Regression Discontinuity in Practice: Straightforward Solutions to Common Problems

Andrew Bertoli

19 September 2017

Although regression discontinuity (RD) is a very simple research design in theory, many applications are riddled with subtle complexities, even when treatment assignment is random close to the cut-point. In this paper, I examine three of the most common problems for regression discontinuity designs: (1) variation in competitiveness across strata, (2) units that appear in the sample multiple times, and (3) independent variables of interest that are not actually assigned by the RD. In each case, I explain the problem formally, offer straightforward solutions, and illustrate the main points with simulations and real-world data.

Over the last decade, regression discontinuity (RD) has gained popularity in many scientific fields. Not only is it one of the most trusted tools of causal inference short of an experiment (Bernardi and Skoufias 2004; Green et al. 2009), but its similarity to a real experiment allows researchers to present their findings to other scholars and the general public in a very compelling way (Hopkins and McCabe 2012; Hall 2015). Given its advantages, the popularity of the regression discontinuity design is likely to continue to grow for the foreseeable future, especially considering the prevalence of scoring systems with cut-points. In recent years, it has even spread to fields like sociology (Rao, Yue, and Ingram 2011; Legewie 2013; Bernardi 2014), international relations (Voeten 2013; Bertoli, Dafoe, and Trager 2017), peace and conflict studies (Crost, Felter, and Johnston 2014), and criminology (Chen and Shapiro 2007; Vollaard 2009).

But despite RD's virtues, applications often face subtle problems that can pose serious threats to inference. The concern that tends to receive the most attention is that certain units close to the cut-point may be able to manipulate their scores and sort to one side or the other (Imbens and Lemieux 2008; Caughey and Sekhon

2011; Eggers et al. 2014). However, even when units cannot sort, other complications can still occur that are just as problematic. These issues usually arise either from the complex rules of the scoring system or because researchers want to make certain comparisons that are not necessarily guaranteed to be valid by the RD.

While applied researchers have made important progress in addressing these problems, there are cases where they have confused the issues, perhaps sometimes simply because they wanted to avoid undesired complexity in their articles. No doubt, students hoping to learn from past examples of RD could easily be misguided by the choices made in some of the most important applications. Unfortunately, there have only been a few attempts to clarify and resolve the complexities that can occur in real world situations (Keele and Titiunik 2014; Cattaneo et al. 2016), and some of the most widespread issues have not been explored in scholarly writing.

In this essay, I investigate three of the most important problems that arise in applied RD, all of which have been largely unaddressed by methodologists. The first occurs when units are organized into strata that vary in terms of their competitiveness. For instance, the units could be candidates running in different districts, and the districts could vary in terms of number of candidates with a realistic chance of winning. In these cases, researchers must be careful which units they include in their sample, or else they risk inducing bias. The second problem occurs when units appear in the sample multiple times, such as when students retake a test or candidates rerun for office. This scenario can cause some units to have multiple scores and possibly sort from one side of the cut-point to the other. The third arises when the independent variable of interest is not actually assigned by the RD. For instance, researchers might want to estimate how the likelihood of international conflict changes when a woman barely defeats a man in an election for head-of-state, even though gender is a basic human characteristic that could never be "as-if" randomly assigned to politicians. This issue can make it difficult to interpret the treatment effect.

Throughout this paper, I often discuss regression discontinuity as though it were a randomize experiment, where the treatment is assigned randomly to units within a certain distance of the cut-point (Cattaneo, Frandsen, and Titiunik 2014). This approach differs from the more common practice of treating units' scores as fixed and estimating the difference at the cut-point with two regression lines. However, the points that I make are all equally applicable to this second approach, which is actually very similar to the natural experiment approach, at least mathematically. I discuss the relationship between these two approaches much more in the next section. The reason that I commonly use the natural experiment approach in this paper is that it makes the problems and solutions that I discuss much clearer.

This paper proceeds as follows: I first lay out the basics of regression discontinuity, including the notation and key concepts, which will serve as the basis of this paper. I also compare the two main approaches to regression discontinuity, highlighting their mathematical similarities. This will clarify how the problems that I raise for one approach will translate into problems for the other. The next three sections are devoted to explaining the three problems, both theoretically and through examples. I also outline solutions that can be used to resolve these problems when possible. The final section concludes.

Section 1: The Basics of Regression Discontinuity

Regression discontinuity is a quasi-experimental research design that allows researchers to estimate causal effects by exploiting scoring systems with important thresholds. The classic example is an exam where every student who scores above a cut-point receives an award and every students who scores below does not (Thistlethwaite and Campbell 1960). The idea is that receiving the award should be close to random for the students who barely surpassed or barely fell short of the cut-point. Thus, the cut-point creates a source of exogeneity that allows researchers to test how the treatment affected the units.

In many cases, all of the units that scored above the threshold received the treatment, while all of the units that scored below did not. These situations are referred to as sharp RDs. However, this clean set-up is not always necessary. We can still estimate the causal effect if some of the units that surpassed the cut-point did not receive the treatment or some units that fell short did. The key is that scoring above the cut-point cannot lower any unit's probability of receiving the treatment, and it must increase the probability that at least some units received the treatment. In these situations, surpassing the cut-point can be thought of as an instrument that increases the probability of treatment assignment for at least some units. These cases are called fuzzy RDs, and they can be analyzed using a combination of regression discontinuity and instrumental variable techniques.

To make this paper more accessible to readers, I write out the notation in terms of the conventional (sharp) RD design where all units to the right of the cut-point received the treatment and all units to the left did not. However, the problems and solutions that I discuss in this paper also apply to fuzzy RDs in a fairly straightforward way, and I will clarify these connections at various points. The key insight to understanding these connections is that the fuzzy RD estimator is calculated by taking the Intention to Treat (ITT) estimator (the effect of surpassing the cut-point on the outcome) and dividing by the estimated effect of surpassing the cut-point on the probability of receiving the treatment. In other words,

$$\text{Fuzzy RD Estimator}=\frac{\text{Effect of surpassing the cut-point on Y}}{\text{Effect of surpassing the cut-point on T}}$$

Thus, the fuzzy RD estimator is just the quotient of two sharp RD estimators. This makes problems for sharp RDs very relevant to fuzzy RDs.

There are two ways that researchers analyze regression discontinuities. I describe each of them below and then discuss their similarities and differences.

The Natural Experiment Set-up. This approach treats regression discontinuity as a natural experiment where the treatment is assigned randomly to units within a certain distance of the cut-point. Under this set-up, analysis is very straightforward. First, researchers must choose a size for their regression discontinuity window. Ideally, they make this decision prior to looking at the outcomes, and they are normally expected to report the results for other reasonable window choices.

Next, researchers focus their attention on the units inside the RD window, excluding the others so that they do not effect the results. This move is valid even if it was somewhat random which units ended up in the RD window. It is similar to a situation where units were selected to be in an experiment in a semi-random manner. Which units ended up in the experiment could be taken as fixed, and then inferences could be drawn about that sample that would be internally valid.

Once the units inside the RD window are identified, researchers treat them as though they were part of a real experiment. The standard notation is as follows. Denote the number of units as n. Unit i's treatment status is $T_i \in \{0, 1\}$ and its score as Z_i. Under the assumption of non-interference between units, each unit has two potential outcomes, one under treatment (Y_{it}) and the other under control (Y_{ic}). The parameter that we are interested in is the Local Average Treatment Effect (LATE), which is written as

$$\bar{\tau}_{RD} = \frac{1}{n} \sum_i (Y_{it} - Y_{ic})$$

This parameter is just the average treatment effect for units inside the RD window. If m is the number of treated units, then the estimator is

$$\hat{\tau}_{RD} = \frac{1}{m} \sum_i Y_{it} T_i - \frac{1}{n-m} \sum_i Y_{ic} (1 - T_i)$$

which is simply the difference in means between the treated and control units inside the RD window. As in a normal experiment, the standard error of this estimator is typically approximated by bootstrapping or by using the formula

$$\hat{SE} = \sqrt{\frac{\hat{\sigma}_t^2}{m} + \frac{\hat{\sigma}_c^2}{n-m}}$$

where $\hat{\sigma}_t^2$ and $\hat{\sigma}_c^2$ are the estimated variances of the treated and control units. These are each calculated using the standard sample-to-population variance formula

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_i (x_i - \bar{x})^2$$

The p-value can be approximated using the estimated standard error, or it can be computed exactly using permutation inference (for the sharp null hypothesis of no treatment effect).

As with real experiments, it is possible to decrease the standard errors and reduce bias from baseline differences between the two groups by controlling for covariates that are predictive of the outcome. For instance, we could run a regression on the sample or do difference-in-differences estimation, as is commonly done with experimental data. However, these adjustment methods are usually considered robustness checks rather than valid procedures to get the main results. The reason is that allowing researchers to control for whatever covariates they want makes it possible for them to manipulate their results so that they can get lower p-values (Masicampo and Lalande 2012).

However, when analyzing a regression discontinuity, there is one additional reason for performing covariate adjustment that does not apply to real experiments. In experiments, the treatment is randomized, making the unadjusted difference in means estimator unbiased. On the other hand, regression discontinuity gives us two groups that were not actually randomized. For instance, if we are comparing leaders who won or lost their elections by less than 100 votes, we should expect the bare winners to differ from the bare losers in small but systematic ways. We might expect the bare winners to be on average slightly higher quality, wealthier, or smarter than the bare losers, and if we had a large enough sample size we would actually be able to pin-point these differences using difference in means tests.

The most direct way to account for these systematic differences is to control for the score, Z, using some linear or non-parametric model. This step should reduce small baseline imbalances that result from differences in the score for units on either side of the cut-point, provided that the relationship between the score and outcome is modeled correctly. I will discuss some possible models in the next section, as well as draw connections between these models and the second regression discontinuity approach.

The Continuity Set-up. Rather than treating regression discontinuity as a local randomized experiment, most researchers include all of the data and use two regression lines to estimate the difference between the average outcomes of the treated and control units at the cut-point. By using regression lines, they are treating the scores as fixed rather than random. The randomness is now in the error terms of the y's, and the regression lines are the conditional mean functions that estimate the expected value of y across different values of the score. I call this approach the continuity set-up. When researchers use it, the major question that they face is how to construct the regression lines. Specifically, they have two key decisions to make.

One of these decisions is what bandwidth to use. The bandwidth, denoted as h, can be thought of as analogous to the size of the RD window. It specifies how far to the right or left of each point z to look to compute the value of the regression line at z. For instance, if the bandwidth is set at $h = 2$, then the value of the regression function at z would be estimated by using observations that lie between (z-2, z+2). The bandwidth would normally be chosen using an optimal bandwidth selection algorithm like the one provided by Imbens and Kalyanaraman (2011) or Calonico, Cattaneo, and Titiunik (2014).

The other decision that researchers need to make is how to estimate the regression function using the points within the bandwidth. One method is to simply take these points and calculate their weighted mean, where the weights discount observations the farther they are from z (kernel smoother). They could also run a (new) regression line through these points and take the $\hat{y}$ value at z (they would then repeat this process for every z to construct their two regression lines). This method is called local regression. For instance, they could run weighted or unweighted regression through the points around in $(z - h,z + h)$ for every z (local linear regression), or they could use polynomial regression (local polynomial regression).

But regardless of how they decide to construct their two regression lines, the regression discontinuity estimator is

$$\hat{\tau}_{RD} = E[Y|Z = c^{+}] - E[Y|Z = c^{-}]$$

which is just the difference between the two regression lines at the cut-point. Thus, when calculating their estimate, they are only focusing on the value of the regression lines at the cut-point, where Z=c. By doing so, they are ignoring all units that are farther than h away from the cut-point. This is why the bandwidth is analogous to the size of the RD window. It specifies how far to the left and right of the cut-point researchers will look to estimate the values of the two regression lines at the cut-point.

Comparing the Two RD Approaches. On a conceptual level, the two regression discontinuity approaches are very different. The natural experiment approach estimates the average treatment effect for the units within the RD window. This is a well-defined causal effect provided that the treatment was as-if random for those units. On the other hand, the continuity approach estimates the average treatment effect at the cut-point, where no units exist. In this sense, there is no sample where this treatment effect is defined, making it radically different than a traditional causal parameter. Moreover, the continuity approach makes no as-if random assumption, except at the cut-point where there are no units.

Nonetheless, the two approaches are very similar mathematically. As mentioned before, the size of the RD window in the natural experiment approach is analogous to the bandwidth in the continuity approach. All points that fall outside the RD window (or bandwidth) have no impact on the results. In the same way, if we take the natural experiment approach and control for only the score, how we control for it is analogous to how we construct the regression lines in the continuity approach. In fact, as long as the bandwidth equals the size of the RD window and we use the same method to compute the standard errors, the natural experiment and continuity set-ups will have mathematically equivalent forms.

The parallels are very straightforward. If we take the natural experiment approach and conduct a weighted difference in means test that weights units by their scores, it is mathematically equivalent to using a kernel smoother in the continuity approach, provided we used the same kernel (or weighting rule) for both approaches. If we take the natural experiment approach and run a regression on the sample within the RD window that controls for Z and the interaction $Z * T$, that is equivalent to using local linear regression (unweighted) in the continuity approach. If we take the natural experiment approach and control for for $Z, Z^2, ..., Z^k$ and the interactions $Z * T$, $Z^2 * T$, ..., $Z^k * T$, that is equivalent to using a local polynomial regression of order k (unweighted) in the continuity approach.

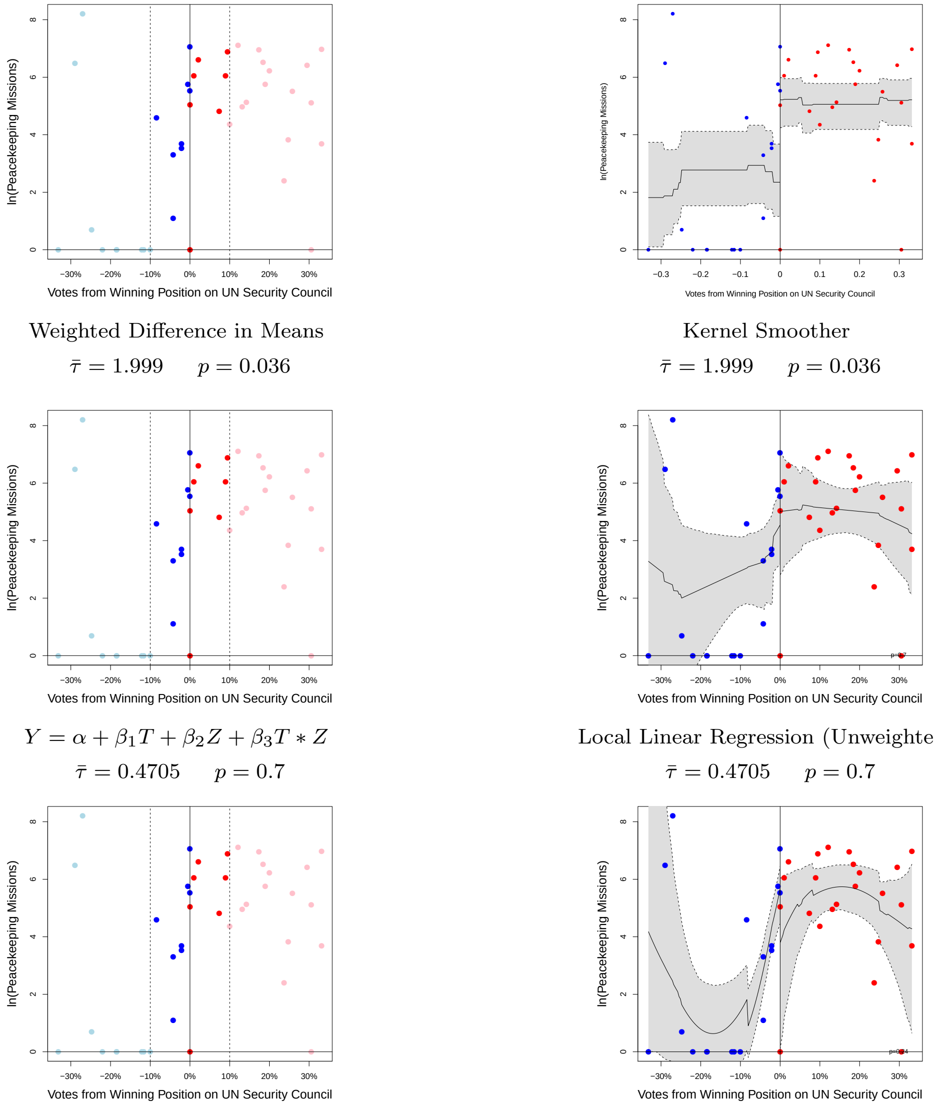
Figure 1 summarizes these relationships graphically, using a study by Voeten (2013) that looks at how being elected to the UN Security Council influences countries' willingness to participate in peace keeping missions. Voting for the UN Security Council is sometimes done through competitive elections, where countries secure a position if they receive at least 50% of the votes. I set the bandwidth (or size of the RD window) at $h = 0.1$, or 10%, which is close to the optimal bandwidth using the algorithm proposed by Calonico, Cattaneo, and Titiunik (2014). Thus, we are focusing on countries that received between 40% and 60% approval. Normally, we would also want to try smaller bandwidths where the as-if random assumption is more plausible, especially if we were not controlling for the score. However, since we are just using this study as an illustrative example, we will fix the bandwidth at 10%.

The figure shows the mathematical similarity between the natural experiment and continuity approaches. If we use the natural experiment approach and control for the systematic discrepancies between the treatment and control groups resulting from small differences in the score, we are really just doing a version of the continuity approach. Which version we are doing simply depends on how we control for the score.

A key point here is that both approaches assume exogeneity (either in the entire RD window or at the cut-point). All that the continuity set-up addresses is the bias that can result from units on one side of the cut-point being slightly different than units on the other side due to the small differences in their scores. If certain units are more likely to barely surpass the cut-point than others for any reason other than the small differences in their scores, than the continuity approach is no longer unbiased. For instance, we can live with the fact that candidates that won by less than 100 votes might tend to be of slightly higher quality than candidates who lost by less than 100 votes. This bias can be eliminated by using the continuity approach, or by taking the natural experiment approach and controlling for the score. However, we cannot live with the situation where certain candidates had other advantages that increased their chances of barely winning their elections.

Now while these two approaches are very similar mathematically, they each have their own advantages conceptually. In particular, the natural experiment approach is very useful for thinking about the design. We can ask whether the units that are very close to the cut-point all had roughly the same probability of being treated, or if some units had an advantage that cannot be accounted for by simply controlling for the score. This question does not make sense in the continuity set-up, as each unit's score is considered fixed. However, it is an essential question for any regression discontinuity design. It can help researchers identify major problems and their solutions, as I will

Figure 1. Similarity of the RD Approaches



Natural Experiment Setup

Continuity Setup

Weighted Difference in Means

$\bar{\tau} = 1.999 \qquad p = 0.036$

Kernel Smoother

$\bar{\tau} = 1.999 \qquad p = 0.036$

$Y = \alpha + \beta_1 T + \beta_2 Z + \beta_3 T * Z$

$\bar{\tau} = 0.4705 \qquad p = 0.7$

Local Linear Regression (Unweighted)

$\bar{\tau} = 0.4705 \qquad p = 0.7$

$Y = \alpha + \beta_1 T + \beta_2 Z + \beta_3 Z^2 + \beta_4 T * Z + \beta_5 T * Z^2$

$\bar{\tau} = -1.885 \qquad p = 0.24$

Local Polynomial Regression (2nd Order)

$\bar{\tau} = -1.885 \qquad p = 0.24$

Note: For the p-values to be equal, the standard errors must be bootstrapped, as there are small differences in some of the methods for computing p-values across the approaches.

illustrate in the next three sections.

Section 2: Variation in Competitiveness Across Strata

In the simplest RD set-up, units are all grouped together into the same pool, such as test-takers competing for a fixed number of scholarships. In these cases, there is only one strata, and all units belong to it. Analysis under this set-up is fairly straightforward. As long as the units around the cut-point could not manipulate their scores in a precise way, the treatment assignment around the cut-off should be "as-if" random. We should not expect imbalances on factors like gender, wealth, or ideology, since these should be balanced by the exogeneity of the treatment.

However, many regression discontinuities have more complicated formats where units are grouped into strata. For example, RDs that involve elections have units that are grouped into electoral strata (districts, states, countries, etc.) and the candidates in each strata compete against each other for office. This format has also appeared in other types of RD, including Van der Klaauw (2002), Niu, Xinchun, and Tienda (2013), and Bertoli (2017a).

The problem is that differences across strata could bias the results if they correlate with the probability of barely winning or barely losing. Before I lay out this problem formally, let me first illustrate it with an example. Consider an electoral system with two major parties. Now imagine that a third party decides to run in districts where it believes it has a good chance of winning, and that in many cases it scores close to the cut-point along with the other two parties. For these districts where three candidates are in the RD window, the candidates would have roughly a 1/3 chance of barely winning and a 2/3 chance of barely losing. Thus, the candidates running in these districts would be more likely to be bare losers. If these candidates differed in systematic ways from the candidates running in districts with only two parties in the RD window, the unequal probability of treatment assignment could bias the results.

This problem can arise anytime the strata vary in terms of their competitiveness. In these cases, more competitive strata will tend to have more bare losers than less competitive strata. In the election example, a district with four competitive parties may have three bare losers, whereas a district with two competitive parties can only have one bare loser. Thus, when we combine a sample of bare winners and bare losers in close elections, we would expect the following two observable implications:

(1) The bare loser group should be larger than the bare winner group.

(2) The bare loser group should have more units from competitive districts than the bare-winner group.

We should also expect these problems to worsen as the number of districts with multiple parties close to the cut-point grows.

This source of imbalance will be particularly dangerous if we are trying to estimate something like the party incumbency advantage, which many recent studies have tried to do in multi-party systems (Uppal 2009; Titiunik 2009). Bare losers will be more likely to be running in districts where they are up against multiple rival parties that have a good chance of winning. Thus, when we look at whether these parties won in their next election, we should expect them to win at lower rates than if they were running against just one competitive party. In short, they are more likely to lose this time and more likely to lose next time. The result is that barely losing will appear to be more costly than it really is, causing us to overestimate the incumbency advantage.

Researchers using the continuity approach might argue that their design avoids this problem because it estimates the LATE exactly at the cut-point, and the probability of a third place candidate existing at this point rapidly converges to 0 for a continuous score. Most importantly, it converges to 0 much faster than the probability for second-place candidate being at the cut-point. In other words, while it is very unlikely that one candidate will lose by exactly one vote, it is far less likely that two candidates from the same district will both lose by exactly one vote. Thus, the influence of third-place candidates should be negligible at the cut-point.

However, in RD applications, inferences about the cut-point are always drawn from the data around the cut-point, so additional bare losers that are not exactly at the cut-point can influence the results. As I will show momentarily, whether controlling for Z helps eliminate this bias depends on whether the model correctly captures the way that the third-place units' influence decreases the closer they get to the cut-point. If it overestimates or underestimates the rate at which their influence changes, then the regression function can be biased at the cut-point.

Before getting to that, I will first write this problem out formally using the simple natural experiment set-up. The estimator we will use is $\hat{\tau}_{Full}$, which will be the difference in means estimator for all units that fall in the RD window. We will assume for now that there can only be one winner, as in electoral cases. Then our estimator is written as

$$\hat{\tau}_{Full} = \frac{1}{m}\sum_{i=1}^{n} Y_{it}T_i - \frac{1}{n-m}\sum_{i=1}^{n} Y_{ic}(1-T_i)$$

Let C_i denote the number of units that are in the RD window in Unit i's strata. So $I(C_i = k)$ is an indicator variable that equals 1 if there are k units in the RD window in Unit i's strata, and equals 0 otherwise. Moreover, let q be the maximum number of units in the RD window in any strata. Then $\hat{\tau}_{Full}$ can be decomposed as follows

$$\begin{aligned}\hat{\tau}_{Full} = {} & \frac{1}{m}\sum_{i=1}^{n} Y_{it}I(C_i = 2) - \frac{1}{n-m}\sum_{i=1}^{n} Y_{ic}I(C_i = 2) + \\ & \frac{1}{m}\sum_{i=1}^{n} Y_{it}I(C_i = 3) - \frac{1}{n-m}\sum_{i=1}^{n} Y_{ic}I(C_i = 3) + \\ & \cdot \\ & \cdot \\ & \cdot \\ & \frac{1}{m}\sum_{i=1}^{n} Y_{it}I(C_i = q) - \frac{1}{n-m}\sum_{i=1}^{n} Y_{ic}I(C_i = q)\end{aligned}$$

Now note that for a strata with k units in the RD window, the probability of barely winning is $P(T_i = 1) = \frac{1}{k}$ and the probability of barely losing is $P(T_i = 0) = \frac{k-1}{k}$. Using these expression, we can write out the expected value of our estimator as follows:

$$\begin{aligned} E[\hat{\tau}_{Full}] = {} & \tfrac{1}{m} \cdot \tfrac{1}{2} \textstyle\sum_{i=1}^{n} Y_{it} I(C_i = 2) - \tfrac{1}{n-m} \cdot \tfrac{1}{2} \textstyle\sum_{i=1}^{n} Y_{ic} I(C_i = 2) + \\ & \tfrac{1}{m} \cdot \tfrac{1}{3} \textstyle\sum_{i=1}^{n} Y_{it} I(C_i = 3) - \tfrac{1}{n-m} \cdot \tfrac{2}{3} \textstyle\sum_{i=1}^{n} Y_{ic} I(C_i = 3) + \\ & \vdots \\ & \tfrac{1}{m} \cdot \tfrac{1}{q} \textstyle\sum_{i=1}^{n} Y_{it} I(C_i = q) - \tfrac{1}{n-m} \cdot \tfrac{q-1}{q} \textstyle\sum_{i=1}^{n} Y_{ic} I(C_i = q) \end{aligned}$$

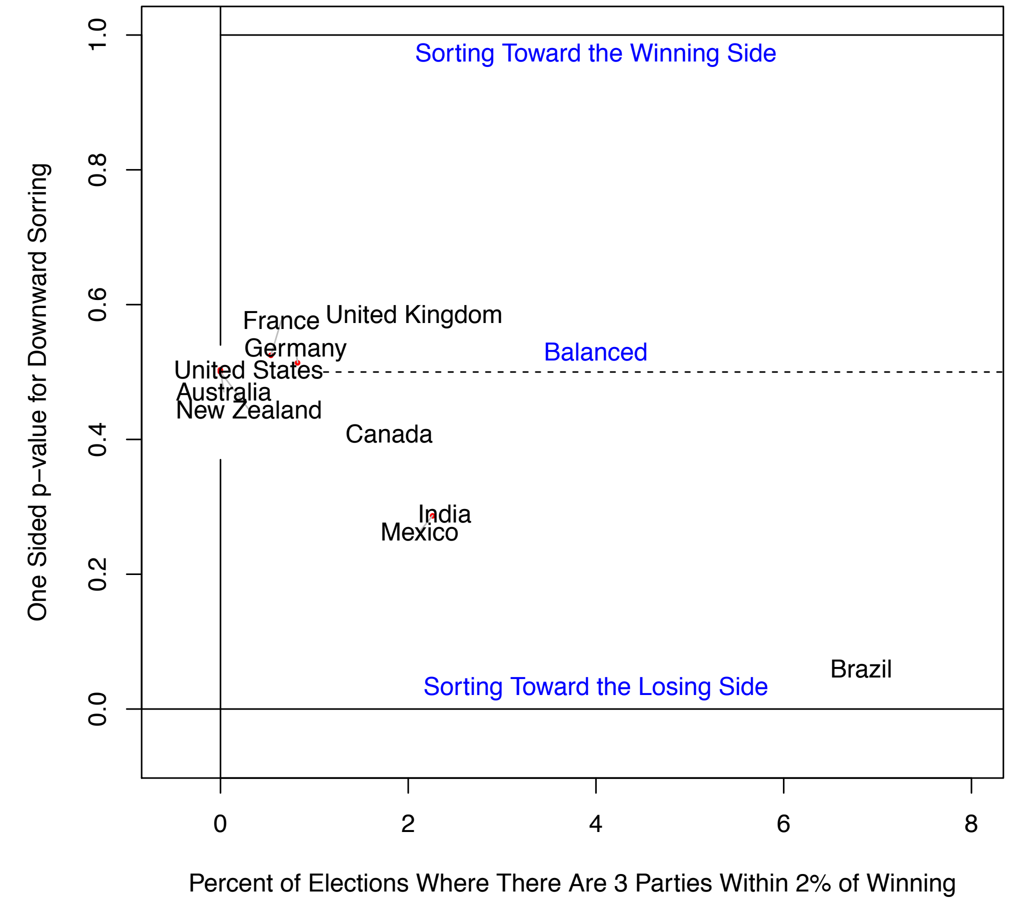
which reduces to

$$\begin{aligned} E[\hat{\tau}_{Full}] = {} & \tfrac{1}{n} \textstyle\sum_{i=1}^{n} [Y_{it} I(C_i = 2) + Y_{it} I(C_i = 3)/(3/2) + ... + Y_{it} I(C_i = k)/(k/2)] - \\ & \tfrac{1}{n} \textstyle\sum_{i=1}^{n} [Y_{ic} I(C_i = 2) - (3/2) \cdot Y_{ic} I(C_i = 3) - (k/2) \cdot Y_{ic} I(C_i = k)] \end{aligned}$$

Of the two terms on the right hand side, the top estimates the average outcome under treatment, $\frac{1}{n}\sum_{i=1}^{n} Y_{it}$, and the bottom estimates the average outcome under control, $\frac{1}{n}\sum_{i=1}^{n} Y_{ic}$. However, the estimate of the average outcome under treatment underweights units from strata with many units in the RD window, while the estimate for the average outcome under control overweights these units.

This problem is evident in the cross-national data on legislative elections that was compiled by Eggers et al. (2014). These scholars constructed a dataset that provides the vote shares of the first three parties in about 200,000 elections across ten countries, which include two-party and multi-party systems. In two-party systems, there are only two candidates close to the cut-point in any election, aside from the rare cases where an independent or minor third-party candidate performs very well. Thus, we should not expect variation in competitiveness across strata to be a major problem when dealing with two-party systems. Without this variation, there should not be many more bare losers than there are bare winners. On the other hand, in countries where there are several competitive parties, we should expect third-party

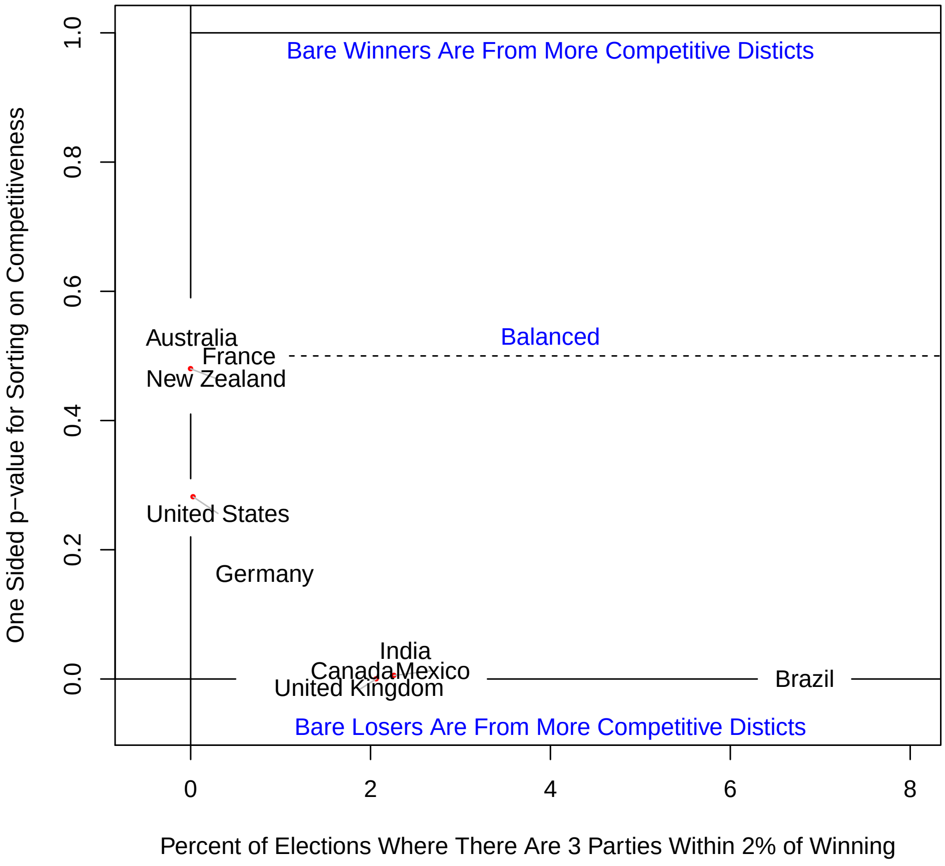
Figure 2: McCrary Sorting Test Results as Third Party Competitiveness Increases



candidates to cause differences in competitiveness across strata, which will lead to a larger bare loser group.

Figure 2 shows whether there tended to be more bare losers in the countries using the McCrary sorting test. This test checks for whether there was an overall tendency for units to sort to one side of the cut-point or the other. Any such sorting would be a threat to inference, unless it was entirely independent of the potential outcomes. As this figure shows, countries that had more districts with multiple candidates near the cut-point tended to fail the McCrary sorting test, Specifically, they had more bare losers than we would expect if each candidate was about equally likely to be on one

Figure 3: Testing for Imbalance on Competitiveness of District Across Countries



Note: p-values values are computed using difference in means tests for the sample of candidates that won or lost by less than 2%.

side of the cut-point or the other.

Figure 3 shows that countries that had more districts with multiple candidates near the cut-point also tended to have more imbalance on the competitiveness of district. There is no imbalance for countries like Australia, France, and New Zealand, where there are only two competitive parties in close elections. However, the p-values are close to 0 for counties like India, Mexico, Brazil, and the United Kingdom (which has the largest number of close elections aside from the United States). For the entire sample, about 2.8% of the bare losing candidates came from competitive districts, compared to about 1.4% of the bare winning candidates. For a t-test in the RD window, the p-value is $p \approx 1.7 \cdot 10^{-11}$.

In fact, it is fairly easy to see that this imbalance must exist. Each winner has exactly one corresponding runner-up. For instance, if the winning party won by 1%, then the runner-up must have lost by 1%. Moreover, both of these units are either from a district with only two parties close to the cut-point, or they are both from a more competitive district with three parties close to the cut-point. Thus, the winners and runner-ups by themselves will always be balanced on competitiveness of district, since they are a mirror image of each other. However, in cases where there is a third-place party in the RD window, that party adds an additional unit in that district. When this happens in many districts, these additional bare losers will create the imbalance that is evident in Figure 3.

Because the density of third-place candidates decreases the closer we get to the cut-point, our estimator might be unbiased if we control for Z with a model that accurately captures the decreasing influence of third place candidates as they approach the cut-point. In the case of elections with three parties, the influence of third-place parties will actually decrease linearly close to the cut-point. For example, imagine that we set the size of the RD window at (-2%,+2%). The score of a third-place candidate in this RD window is (roughly) the minimum of two draws from the random variable Unif(-2%,0%), which has a linear density (in this case, $f(x) = -x/2$ for $x \in (-2, 0)$)). Thus, local linear regression can substantially reduce bias in the three-party case. The kernel smoother, however, can lead to badly biased results, as it incorrectly models the relationship between Z and competitiveness.

While the local linear smoother can help us avoid bias when some strata have three candidates in the RD window, it cannot resolve cases where some strata have four or more candidates close to the cut-point. Put simply, these strata will induce a non-linear relationship between Z and the influence of additional bare losers. The fourth place candidates will be distributed $f_1(x) = 3x^2/8$ and the third place candidates will be distributed for $f_2(x) = -3x^2/2 - 3x$ for $x \in (-2, 0)$. These distributions are easily derived using order statistics formulas. Without the linearly decreasing influence of the additional bare losers, the local linear smoother can be badly biased. For instance, if you take Eggers et al. dataset of roughly 20,000 bare winners and bare losers and add just 91 fourth-place candidates, each of whom scores half-way between -2% and the third-place candidates, the local linear estimator returns significant imbalance on competitiveness of district. While high-order polynomial smoothers may be able to capture the non-linear influence of the additional bare losers to some extent, they will also be more sensitive to noise in the data and have a much higher variance (Gelman and Imbens 2014).

In short, when strata have varying numbers of units in the RD window, how we control for Z matters. It is no longer a robustness check, but a key decision about the design that can greatly impact the results. In the three-candidate case, we can go from no bias with a local linear smoother to substantial bias with a kernel or nearest-neighbor smoother. If we add some fourth-place candidates into the dataset, local linear regression will also become biased. Put simply, the additional bare losers make our results highly sensitive to our modeling assumptions.

Since one of the most appealing features of RD is that it tends to be much less sensitive to how we model the data, it is worth considering some ways of resolving this problem. The easiest solution is just to restrict the analysis to the two units in each strata that were closest to making and closest to missing the cut-point. While this approach involves dropping units, and therefore losing some information, it maintains balance over strata and time. In other words, the bare winners and bare losers in the election example will be perfectly balanced on district and year, since every bare winner has a corresponding bare loser. Thus, this procedure creates a blocking scheme that should increase balance in the sample. Researchers can also still adjust for Z

and present the normal continuity graphs with two regression lines that meet at the cut-point. It is just that all units that were not the closest to making or missing the cut-point would be left out of this graph.

The second solution is to reweight units based on their probability of treatment assignment given their strata. If Unit i is from a strata with k units in the RD window and j bare winners, then we would give Unit i weight $\frac{k}{n} \cdot \frac{k-j}{k}$ if it barely won and weight $\frac{k}{n} \cdot \frac{j}{k}$ if it barely lost. We could then use a difference in means test under the natural experiment set-up. In electoral settings, this estimator would be

$$\begin{aligned}
\hat{\tau}_{ReweightedFull} = {} & \tfrac{1}{n_2/2} \textstyle\sum_{i=1}^{n} Y_{it} T_i I(C_i = 2) - \tfrac{1}{n_2/2} \textstyle\sum_{i=1}^{n} Y_{ic}(1 - T_i) I(C_i = 2) + \\
& \tfrac{1}{n_3/3} \textstyle\sum_{i=1}^{n} Y_{it} T_i I(C_i = 3) - \tfrac{1}{2n_3/3} \textstyle\sum_{i=1}^{n} Y_{ic}(1 - T_i) I(C_i = 3) + \\
& \vdots \\
& \tfrac{1}{n_q/q} \textstyle\sum_{i=1}^{n} Y_{it} T_i I(C_i = q) - \tfrac{1}{(q-1)n_q/q} \textstyle\sum_{i=1}^{n} Y_{ic}(1 - T_i) I(C_i = q)
\end{aligned}$$

where n_k is the number of units from strata with k outcomes close to the cut-point. This estimator is easily proven to be unbiased under the as-if random assumption by taking the expected value of $\hat{\tau}_{ReweighedFull}$. It is also still possible to control for the score by using local regression with this weighting scheme, thus getting the benefits of local linear or local polynomial regression. By reweighting, this procedure maintains balance across strata while also using all of the data in the RD window.

Section 3: Units That Reappear in the Sample

There are two versions of the multiple-score problem. The simple version occurs when units have outcomes for each scoring round, such that each score is associated with a single outcome. In these cases, researchers can simply treat every time that a unit receives a score as a separate observation. The more problematic version arises when each unit has only one outcome, meaning that multiple scores will now correspond to a single outcome. For instance, if we wanted to estimate how passing an exam influenced the chances that students went to college, we would have a problem if the students could take the exam multiple times. However, there is a fairly straightforward solution to this problem that I will discuss in the second half of this section.

Unique Outcomes for Every Score. This issue arises for many studies, including (Lee 2008), Hainmueller and Kern (2008), Lalive (2008), Lemieux and Milligan (2008), Broockman (2009), Titiunik (2009), Cellini, Ferreira, and Rothstein (2010), Hall (2014), Eggers et al. (2014), and Bertoli (2017a; 2017b). For example, consider the RD studies that look at the incumbency advantage. Their goal is to estimate how winning an election at time t affects a party's vote share and winning probability at time $t + 1$. To increase the sample size, these studies look at long historical periods rather than a single election year, treating each election as a separate observation. The units are parties within each district. These parties' treatment assignment is whether they won, their score is how close they were to winning, and their outcome is how they did in the next election. One entry in the dataset would be Party k in District i with the treatment assigned at time t and the outcome recorded at time $t + 1$, another would be the same party and district with the treatment assigned at $t+1$ and the outcome recorded at $t+2$, and so on. Thus, the problem of having multiple scores for single outcomes does not arise, but there is still the issue of rerunning units.

In these cases, the key question is whether the treatment effect in each round influences the LATE in future rounds. If there is good reason to believe that it does not, then researchers can consider each case where a unit received a score as a single observation. However, if the treatment assignment in one round affects either who is in the RD window later or the size of future individual treatment effects, then researchers need to be more careful about how they proceed.

Since this issue is probably clearest in the party incumbency advantage literature, I will stick with that example. First, let us assume that treatment assignment is as good as random within a small window around the cut-point. Now imagine that winning an election at time t increases vote share at $t+1$ by 3 points. This implies that treatment assignment in round t can affect which units are very close to the cut-point at time $t+1$, so treatment in early rounds affects who is influencing the LATE in later rounds. Furthermore, imagine that some units also receive a 5-point bump for winning if they won in the previous round, but they only experience a 3-point bump if they lost in the previous round. If some of these units are in the RD window back-to-back years, then treatment assignment affects both who influences the LATE and the size of treatment effects for some units after the first round.

We can still test the sharp null hypothesis that incumbency has no impact of vote share in future rounds, since this hypothesis guarantees independence between rounds. However, deriving point estimates and confidence intervals for the LATE is no longer a simple matter. On one hand, there is no longer a single LATE, since both who is close to the cut-point and how large some of the individual treatment effects are in later rounds depend on treatment assignment in earlier rounds. Thus, the **LATE** is a random variable. The parameter that we might be interested in estimating is the expected value of the **LATE** over all treatment assignments:

Surprisingly, the normal RD procedures give us an unbiased estimate of this value, assuming that the treatment is random within the RD window and we restrict our attention to the pair of candidates closest to the cut-point. Let $\mathbf{T}$ be the treatment assignment vector, which is composed of the treatment assignment vectors in each of the k rounds.

$$\mathbf{T} = \mathbf{T_1} \cup \mathbf{T_2} \cup ... \cup \mathbf{T_k}$$

Let the number of units per round be denoted by $n_1, \mathbf{n_2}, ..., \mathbf{n_k}$ units, which are also random variables that depend on the treatment assignments in previous rounds, with the exception of n_1. Furthermore, the **LATE** is a random variable that is the sum of the LATE's from each round, $LATE_1$, $\mathbf{LATE_2}$, ..., $\mathbf{LATE_K}$, where each of these terms is multiplied by the percentage of units that are coming from that round:

$$\mathbf{LATE} = \frac{n_1}{\mathbf{n}} LATE_1 + \mathbf{\frac{n_2}{n} LATE_2} + ... + \mathbf{\frac{n_k}{n} LATE_k}$$

So the parameter we are interested in is

$$E[\mathbf{LATE}] = E[\frac{n_1}{\mathbf{n}} LATE_1] + E[\mathbf{\frac{n_2}{n} LATE_2}] + ... + E[\mathbf{\frac{n_k}{n} LATE_k}]$$

The normal RD estimator can also be broken down into the estimates for each round. The estimate for the first round is clearly unbiased under the normal as-if randomness assumption, since there is no rerunning problem in this round. However, T_1 is not only one of $2^{\frac{n_1}{2}}$ possible treatment assignments for Round 1, but also one of $2^{\frac{n_1}{2}}$ possible paths to Round 2, which is selected at random. Moreover, the RD estimate for Round 2 is an unbiased estimator for all possible treatment assignments in Round 2 on that path. So we randomly selected our path to Round 2, and then got an unbiased estimate for Round 2 conditional on that path. Similarly, T_2 is a randomly selected path to Round 3, and we again get an unbiased estimate of the LATE in that round conditional on that path. This pattern continues until we reach the last round.

The result is that the full path from Round 1 to Round k is randomly selected from all possible paths, and in each round along the way we have an unbiased estimator for the LATE in that round conditional on being on that path. So we have

$$E[\tau_{\mathbf{RD}}] = E[\frac{n_1}{\mathbf{n}} LATE_1 + E[\mathbf{\frac{n_2}{n} LATE_2}|T_1] + ... + E[\mathbf{\frac{n_k}{n} LATE_k}|T_{k-1}, ..., T_1]]$$

The tower property gives us

$$E[\tau_{\mathbf{RD}}] = E[\frac{n_1}{\mathbf{n}} LATE_1] + E[\mathbf{\frac{n_2}{n} LATE_2}] + ... + E[\mathbf{\frac{n_k}{n} LATE_k}]$$

$$E[\tau_{\mathbf{RD}}] = E[\mathbf{LATE}]$$

So even when the size of the effects and observations in the RD window depend on the treatment, we can still estimate an important parameter of interest. It should be noted, however, that we do not have an unbiased estimate of the realized LATE that occurred on the observed path unconditional on the path, unless that LATE equals $E[\mathbf{LATE}]$. It will also not necessarily be an unbiased estimate of the realized LATE conditional on the path. For instance, some LATE's may only come about if the bare winners and losers in the first round are very different on baseline covariates, and in these cases $\tau_{\mathbf{RD}}$ will be biased from the chance imbalance early on.

While we can obtain an unbiased point estimate of $E[\mathbf{LATE}]$, confidence intervals are more complicated? Unfortunately, we have no way of determining what the distribution of the **LATE** looks like. We only have unbiased estimates for each round on a randomly determined path, but with possible correlation between treatment assignment and the LATE in future rounds. Simulations suggest that the normal RD confidence intervals can be misleading if the **LATE** in future rounds is highly dependent on treatment assignment in earlier rounds. Of course, this problem does not apply if we are simply testing the sharp null hypothesis.

This issue is a major limitation if we care about quantifying the uncertainty of our point estimate, which is certainly true in the party incumbency literature. In cases like this, where we both care about the confidence interval and believe that treatment assignment influences the **LATE** in later rounds, it would probably make more sense to take a different approach. One option would be to drop every district after it appears in the RD window once, which would limit the sample size and change the parameter to a different LATE, but would provide us with a valid confidence interval under the normal RD assumptions.

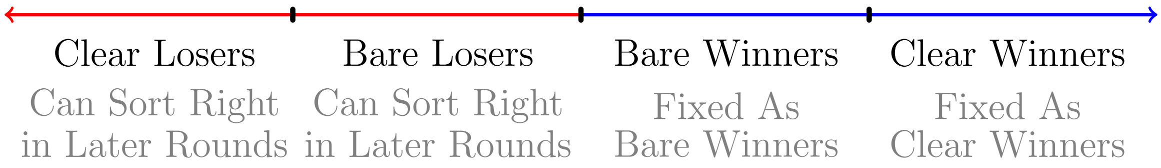
Only One Outcome for Each Unit. A more complicated scenario arises when units have multiple scores but only one outcome. In these cases, there are usually units that both make and fall short of the cut-point at different times. The problem is that it is not entirely clear whether they belong in the treatment or control group, or what values of their scores should be used.

The most famous application that deals with this problem is Eggers and Hainmueller (2009) APSR article “MP’s for Sale: Returns to Office in Postwar British Politics,” which tests how winning a seat in British Parliament influences wealth at death. To get around the rerunning issue, they count all candidates who won at least once as winners and use their scores from their first winning races, while counting all candidates who never won as losers and using their scores from their best losing races. For instance, a candidate who ran twice and lost with scores $\{-30\%, -2\%\}$ would be given the score -2%, and a candidate who ran four times with scores $\{-5\%, 14\%, -1\%, 4\%\}$ would be given the first winning score of 14%. Eggers and Hainmueller then use the usual RD procedures on this new sample, where each candidate has only one score and outcome. I call the estimator that is obtained by these procedures the First-Winning-Best-Losing (FWBL) estimator.

While the FWBL estimator ensures that no unit will have multiple scores, it creates several other problems. The first is that the expected size of the treatment group will now be much larger than the expected size of the control group. This problem occurs because bare losers can switch to winners by rerunning, but bare winners are immediately fixed as bare winners provided that they did not already win in a previous round.

Thus, there are four types of candidates: (1) candidates who lose on their first attempts and fall outside the RD window, and who can rerun in later rounds and move right; (2) candidates who lose on their first attempt and fall inside the RD window, and who can also rerun in later rounds and move right; (3) candidates who win on their first attempts and fall inside the RD window, and who now have fixed scores in the treatment group; and (4) candidates who win on their first attempts and fall outside the RD window, whose scores are now also fixed at their first winning

Figure 2: Types of Candidates Based on First Score



score. Although the sizes of the bare winner and bare loser groups should be about equal after every candidate's first run, candidates in the second group (bare losers) can rerun and switch to the bare winner group or to the group of winners who scored outside the RD window. The result is that the bare loser group gets depleted, with some of their members switching to the bare winning group. While initial losers in the first group can rerun and replace them, these units are just as likely to be bare winners as bare losers in any future round. Moreover, even if they move up to become bare losers, they can still sort out of this group by winning an election in a future round.

In short, the more opportunities that candidates have to win, the more likely they are to be counted as a bare winners and the less likely they are to be counted as bare losers. Thus, within the RD window, treatment should be correlated with the likelihood of rerunning. More specifically, it should be correlated with their probabilities of rerunning after losing, since they become fixed in the sample after their first win. Since a candidate's probability of rerunning after losing is likely to be related to individual resources and determination, there is reason to suspect that the FWBL estimator will be biased in many cases.

There is also another potential source of bias that would arise even if all units were guaranteed to run the same number of times. Imagine that there are two types of candidates, those who respond to barely losing by working hard and succeeding later, and those who get frustrated and put little effort into future campaigns. The more resilient candidates would tend to sort out of the bare loser category in later rounds, whereas their more easily discouraged counterparts would lose badly and be counted as bare losers based on their performance in the first round. Thus, even if there was a predetermined rule that required each candidate to run exactly two times, the FWBL could be confounded by this additional source of bias. Specifically, we would expect bare losers to be less resilient than bare winners.

The nature of this bias is fundamentally more problematic than the imbalance on competitiveness across strata in Section 2 because a well-chosen smoother cannot eliminate this type of bias. In the prior example, the influence of the additional bare losers converged to 0 at the cut-point, since the probability of multiple candidates from one strata barely losing rapidly converged to 0 when Z is continuous. Thus, we could eliminate bias if we chose a smoother that accurately captured the relationship between Z and the decreasing influence of the additional bare losers. However, in this case the influence of candidates rerunning does not converge to 0 the closer we get to the cut-point. Candidates that lose by one vote would probably be very likely to rerun, and would bias the FWBL estimator if they won in a later election.

One potential solution to this problem is to use a different estimator that is proposed by Querubin and Snyder (2011) in a similar study. Their study looks at how winning a seat in US Congress affected individual wealth during the mid-nineteenth century. Rather than looking at each candidates first win or best loss, they look at the first time each candidate ran. Thus, they count any candidate who lost in his first round as a loser, even if he won and held office in a later round. Their justification for using this estimator is that only about 9% of candidates who lost on their first try attempted to rerun, so very few of their losers ended up holding office later.

This estimator can be thought of as an intention-to-treat (ITT) estimator. The idea here is that a candidate's outcome on his first try is an instrument for whether he ever held office. Winning on the first attempt guarantees that the candidate would hold office, whereas losing increases the probability that the candidate would never hold office. Thus, we can think of the candidates who lost on their first attempt and never held office as compliers, and the candidates who lost on their first attempt but won later as non-compliers. Of course, all units who won on their first attempt are compliers, since they were guaranteed to hold office. Thus, we have an instrument (winning or losing on the first attempt) and one-way non-compliance, making the estimator that Querubin and Snyder use an ITT estimator.

A third approach is just to acknowledge that the outcome in the first round is an instrument and proceed with the fuzzy regression discontinuity design. Fuzzy RDs are situations where the cut-point determines who receives an instrument instead of who receives the treatment. The Fuzzy RD estimator is the ITT estimator divided by the estimated local compliance rate, which is defined as the percentage of first-time losing candidates in the RD window who never held office. I denote this parameter as α_{LATE}. So the estimator is

$$LATE_{FRD} = \frac{LATE_{ITT}}{\alpha_{LATE}}$$

Since $\alpha_{LATE} < 1$, this value will have a larger absolute value than the ITT estimator. As with any fuzzy RD, the p-value for this estimator is the same as the ITT estimator p-value, which tests whether the instrument has an affect on the outcome.

The difference between the ITT and Fuzzy RD estimators is that the ITT estimator captures the impact of the instrument (succeeding the first time), whereas the Fuzzy RD estimator captures the effect of the treatment (succeeding at some point). In most applications, the Fuzzy RD estimator will probably be preferable, even when the compliance rate is high. Since the p-values from the two approaches are the same, the decision to use one estimator over the other matters for the size of the effect but not the statistical significance of the results.

There are two drawbacks of the Fuzzy RD estimator. First, it requires researchers to assume that the exclusion restriction holds. The instrument cannot affect the outcome in any way aside from affecting the likelihood of the treatment. This assumption may be invalid in many cases. In the economic returns to office example, we would have to assume that the candidates' outcome in their first races do not affect their future wealth except in how they change the probability of holding office at some point. This assumption would be violated if there were candidates who would win a future election if they barely fail the first time, and whether they succeeded the first time affects their future wealth. In this scenario, their performance on the first run does not affect their probability of holding office, which is 1 regardless of whether they win or lose. However, it does influence their outcomes, thus violating the exclusion restriction. It is therefore very important that researchers consider whether the exclusion restriction holds before using the Fuzzy RD estimator.

Second, the Fuzzy RD estimator will usually be biased in finite samples, as is normally the case with instrumental variable estimators. The reason is that we can estimate both $LATE_{ITT}$ and α_{LATE} without bias, but we cannot convert these values into an unbiased estimate of $\frac{LATE_{ITT}}{\alpha_{LATE}}$. This problem results from the rule in probability that $\frac{E[A]}{E[B]} \neq E[\frac{A}{B}]$ in general. However, if the sample size were to go to infinity, the variances of the numerator and denominator would go to 0, giving us an unbiased estimate of the $\frac{LATE_{ITT}}{\alpha_{LATE}}$. Therefore, the Fuzzy RD estimator is consistent, albeit not unbiased.

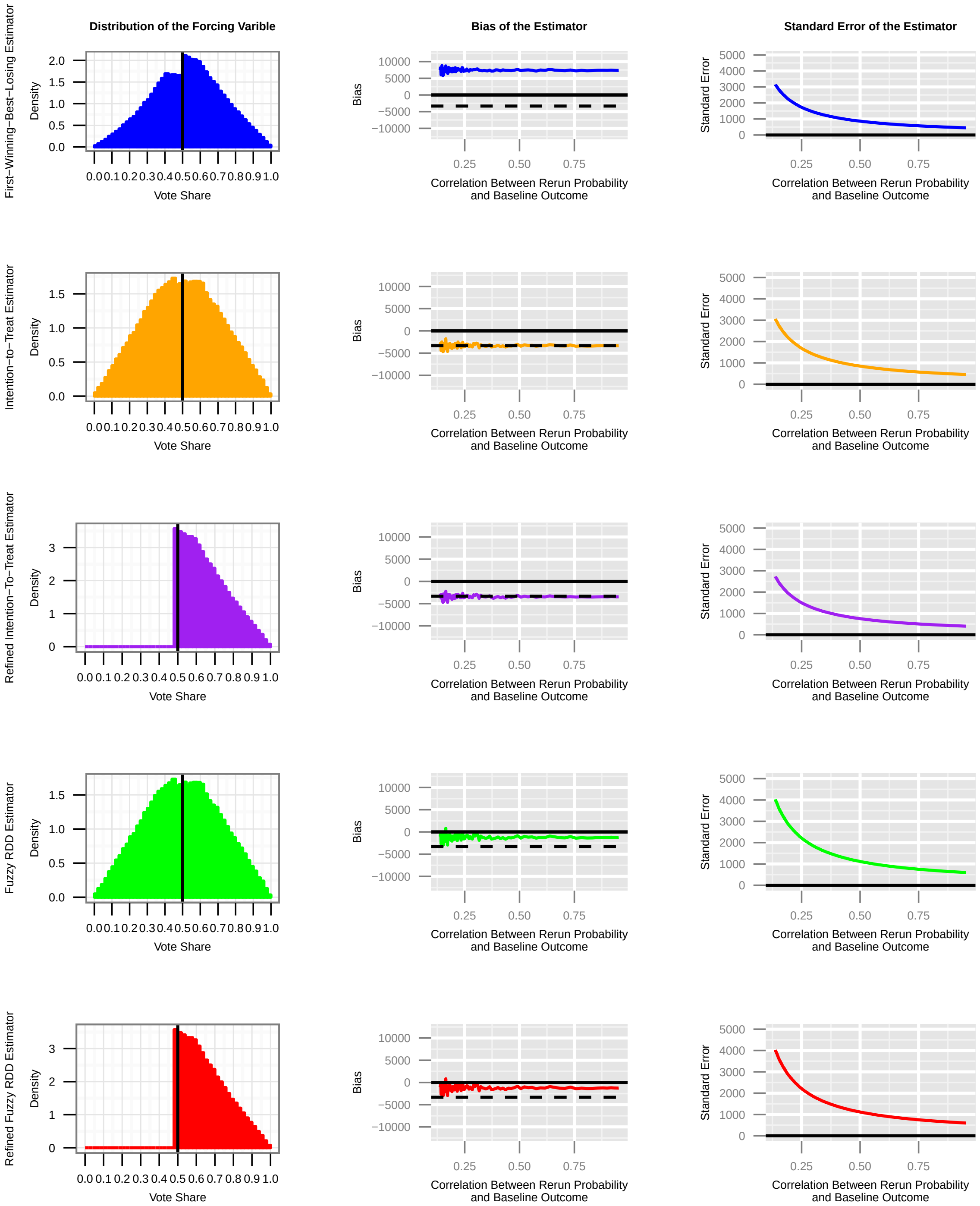
It is also possible to increase the power of both of the ITT and Fuzzy RD estimators by disregarding any attempt where a candidate scored below the RD window. In other words, only count an attempt as an attempt if the unit won or scored within the RD window. Thus, any early attempt where a candidate scored below the RD window would not disqualify that unit from the analysis. This rule will increase the sample size by retaining any candidate who had a bad early loss but produced a more competitive score later. I call these estimators the Refined ITT and Refined Fuzzy RD estimators.

In Figure 1, I compare the five estimators using a simple simulation. The simulation procedures are as follows, although the patterns observed here do not depend on the specific details, but will arise anytime the probability of rerunning is correlated with baseline determinants of the outcome. There are four election years, each with 10,000 candidates who run against unnamed opponents. Candidates who score above 50% win their elections, and candidates that score below 50% lose. Each candidate has a quality level $Q_i \sim Unif(0.2, 0.8)$. The vote share for a candidate in any election is $Z_i = Q_i + \epsilon$, where $\epsilon \sim Unif(-0.2, 0.2)$. Every candidate reruns with probability $P_i \sim Unif(0, 1)$, and candidates that drop out are replaced by new candidates. T_i is the treatment indicator which equals 1 if the candidate ever held office and 0 if not. A candidate's baseline wealth is $B_i = 100,000 \cdot P_i + \epsilon^*$, where ϵ^* is another random error term. Finally, wealth at death is $W_i = B_i + 10,000 \cdot T_i$. Thus, the treatment effect of holding office is a constant of \$10,000, and wealthier candidates are slightly more likely to rerun than poorer candidates.

The left-hand column shows the density of the running variable for each estimator. All densities are smooth around the cut-point except for the FWBL estimator, which we expected to be discontinuous since bare losers can sort right. The middle column shows the bias for each estimator as a function of the correlation between the baseline outcome and the probability of rerunning. As expected, the FWBL estimator is substantially biased, since wealthier candidates are more likely to rerun and become bare winners if we use their first-winning or best-losing vote share. The Fuzzy RD and Refined Fuzzy RD estimators are also biased, even though the exclusion restriction holds in this example. The other two estimators perform very well, although they focus on the intention to treat effect (or the effect of winning the first time). The third column presents the standard errors for each estimator. The standard errors for the Refined ITT and Fuzzy RD and estimators are slightly lower than the standard errors for the unrefined estimators, since these estimators increase the sample size.

It should be noted, however, that this simulation may not capture what is hap-

Figure 4: Comparing the Estimators with a Simulation



Note: The dotted horizontal line in the middle graph shows the intention-to-treat effect. Thus, the Refined Intention-to-Treat Estimator provides the best estimate of the intention-to-treat effect, and the Refined Fuzzy RD estimator provides the best estimate of the treatment effect.

pening in Hainmueller's and Eggers's study. It depends on the extent to which the probability of rerunning correlates with wealth, and they only have a measure of wealth at the time of each candidate's death. Querubin and Snyder do have an estimate of baseline wealth. However, very few of their candidates ever won after losing the first-time, so the compliance rate is too large to distinguish between these four types of estimators using their data. Thus, it is difficult to illustrate this problem with a real-world example. However, there are clear theoretical reasons to believe that the FWBL estimator will provide misleading results in cases where the compliance rate is lower (many units win after their first loss) and their is a strong correlation between the outcome and the probability of rerunning. The simulation illustrates this problem very clearly.

Section 4: Independent Variables Not Assigned by the RD

In the previous examples, scoring above or below the cut-point determined who received some treatment, like holding office or earning a scholarship. However, there are other cases where the main variable that researchers care about is not actually assigned by the RD. For instance, Meyersson (2014) compares Turkish municipalities where an Islamist party barely won and barely lost elections to test how Islamist rule affects women's rights. Other scholars have focused on cases where Democrats or Republicans narrowly won or lost gubernatorial races, looking at outcomes like taxation (Fredriksson, Wang, and Warren 2013), unemployment (Leigh 2007), and racial inequality (Beland 2014). Another example is Clots-Figuerasa's (2012) use of close races where women barely won or lost their elections to test whether female politicians increase the likelihood that children will receive a primary education.

This type of regression discontinuity can be very informative, but it is first important to recognize what we cannot learn from it using the normal RD assumptions alone. That is, it does not allow us to estimate the impact of the independent variable of interest on the outcome. For instance, Meyersson finds that the municipalities where an Islamist party barely won had better women's rights records afterward, but this finding tells us little about the effect of Islamism on women's rights. After all, Islamist parties differ from the non-Islamist parties in many ways besides religion and ideology, and the observed treatment effect could be due to some of these other differences. Put another way, his study tells us very little about how convincing a secular politician to join an Islamist party would affect women's rights. Similarly, Leigh finds that electing Democratic governors tends to result in less unemployment, but this does not mean that if we convinced a Republican governor to become a Democrat we should expect a similar effect. After all, Republican and Democratic governors differ in many ways besides their ideologies.

To understand what this type of regression discontinuity tells us under the normal assumptions, we must think of the RD set-up in a different way than before. In the more traditional RD cases, like when test-takers try to earn a scholarship, the units are the individuals and the treatment is getting the scholarship. However, in cases where the independent variable of interest is an attribute, the unit is no longer the individual. Instead, the units are the open "spots" that individuals or parties can fill, such as seats in Congress. The treatment is whether that open spot received an individual or party with a particular characteristic, although that characteristic was not randomly assigned and is likely confounded with other important factors.

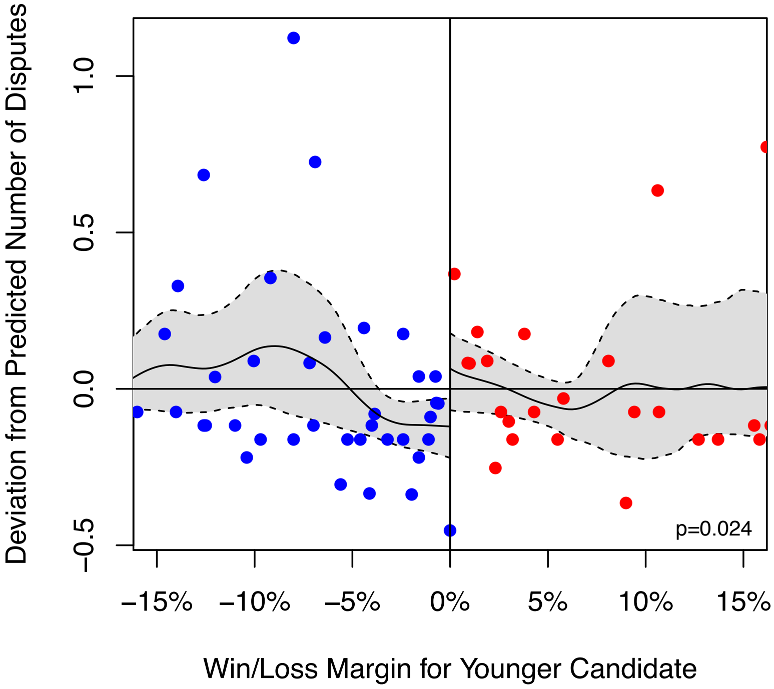
In this context, the regression discontinuity gives us causal leverage by creating exchangeability across the open spots, rather than across the units that fill them. For instance, municipalities where Islamist parties barely defeated secular parties should be similar to municipalities where they barely lost to secular parties. This exchangeability allows us to be confident that there are not systematic differences between districts controlled by Islamist and secular parties, but we still must be cautious that differences between the parties besides Islamism may be driving the results.

Thus, this type of regression discontinuity tells us how the outcome will tend to change if the seat to be occupied by someone of Type A or Type B, given that the race is tight. For example, if we were voters in Turkey who cared about women's rights, we should consider voting for the Islamist party if the election was a toss-up. However, whether the causal mechanism has anything to do with the party being Islamist is an open question. Similarly, Clots-Figuerasa's study tells us whether voting for a female politician in a close election is good for education levels. Nevertheless, it tells us very little about the effect of gender on education, since the female politicians in her study are not being compared to a similar group of counterfactual male candidates. We simply do not know what that comparison might reveal. Therefore, this type of regression discontinuity is more useful for predicting the impact of a certain decision, in these cases who to elect in tight races, than it is for understanding the effect of characteristics that were never assigned by the RD.

However, the RD can make it easier to estimate the effect of the characteristic when combined with observational methods. What the RD does is guarantee that the environments where units of Type A win are comparable to the environments where units of Type B win. If we also control for some of the confounding factors that make candidates of Type A different than candidates of Type B, then we may have a better case for arguing differences in outcomes are explained by the characteristics of interest.

Consider the study by Bertoli, Dafoe, and Trager (2017) that looks at the relationship between leader age and state aggression. Since younger leaders tend to have more testosterone than older leaders, it is possible that countries will start more military conflicts after electing younger leaders. Specifically, Bertoli, Dafoe, and Trager look at close elections between presidential candidates with large differences in age. The findings indicate that countries do behave more aggressively after barely electing

Figure 5: Testing How Electing Younger Leaders Affects State Aggression



younger leaders. However, it is possible that some differences between the younger and older leaders besides age are driving the results. Thus, it is safe to conclude that electing younger leaders tends to make countries more aggressive, but it is not certain that age is the key explanatory factor.

It is possible to control for other possible explanatory variables using linear regression. The first step is to create a linear model that predicts how many military conflicts a country will initiate given the other independent variables of interest, such as the leader's ideology, military experience, and whether they were an incumbent in the most recent election. The next step is to take the residuals from this linear model. These residuals represent every country's deviation from their expected level of conflict based on the leader variables that were controlled for. Lastly, the residuals can be used as the outcomes in the regression discontinuity. Thus, the RD controls for the environment while the linear model controls for the leader characteristics that might be correlated with age. Figure 5 shows the results from this procedure in the age example.

A key point to remember here is the importance of having an implied experiment. For example, while it is easy to imagine an experiment where presidential candidates with different ages were assigned to be leaders, it is much harder to think of an experiment where age was randomly assigned to leaders. Without an implied experiment, it is difficult to know what the real counterfactual is. Thus, the age example is not entirely unproblematic, and it would be better to use this method on a leader characteristic that was in theory more manipulable. However, even when the independent variable of interest is an attribute, controlling for some baseline characteristics in this way may alleviate concerns that results are explained by other factors.

Conclusion

This paper has covered several key points, which can be summarized as follows:

(1) The natural experiment approach and continuity approach are mathematically similar despite their conceptual differences. The question is not about which of these approaches researchers should use, but whether they should control for the score, and if so, how. A major advantage of the natural experiment approach is that it can help researchers evaluate their research designs.

(2) In cases where units are grouped together into strata that vary in terms of their competitiveness, researchers should restrict their focus to the pairs of units that are closest to the cut-point—the highest scoring loser and the lowest scoring winner—or else weight the observations based on their probabilities of treatment assignment. Failing to do so can result in bias if the smoother does not accurately capture the relationship between Z and the changing influence of additional bare winners and losers on the estimates.

(3) Units that are in the sample multiple times do not pose a problem if they have one outcome for every score, although they can pose a threat to quantifying the

uncertainty of the estimates. This problem can be managed by dropping certain units from the sample.
(4) When units with multiple scores have only one outcome, researchers should use the Refined Fuzzy RD design, where the instrument is whether the unit won or lost in the first case where it scored in the RD window or above it. It is important to consider whether the exclusion restriction holds in these cases.
(5) RDs do not provide a valid causal estimate of the impact of a characteristic that was not assigned by the RD. They can help researchers determine what will happen if an individual or party with that characteristic barely wins or barely loses. They can also be combined with observational methods to shed light on how the characteristic of interest influences the outcome.

Keeping these points in mind can help researchers avoid bias in applications of RD.

The subtle problems explored in this paper also underscore how important it is for researchers to offer a clear theoretical explanation for why the treatment should be approximately random around the cut-point. Because of the emphasis on balance tests, not much attention is payed to whether the design makes sense in theory. The key is to show that all units close to the cut-point should have roughly the same probability of treatment assignment. The only possible systematic difference would result from the small imbalance in the score, and that could be accounted for by controlling for it.

If researchers take advantage of the natural experiment approach to identify potential problems with their designs, then RD can be a very powerful tool for inference. Scoring systems are a central feature of modern society, used in areas like education, sports, public health, criminal justice, domestic politics, and international relations. Although the scoring system will often be more complicated than the classic example of high-school students taking a test, researchers who are careful about their designs can still use RD to make new and fascinating discoveries.

References

Bernardi, Fabrizio and Emmanuel Skoufias. 2014. "Compensatory Advantage as a Mechanism of Educational Inequality A Regression Discontinuity Based on Month of Birth." *Sociology of Education* 87(2) 74-88.

Bertoli, Andrew. 2017a. "Nationalism and Interstate Conflict: A Regression Discontinuity Analysis." Working Paper. 2015.

Bertoli, Andrew. 2017b. "United We Fight: Democratic Unity and State Aggression." Working Paper. 2015.

Broockman, David E. 2009. "Do congressional candidates have reverse coattails? Evidence from a regression discontinuity design." *Political Analysis* 17(4): 418-434.

Brollo, Fernanda, and Tommaso Nannicini. 2012. "Tying your enemy's hands in close races: The politics of federal transfers in Brazil." *American Political Science Review* 106(04): 742-761.

Buddelmeyer, Hielke, and Emmanuel Skoufias. 2004. *An evaluation of the performance of regression discontinuity design on PROGRESA*. Vol. 827. World Bank Publications.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik. 2013. "Robust nonparametric confidence intervals for regression-discontinuity designs." Manuscript.

Cattaneo, Matias, Brigham Frandsen, and Rocio Titiunik. 2014. "Randomization inference in the regression discontinuity design: An application to the study of party advantages in the US Senate." *Journal of Causal Inference.*

Cattaneo, Matias D., Luke Keele, Roco Titiunik, and Gonzalo Vazquez-Bare. "Interpreting regression discontinuity designs with multiple cutoffs." *The Journal of Politics* 78(3).

Caughey, D., and J. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008." *Political Analysis* 19(4): 385-408.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010."The value of school facility investments: Evidence from a dynamic regression discontinuity design." *The Quarterly Journal of Economics* 125(1): 215-261.

Chen, M. Keith, and Jesse M. Shapiro. 2007."Do harsher prison conditions reduce recidivism? A discontinuity-based approach." *American Law and Economics Review* 9(1): 1-29.

Crost, Benjamin, Joseph Felter and Patrick Johnston. 2014. "Aid Under Fire: Development Projects and Civil Conflict." *American Economic Review* 104(6): 1833-1856.

Do, Quoc-Anh, Yen-Teik Lee, and Bang Dang Nguyen. 2014. "Political connections and firm value: evidence from the regression discontinuity design of close gubernatorial elections." Working paper.

Eggers, Andrew, Anthony Fowler, Jens Hainmueller, Andrew B. Hall, and James M. Snyder Jr. 2014. "On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races." Formerly MIT Political Science Department Working Paper Series 2013-26.

Eggers, A., and J. Hainmueller. 2009. "MP's for Sale: Returns to Office in Postwar British Politics." *American Political Science Review* 103(4): 513-543.

Green, D. P., Leong, T. Y., Kern, H. L., Gerber, A. S., & Larimer, C. W. (2009). "Testing the accuracy of regression discontinuity analysis using experimental benchmarks." *Political Analysis* 17(4): 400-417.

Hall, Andrew B. 2015. "What Happens When Extremists Win Primaries?" Forthcoming, *American Political Science Review.*

Hainmueller, Jens, and Holger Lutz Kern. 2008. "Incumbency as a source of spillover effects in mixed electoral systems: Evidence from a regression-discontinuity design." *Electoral Studies* 27(2): 213-227.

Hopkins, Daniel J. and Katherine T. McCabe. 2012. "After Its Too Late Estimating the Policy Impacts of Black Mayoralties in US Cities." *American Politics Research* 40(4): 665-700.

Imbens, Guido, and Karthik Kalyanaraman. 2011. "Optimal bandwidth choice for the regression discontinuity estimator." *The Review of Economic Studies*: rdr043.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics* 142(2): 615-635.

Keele, Luke, and Rocio Titiunik. 2014. "Geographic boundaries as regression discontinuities." *Political Analysis.*

Lalive, Rafael. 2008. "How do extended benefits affect unemployment duration? A regression discontinuity approach." *Journal of Econometrics* 142(2): 785-806.

Lee, D. 2008. "Randomized experiments from non-random selection in U.S. House elections." *Journal of Econometrics* 142: 675-97.

Lemieux, Thomas, and Kevin Milligan. 2008. "Incentive effects of social assistance: A regression discontinuity approach." *Journal of Econometrics* 142(2): 807-828.

Masicampo, E. J., and Daniel R. Lalande. 2012. "A Peculiar Prevalence of P-Values Just Below 05." *The Quarterly Journal of Experimental Psychology* 65(11): 2271-2279.

Niu, Sunny Xinchun, and Marta Tienda. 2010. "The impact of the Texas top ten percent law on college enrollment: A regression discontinuity approach." *Journal of Policy Analysis and Management* 29(1): 84-110.

Querubin, Pablo, and James M. Snyder Jr. 2013. "The Control of Politicians in Normal Times and Times of Crisis." *Quarterly Journal of Political Science* 8(4): 409-450.

Rao, Hayagreeva, Lori Qingyuan Yue, and Paul Ingram. 2011. "Laws of Attraction Regulatory Arbitrage in the Face of Activism in Right-to-Work States." *American Sociological Review* 76(3): 365-385.

Redmond, Paul, and John Regan. 2013. "Incumbency Advantage in Irish Elections: A Regression Discontinuity Analysis."

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

Titiunik, Rocio. 2009. "Incumbency Advantage in Brazil: Evidence from Municipal Mayor Elections." University of California-Berkeley, mimeo.

Uppal, Yogesh. 2009. “The disadvantaged incumbents: estimating incumbency effects in Indian state legislatures.” *Public Choice* 138 (1-2): 9-27.

Van der Klaauw, Wilbert. 2002. “Estimating the Effect of Financial Aid Offers on College Enrollment: A RegressionDiscontinuity Approach.” *International Economic Review* 43(4): 1249-1287.

Voeten, Erik. 2013. “Does Participation in International Organizations Increase Cooperation?” *The Review of International Organizations* 8(3), 1-24.

Vollaard, B. 2009. “Does regulation of built-in security reduce crime? Evidence from a regression discontinuity approach.” first Bonn/Paris Workshop on Law and Economics.