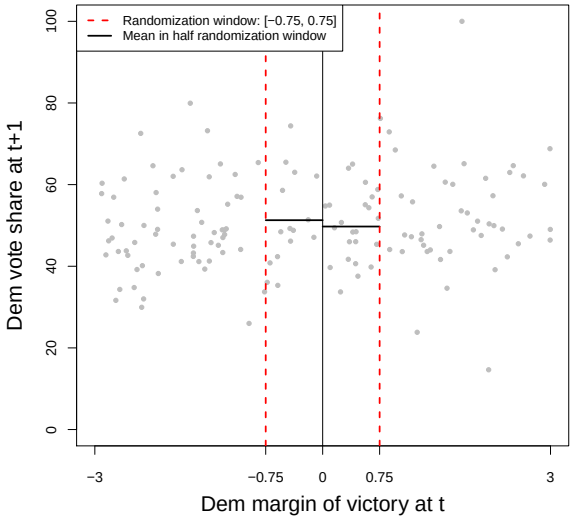


(b) Randomization-based Estimation

DESIGN II: DEMOCRATIC VOTE SHARE AT $t+1$



(c) Local Linear Estimation



(d) Randomization-based Estimation

1 holds inside our chosen window, it also must hold in all windows contained in it. Thus, analyzing the smaller window $[-0.5, 0.5]$ can provide evidence on whether there is heterogeneity in the results found in the originally chosen window.

Second, we perform the test of the sharp null using different test-statistics. Under Assumption 1, there is no relationship between the outcome and the score on either side of the threshold within W_0. In this situation, performing randomization-based tests using the difference-in-means as a test statistic should yield the same results as using other test-statistics that allow for a relationship between the conditional regression function and the score. This suggests using different test-statistics in the same window as a robustness check. In a similar spirit to conventional parametric and non-parametric RD methods, we consider four different additional tests-statistics. Let the window considered be $[w_l, w_r]$ and the recall that the cutoff is r_0. For each window, the four test-statistics considered are the following: the difference in the estimated intercepts of regressions of Y_i on $R_i - r_0$ on either side of the cutoff (that is, the standard local linear estimator), the difference in the predicted values $\hat{Y}_i$ of a regression of Y_i on $R_i - r_0$ evaluated at the midpoints of the half-windows to the left and right of the cutoff ($R_i = (r_0 - w_l)/2$ and $R_i = (w_r - r_0)/2$, respectively), the difference in the estimated intercepts of a regression of Y_i on $R_i - r_0$ and $(R_i - r_0)^2$, and the difference in the predicted values $\hat{Y}_i$ of a regression of Y_i on $R_i - r_0$ and $(R_i - r_0)^2$ also evaluated at the midpoints of the half-windows to the left and right of the cutoff. Below, we call the p-values based on these test-statistics 'p-value linear-cutoff, 'p-value linear-midpoint', 'p-value quadratic-cutoff' and 'p-value quadratic-midpoint', respectively.

Table 4 presents the results from our sensitivity analysis. Panel A shows results for Democratic Vote Share at $t+2$ (Design I), and Panel B for for Democratic Vote Share at $t+1$ (Design II). For each panel, we reproduce the results in our chosen $[-0.75, 0.75]$ window, and show results for the three additional windows mentioned above: $[-0.5, 0.5]$, $[-1.0, 1.0]$ and $[-2.0, 2.0]$. In all cases, the point estimate (Hodghes-Lehmnan estimate), treatment effect confidence interval (obtained inverting hypothesis tests based on constant treatment effect model), and the quantile treatment effect confidence intervals are calculated as in Table 3. The 'p-value diffmeans' is equivalent to the p-value reported in Table 3, which corresponds to a test of the sharp null hypothesis based on the difference-in-means test statistic. The four additional p-values reported correspond to a test of the sharp-null hypothesis based on the four test-statistics described above. All p-values less than or equal to 0.05 are shown in bold in the table.

There are important differences between our two outcomes. The results in Design I (Panel A) are remarkably robust to the choice of the test-statistic in the originally chosen $[-0.75, 0.75]$ window and in the smaller $[-0.50, 0.50]$ window. The results are also insensitive to increasing the window, as seen in the last two columns of Panel A. In contrast, the null results found in Design II seem more fragile. First, the sharp

null hypothesis is often rejected in larger windows when alternative test-statistics are considered, as shown in the last two columns in Panel B. Second, in our chosen window, the sharp null hypothesis is rejected with three of the four additional test-statistics considered. As we showed before in Table 3 and reproduce in Table 4, this does not translate into a statistically significant constant or quantile treatment effect – all confidence intervals are roughly centered around zero. An interesting phenomenon that might explain these rejections occurs when we consider the smaller $[-0.5, 0.5]$ window. In this window, the point estimate and confidence intervals show a negative effect and provide support for the opposite-party advantage hypothesis. The Hodges-Lehmann point estimate is about -8 percentage points, more than a 10-fold increase in absolute value with respect to the conventional estimates, and we reject the sharp null hypothesis of no effect at 5% level with three of the five different test-statistics considered. Our randomization-based confidence interval of the constant treatment effect ranges from -16.65 to -0.04, ruling out a non-negative effect. The confidence interval for the 25th quantile treatment effect also excludes zero and again provides support for the opposite party advantage.

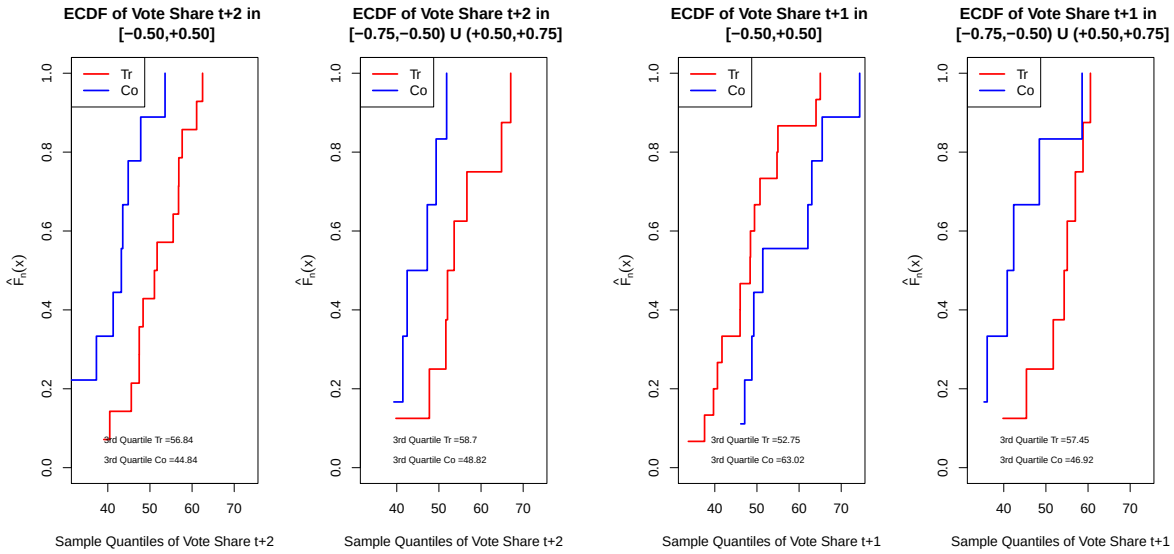
As mentioned above, if Assumption 1 holds in our chosen window it must hold in a smaller window, which implies that these results are in principle valid. To investigate this issue further, Figure 4 plots the empirical cumulative distribution functions (ECDF) of our two outcomes in two different windows: the small $[-0.5, 0.5]$ window and the window defined by $[-0.75, -0.50) \cup (0.5, 0.75]$. The union of these two windows is our chosen $[-.75, 0.75]$ window. Figure 4(a) shows that for Democratic Vote Share $t+2$, the ECDF of the treatment group is shifted to the right of the ECDF of the control group everywhere in both windows, showing that the treated quantiles are larger than the control quantiles. Since the treated outcome dominates the control outcome in both windows, combining the observations into our chosen window produces the robust incumbent-party advantage results that we see in the first two columns of Table 4. In contrast, for Democratic Vote Share $t+1$, the outcome in Design I, the smaller $[-0.5, 0.5]$ window exhibits a very different pattern from that observed for the observations that are added to it when our chosen window is considered. The left plot in Figure 4(b) shows that the ECDF of the control group is shifted to the right of the ECDF of the treatment group everywhere, showing support for the negative effect (opposite-party advantage) reported in the first column of Table 4. But the right plot in Figure 4(b) shows that this situation reverses in the window $[-0.75, -0.50) \cup (0.5, 0.75]$, where treated quantiles are larger than control quantiles almost everywhere. The combination of the observations in both windows is what produces the null effects in our chosen $[-.75, 0.75]$ window.

The results in $[-0.5, 0.5]$ suggest some support for the opposite-party advantage, and show that our chosen window combines possibly heterogeneous treatment effects for Vote Share $t+1$ – but not for Vote Share $t+2$. It is beyond the scope of this article to determine whether this limited empirical evidence

supporting the opposite party advantage reflects a true underlying effect or is simply the result of random noise. We note, however, that the phenomenon of heterogeneity illustrated in Figure 4(b) could arise because we are trying to detect a *negative* treatment effect when the slope of the conditional regression function of Democratic Vote Share $t+1$ is still *positively* related to the Democratic margin of victory at t (our running variable), as shown in Figure 1 above. In this situation, if the chosen window is slightly larger than it should be, the negative effect would quickly become null.

In sum, our sensitivity and robustness analysis in this section show that the incumbent-party advantage results are remarkably robust but our opposite-party advantage results are more fragile and suggest some avenues for future research.

Figure 4: Empirical CDFs of outcomes for treated and control in different windows – U.S. Senate Elections, 1914-2010



(a) Democratic Vote Share at $t+2$ (b) Democratic Vote Share at $t+1$

Table 4: Sensitivity of randomization-based RD results: incumbent-party and opposite-party advantages in the U.S. Senate in different window choices

	A. Design I (outcome = Dem Vote Share at t+2)			
	Smaller window	Chosen Window	Larger Windows	
Window	$[-0.50, 0.50]$	$[-0.75, 0.75]$	$[-1.00, 1.00]$	$[-2.00, 2.00]$
Point estimate	10.16	9.32	9.61	8.90
p-value diffmeans	**0.0042**	**0.0006**	**0.0000**	**0.0000**
p-value linear-cutoff	**0.0001**	**0.0000**	**0.0000**	**0.0000**
p-value linear-midpoint	**0.0006**	**0.0000**	**0.0002**	**0.0000**
p-value quadratic-cutoff	**0.0080**	**0.0000**	**0.0000**	**0.0002**
p-value quadratic-midpoint	0.1187	**0.0000**	**0.0000**	**0.0000**
Treatment effect CI	[3.61 , 17.18]	[4.56 , 14.84]	[5.81 , 15.16]	[6.33 , 14.04]
.25-QTE CI	[-2.75 , 19.42]	[-2.00 , 21.12]	[4.13 , 21.25]	[4.88 , 18.57]
.75-QTE CI	[1.93 , 17.87]	[3.68 , 18.94]	[1.78 , 17.53]	[0.42 , 13.69]
Sample size treated	14	22	25	47
Sample size control	9	15	18	49
	B. Design II (outcome = Dem Vote Share at t+1)			
	Smaller window	Chosen Window	Larger Windows	
Window	$[-0.50, 0.50]$	$[-0.75, 0.75]$	$[-1.00, 1.00]$	$[-2.00, 2.00]$
Point estimate	-8.17	-0.79	2.32	0.56
p-value diffmeans	**0.0426**	0.6253	0.5138	0.7311
p-value linear-cutoff	0.6502	**0.0000**	**0.0000**	0.2904
p-value linear-midpoint	0.9486	**0.0000**	**0.0000**	0.9290
p-value quadratic-cutoff	**0.0129**	0.7256	0.4724	0.0610
p-value quadratic-midpoint	**0.0000**	**0.0000**	**0.0107**	**0.0000**
Treatment effect CI	[-16.65 , -.04]	[-8.11 , 5.05]	[-4.99 , 9.71]	[-3.89 , 5.57]
.25-QTE CI	[-13.82 , -.16]	[-8.75 , 9.96]	[-8.63 , 14.65]	[-4.14 , 4.85]
.75-QTE CI	[-25.92 , 12.63]	[-11.15 , 11.31]	[-10.72 , 16.23]	[-8.26 , 12.63]
Sample size treated	15	23	27	50
Sample size control	9	15	18	49

Notes: Results based on U.S. Senate elections from 1914 to 2010. Point estimate is from Hodges-Lehmann estimator. Treatment effect confidence intervals are obtained by inverting a test of a constant treatment effect model, using the difference-in-means test statistic and assuming a fixed margins randomization mechanism. P-values are randomization-based and correspond to a test of the sharp null hypothesis of no treatment effect, assuming a fixed margins randomization mechanism. Each p-value reported corresponds to a test based on a different test-statistic: 'p-value diffmeans' uses the difference-in-means; 'p-value linear-cuttoff' uses the difference in intercepts in a fitted linear polynomial of the outcome on the score; 'p-value linear-midpoint' uses same statistic as 'p-value linear-cutoff', except that predicted values are calculated at the midpoint of the window on either side of cutoff; 'p-value quadratic-cuttoff' uses the difference in intercepts in a fitted quadratic polynomial of outcome on score; 'p-quadratic linear-midpoint' uses the same statistic as 'p-value quadratic-cutoff', except that predicted values are calculated at the midpoint of the window on either side of cutoff. CI denotes 95% confidence intervals (CI). The quantities '.25-QTE CI' and '.75-QTE CI' denote the 95% CI for the 25th-quantile and 75th-quantile treatment effects, respectively, and are constructed as described in the text.

6 Extensions, Applications and Discussion

We introduced a framework to analyze regression discontinuity designs employing a "local" randomization approach and proposed using randomization inference techniques to conduct finite-sample exact inference. Our methodological approach can be extended to a variety of empirically relevant contexts. In this section, we discuss four natural extensions focusing on fuzzy RD designs, discrete-valued and multiple running variables and matching techniques, among other possibilities. In addition, the last Subsection 6.5 discusses the connection between our approach and the conventional large-sample RD approach available in the literature.

6.1 Fuzzy RD with Possibly Weak Instruments

In the sharp RD design, treatment assignment is equal to $Z_i = \mathbb{1}(R_i \geq r_0)$, and treatment assignment is equal to actual treatment status. In the fuzzy design, treatment status D_i, with observations collected in n-vector $\mathbf{D}$ as above, is not completely determined by placement relative to r_0, so D_i may differ from Z_i. Our framework extends directly to the fuzzy RD designs and, as we discuss in more detail below, it also offers a robust inference alternative to the traditional approaches when the instrument (i.e., the relationship between D_i and Z_i) is regarded as "weak".

Let $d_i(\mathbf{r})$ be unit i's potential treatment status when the vector of scores is $\mathbf{R} = \mathbf{r}$. Similarly, we let $y_i(\mathbf{r}, \mathbf{d})$ be unit i's potential outcome when the vector of scores is $\mathbf{R} = \mathbf{r}$ and the treatment status vector is $\mathbf{D} = \mathbf{d}$. Observed treatment status and outcomes are $D_i = d_i(\mathbf{R})$ and $Y_i = y_i(\mathbf{R}, \mathbf{D})$. This generalization leads to a framework analogous to an experiment with non-compliance, where Z_i is used as an instrument for D_i and randomization-based inferences are based on the distribution of Z_i. Assumption 1 generalizes as follows.

Assumption 1′: 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) $F_{R_i|R_i \in W_0}(r) = F(r)$, and

(b) $d_i(\mathbf{r}) = d_i(\mathbf{z}_{W_0})$ and $y_i(\mathbf{r}, \mathbf{d}) = y_i(\mathbf{z}_{W_0}, \mathbf{d}_{W_0})$ for all $\mathbf{r}, \mathbf{d}$.

This assumption permits testing the null hypothesis of no effect exactly as described above, although the interpretation of the test differs, as now it can only be considered a test of no effect of treatment among those units whose potential treatment status $d_i(\mathbf{z}_{W_0})$ varies with $\mathbf{z}_{W_0}$.

Constructing confidence intervals and point estimates in the fuzzy design requires generalizing Assumption 2 and introducing an additional assumption.

Assumption 2′: Local SUTVA (LSUTVA). For all i with $R_i \in W_0$:

(a) If $z_i = \tilde{z}_i$, then $d_i(\mathbf{z}_{W_0}) = d_i(\tilde{\mathbf{z}}_{W_0})$, and

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

Assumption 6: Local Exclusion Restriction. For all i with $R_i \in W_0$: $y_i(\mathbf{z}, \mathbf{d}) = y_i(\tilde{\mathbf{z}}, \mathbf{d})$ for all $(\mathbf{z}, \tilde{\mathbf{z}})$ and for all $\mathbf{d}$.

Assumption 6 means potential responses depend on placement with respect to the threshold only through its effect on treatment status. Under assumptions $1'$–$2'$ and Assumption 6, we can write potential responses within the window as $y_i(\mathbf{z}, \mathbf{d}) = y_i(d_i)$. Furthermore, under the constant treatment effect model in Assumption 3, estimation and inference proceeds exactly as before, but defining the adjusted responses as $\mathbf{Y}_{W_0} - \tau_0 \mathbf{D}_{W_0}$. Inference on quantiles in the fuzzy design is more involved, and requires an additional monotonicity assumption (e.g., Frandsen, Frölich, and Melly, 2012).

Fuzzy RD designs are local versions of the usual instrumental variables (IV) model and thus concerns about weak instruments may arise in this context as well (Marmer, Feir, and Lemieux, 2012). Our randomization inference framework, however, circumvents this concern because it enable us to conduct exact finite-sample inference, as discussed in Imbens and Rosenbaum (2005) for the usual IV setting. Therefore, this paper also offers an alternative, robust inference approach for fuzzy RD designs under possibly weak instruments.

6.2 Discrete and Multiple Running Variables

Another important feature of our framework is that it can handle RD settings where the running variable is not univariate and continuous. Our results provide an alternative inference approach when the running variable is discrete or has mass points in its support (see, for example, Lee and Card, 2008). While conventional, nonparametric smoothing techniques are usually unable to handle this case without appropriate technical modifications, our randomization inference approach applies immediately to this case and offers the researcher a fully data-driven approach for inference when the running variable is not continuously distributed.

Our approach also extends naturally to settings where multiple running variables are present or, equivalently, to the so-called multiple RD design case (see Keele and Titiunik (2013) and references therein). For example, in the case of geographic RD designs, which involves two running variables, Keele, Titiunik, and Zubizarreta (2014) discuss how the methodological framework introduced herein can be used to conduct inference employing both geographic variation and matching on observables techniques.

6.3 Matching and Parametric Modeling

Conventional approaches to RD employ continuity of the running variable and large sample approximations, and typically do not emphasize the role of covariates and parametric modeling, relying instead on nonparametric smoothing techniques local to the discontinuity. However, in practice, researchers often incorporate covariates and employ parametric models in a "small" neighborhood around the cutoff when conducting inference. Our framework gives a formal justification (i.e., "local randomization") and an alternative inference approach (i.e., randomization inference) for this common empirical practice. For example, our approach can be easily extended to justify (finite-sample exact) inference in RD contexts using panel or longitudinal data, specifying nonlinear models or, perhaps, relying on flexible "matching" on covariates techniques. For a recent example of such an approach see Keele et al. (2014).

6.4 Sensitivity Analysis and Related Techniques

In the context of randomization-based inference, a useful tool to asses the plausibility of the results is a sensitivity analysis that considers how the results vary under deviations from the randomization assumption. Rosenbaum (2002, 2010) provides details of such an approach when the treatment is assumed to be randomly assigned conditionally on covariates. Under a randomization-type assumption, the probability of receiving treatment is equal for treated and control units; a sensitivity analysis proposes a model for the odds of receiving treatment and allows the probability of receiving treatment to differ between groups and recalculates the p-values, confidence intervals or point estimates of interest. The analysis asks whether small departures from the randomization-type assumption would alter the conclusions from the study. If, for example, small differences in the probability of receiving treatment between treatment and control units lead to markedly different conclusions (i.e., if the null hypothesis of no effect is initially rejected but then ceases to be rejected), then we conclude that the results are sensitive and appropriately temper our confidence in the results. This kind of analysis could be directly applied in our context inside W_0. In this window, our assumption is that the probability of receiving treatment is equal for all units (and that we can estimate such probability); thus, a sensitivity analysis of this type could be applied directly to establish whether our conclusions survive under progressively different probabilities of receiving treatment for treated and control units inside W_0.

6.5 Connection to Standard RD Setup

Our finite-sample RD inference framework may be regarded as an alternative approximation to the conventional RD identifying conditions in Hahn et al. (2001). This section defines a large-sample identification framework similar to the conventional one and discusses its connection to the finite-sample Assumption 1.

In the conventional RD setup, individuals have *random* potential outcomes $Y_i(r, d)$ which depend on

the value of a running variable, $r \in \mathbb{R}$, and treatment status $d \in \{0,1\}$. The observed outcome is $Y_i \equiv Y_i(R_i, D_i)$, and identification is achieved by imposing continuity, near the cutoff r_0, on $\mathbb{E}[Y_i(r,d)|R_i = r]$ or $F_{Y_i(r,d)|R_i=r}(y) = \Pr[Y_i(r,d) \leq y|R_i = r]$. Consider the following alternative sufficient identifying condition.

Assumption 7: Conventional RD Assumption. For all $d \in \{0,1\}$ and $i = 1, 2, \cdots, n$:

(a) R_i is continuously distributed,

(b) $Y_i(r,d)$ is Lipschitz continuous in r at r_0,

(c) $F_{Y_i(r_0,d)|R_i=r}(y) = \Pr[Y_i(r_0,d) \leq y|R_i = r]$ is Lipschitz continuous in r at r_0.

Part (a) requires a continuous running variable, part (b) ensures there are no other confounding effects on the outcome at the threshold, and part (c) ensures there is no sorting based on potential outcomes around the threshold. These conditions are very similar to those in Hahn et al. (2001) and other (large-sample-type) approaches to RD. The main difference is that we require continuity of potential outcome functions, as opposed to just continuity of the conditional expectation of potential outcomes. Continuity rules out knife-edge cases where confounding differences in potential outcomes at the threshold (that is, discontinuities in $Y_i(r,d)$) exactly offset sorting in the running variable at the threshold so that the conditional expectation of potential outcomes is still continuous at the threshold. In ruling out this knife-edge case, our condition is technically stronger, but arguably not stronger in substance, than conventional identifying conditions.

Estimation and inference are based on approximations to the underlying identifying conditions in both the conventional large-sample approach and our finite-sample approach. The conventional approach approximates the conditional distribution of outcomes near the threshold as locally linear, and relies on large-sample asymptotics for inference. Our approach proposes an alternative local constant approximation and uses finite-sample inference techniques. The local linear approximation may be more accurate than local constant farther from the threshold but the large-sample sample approximations may be poor. The local constant approximation will likely be appropriate only very near the threshold, but the inference will remain valid for small samples.

The following result shows that the finite-sample condition in Assumption 1 (Local Randomization) can be seen as an approximation obtained from the more-conventional RD identifying conditions given in Assumption 7, with an approximation error that is controlled by the window width.

Result 1: Connection between RD frameworks. Suppose Assumption 7 holds. Then:

(i) $F_{R_i|R_i\in[\underline{r},\overline{r}],Y_i(r_0,d)}(r) = F_{R_i|R_i\in[\underline{r},\overline{r}]}(r) + O(\overline{r} - \underline{r})$, and

(ii) $Y_i(r,d) = Y_i(r_0,d) + O(\overline{r} - \underline{r})$.

Part (i) of this result says that the running variable is approximately independent of potential outcomes near the threshold, or, in the finite-sample framework where potential outcomes are fixed, each unit's running

variable has approximately the same distribution (under i.i.d. sampling). This corresponds to part (a) of Assumption 1 (Local Randomization), and gives a formal connection between the usual RD framework and our randomization-inference framework. Similarly, part (ii) implies that potential outcomes depend approximately on treatment status only near the threshold r_0, as assumed in Assumption 1(b).

7 Conclusion

Motivated by the interpretation of regression discontinuity designs as local experiments, we proposed a randomization-inference framework to conduct exact finite-sample inference in this design. Our approach is especially useful when only a few observations are available in the neighborhood of the cutoff where local randomization is plausible. Our randomization-based methodology can be used both for validating (and even selecting) this window around the RD threshold and performing statistical inference about the effects in this window. Our analysis of party-level advantages in U.S. Senate elections illustrated our methodology and showed that a randomization-based analysis can lead to different conclusions than standard RD methods based on large-sample approximations.

We envision our approach as complementary to existing parametric and non-parametric methods for the analysis of RD designs. Employing our proposed methodological approach, scholars can provide evidence about the plausibility of the as-good-as-random interpretation of their RD designs, and also conduct exact finite-sample inference employing only those few observations very close to the RD cutoff. If even in a small window around the cutoff the sharp null hypothesis of no effect can be rejected for predetermined covariates, scholars should not rely on the local randomization interpretation of their designs, and hence should pay special attention to the plausibility of the continuity assumptions imposed by the standard approach.

References

Abramowitz, A. (1980): "A comparison of voting for u.s. senator and representative in 1978," *American Political Science Review*, 74, 633–640.

Alesina, A., M. Fiorina, and H. Rosenthal (1991): "Why are there so many divided senate delegations?" Working Paper 3663, National Bureau of Economic Research, URL `http://www.nber.org/papers/w3663`.

Ansolabehere, S., J. M. Hansen, S. Hirano, and J. M. Snyder (2007): "The incumbency advantage in u.s. primary elections," *Electoral Studies*, 26, 660–668.

Ansolabehere, S. and J. M. Snyder (2002): "The incumbency advantage in u.s. elections: An analysis of state and federal offices, 1942-2000," *Election Law Journal: Rules, Politics, and Policy*, 1, 315–338.

Barrios, T., R. Diamond, G. W. Imbens, and M. Kolesar (2012): "Clustering, spatial correlations and randomization inference," *Journal of the American Statistical Association*, 107, 578–591.

Black, D. A., J. Galdo, and J. A. Smith (2007): "Evaluating the bias of the regression discontinuity design using experimental data," *American Economic Review Papers and Proceedings*, 97, 104–107.

Buddelmeyer, H. and E. Skoufias (2003): "An evaluation of the performance of regression discontinuity design on progresa," *IZA Discussion Paper 827*.

Butler, D. and M. Butler (2006): "Splitting the difference? causal inference and theories of split-party delegations," *Political Analysis*, 14, 439–455.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014a): "Robust data-driven inference in the regression-discontinuity design," revision requested by *Stata Journal*.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014b): "Robust nonparametric confidence intervals for regression-discontinuity designs," working paper, University of Michigan.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014c): "`rdrobust`: An r package for robust inference in regression-discontinuity designs," in preparation for *Journal of Statistical Software*.

Caughey, D. and J. S. Sekhon (2011): "Elections and the regression-discontinuity design: Lessons from close u.s. house races, 1942–2008," *Political Analysis*, 19, 385–408.

Collier, K. and M. Munger (1994): "A comparison of incumbent security in the house and senate," *Public Choice*, 78, 145–154.

Cook, T. D. (2008): ""waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics," *Journal of Econometrics*, 142, 636–654.

Cook, T. D., W. R. Shadish, and V. C. Wong (2008): "Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons," *Journal of Policy Analysis and Management*, 27, 724–750.

Craiu, R. V. and L. Sun (2008): "Choosing the lesser evil: Trade-off between false discovery rate and non-discovery rate," *Statistica Sinica*, 18, 861–879.

Diamond, A. and J. S. Sekhon (2013): "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies," *Review of Economics and Statistics*, 95, 932–945.

Dinardo, J. and D. S. Lee (2011): "Program evaluation and research designs," in O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics*, volume 4A, Elsevier Science B.V., 463–536.

Efron, B. (2010): *Large-Scale Inference*, Cambridge, UK: Cambridge.

Erikson, R. and R. Titiunik (2013): "Using regression discontinuity to uncover the personal incumbency advantage," working paper, University of Michigan.

Erikson, R. S. (1971): "The advantage of incumbency in congressional elections," *Polity*, 3, 395–405.

Frandsen, B., M. Frölich, and B. Melly (2012): "Quantile treatments effects in the regression discontinuity design," *Journal of Econometrics*, 168, 382–395.

Gelman, A. and G. King (1990): "Estimating incumbency advantage without bias," *American Journal of Political Science*, 34, 1142–1164.

Green, D. P., T. Y. Leong, H. Kern, A. S. Gerber, and C. W. Larimer (2009): "Testing the accuracy of regression discontinuity analysis using experimental benchmarks," *Political Analysis*, 17, 400–417.

Hahn, J., P. Todd, and W. van der Klaauw (2001): "Identification and estimation of treatment effects with a regression-discontinuity design," *Econometrica*, 69, 201–209.

Hansen, B. B. and J. Bowers (2009): "Attributing effects to a cluster randomized get-out-the-vote campaign," *Journal of the American Statistical Association*, 104, 873–885.

Ho, D. E. and K. Imai (2006): "Randomization inference with natural experiments: An analysis of ballot effects in the 2003 election," *Journal of the American Statistical Association*, 101, 888–900.

Holland, P. W. (1986): "Statistics and causal inference," *Journal of the American Statistical Association*, 81, 945–960.

Imbens, G. and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142, 615–635.

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

Imbens, G. W. and P. Rosenbaum (2005): "Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education," *Journal of the Royal Statistical Society, Series A*, 168, 109–126.

Jung, G.-R., L. W. Kenny, and J. R. Lott (1994): "An explanation for why senators from the same state vote differently so frequently," *Journal of Public Economics*, 54, 65–96.

Keele, L. and R. Titiunik (2013): "Geographic boundaries as regression discontinuities," working paper, University of Michigan.

Keele, L., R. Titiunik, and J. Zubizarreta (2014): "Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout," *Journal of the Royal Statistical Society, Series A*, forthcoming.

Krasno, J. (1994): *Challengers, Competition, and Re-election*, New Haven: Yale University Press.

Lee, D. S. (2008): "Randomized experiments from non-random selection in u.s. house elections," *Journal of Econometrics*, 142, 675–697.

Lee, D. S. and D. Card (2008): "Regression discontinuity inference with specification error," *Journal of Econometrics*, 142, 655–674.

Lee, D. S. and T. Lemieux (2010): "Regression discontinuity designs in economics," *Journal of Economic Literature*, 48, 281–355.

Lehmann, E. L. (2006): *Nonparametrics: Statistical Methods Based on Ranks*, New York: Springer.

Lublin, D. I. (1994): "Quality, not quantity: Strategic politicians in u.s. senate elections, 1952-1990," *Journal of Politics*, 56, 228–241.

Marmer, V., D. Feir, and T. Lemieux (2012): "Weak identification in fuzzy regression discontinuity designs," working paper, Univeristy of British Columbia.

McCrary, J. (2008): "Manipulation of the running variable in the regression discontinuity design: A density test," *Journal of Econometrics*, 142, 698–714.

Porter, J. (2003): "Estimation in the regression discontinuity model," working paper, University of Wisconsin.

Rosenbaum, P. R. (2002): *Observational Studies*, New York: Springer, 2nd edition.

Rosenbaum, P. R. (2010): *Design of Observational Studies*, New York: Springer.

Segura, G. M. and S. P. Nicholson (1995): "Sequential choices and partisan transitions in u.s. senate delegations: 1972–1988," *Journal of Politics*, 57, 86–100.

Senate Historical Office, U. S. (2012): "Direct election of senators," U.s. senate, U.S. Senate, `http://www.senate.gov/artandhistory/history/common/briefing/Direct_Election_Senators.htm`.

Thistlethwaite, D. L. and D. T. Campbell (1960): "Regression-discontinuity analysis: An alternative to the ex-post facto experiment," *Journal of Educational Psychology*, 51, 309–317.

Wellek, S. (2010): *Testing Statistical Hypotheses of Equivalence and Noninferiority*, Boca Raton, FL: Chapman & Hall/CRC, 2nd edition.