Non-monotonic Selection Issues in Regression Discontinuity Designs

Jakob de Lazzer

Abstract

The Regression Discontinuity Design (RDD) has become a popular method for program evaluation in recent years. While it is compelling in its simplicity and requires little in terms of a priori assumptions, it is vulnerable to bias introduced by self-selection into treatment or control group. The purpose of this article is to discuss the issue of non-monotonic self-selection, by which similar numbers of individuals select into and out of treatment simultaneously. This kind of selection has not been discussed in detail so far in the literature, and can be hard to detect with the commonly used methods for data-driven RDD specification testing. The focus of this article lies on selection in the context of close elections, since those are popular natural experiments for RDD applications, and because in this context the issue of non-monotonic selection is rarely considered in practise. I will present a slightly modified approach to specification testing, designed to detect non-monotonic self selection and based on the density test by McCrary (2008). In order to demonstrate how RDDs can be affected by the issue, two existing RDD applications are analysed with respect to non-monotonic sorting. In the first, this article follows up and expands on the remarks made by Caughey & Sekhon (2011) about selection issues in the well known RDD application by D. Lee (2008). The second application is based on the Mexican mayoral election RDD by Dell (2015).

1 Introduction

The Regression Discontinuity Design (RDD) has rapidly risen in popularity among researchers in recent years.[1] It allows for causal inference on treatment effects from natural experiments. Much like other causal program evaluation methods the RDD can be biased by endogenous selection and lose internal validity. Therefore, it is of concern to researchers how cases of self-selection can be detected in advance.

In the case of the Regression Discontinuity Design, self selection invalidates the identifying assumption that the sub-populations near the assignment threshold are perfectly comparable in the absence of treatment.[2] When implementing RDDs, it is common practice in the

[1] See Cook (2008).
[2] See Lee (2008).

literature to consider channels of influence trough which units of observation can influence their treatment status and perform data-driven tests of the identifying assumption.[3]

We have to distinguish between monotonic and non-monotonic sorting dynamics. Non-monotonic sorting occurs when some individuals select into treatment while similar numbers of individuals select out of treatment. Such sorting can also happen in the form of forced deselection by an external influence, even if all the individuals have the same treatment preference. In the literature, two kinds of tests for internal validity of the RDD are typically applied: The density based test by McCrary (2008) and cheks for balanced covariate levels. Non-monotonic sorting can not be detected with current implementations of the former and is sometimes difficult to identify with the latter.

I will motivate the importance of finding selection issues in advance, by presenting several likely channels of influence through which individuals can manipulate their treatment assignment in an RDD. This article contributes to the literature by discussing non-monotonic selection in the RDD and developing a modified application of the McCrary specification test which can reliably detect non-monotonic sorting at the threshold. To my knowledge, the problem of non-monotonic sorting dynamics in RDD applications has, not been studied in detail before. The test for non-monotonic selection works by identifying sub-samples of data whose likelihood of sorting in one direction is higher than their likelihood of sorting in the other. I then perform the density analysis on these sub-samples. If the sub-samples display an uneven density at the threshold while the full sample does not, then non-monotonic sorting is present at the threshold.

In order to illustrate the considerations for using and the workings of the specification test, I have applied it to two RDD analysis, one by Lee (2008), about the incumbency advantage in United States Senate elections. By applying the test to this dataset, I follow up on the findings of Caughey & Sekohn (2011), which indicate that results of close elections for US congressmen are not as randomly distributed as one would expect them to be. Also, this application illustrates that selection problems can be present even in well-established RDDs and in environments where one would, at first glance, think them unlikely. The second application is based on the first stage RDD of Dell (2015), which is also an RDD which exploits close elections.

In election settings, such as this, it is not intuitively obvious why non-monotonic selection should be an issue. The individual units of observation only have incentives to attain higher election outcomes and therefore sort themselves monotonically. However, the data contains only candidates from one party. In this case, successful monotonic sorting by each party's candidates amounts to non-monotonic sorting in the analysed sample. When applying the modified test to the election data, a suitable sub-sample which is likely to be more successful at sorting themselves above the threshold, are those candidates who's party was already in office at the time of the assignment process. The results from testing of this sub-sample indicate that a degree of sorting appears to be present. The magnitude of the estimated density discontinuity depends in part on the exact specification of the test,

[3] See Imbens & Lemieux (2008).

but the overall indications point strongly towards sorting effects. The reason why such sorting should be present is not clear cut. I arrive at the conclusion that no single factor seems to be primarily responsible, but rather a the cumulative effect of several actions with individually limited influence on election results.

The remainder of the text is organised as follows: The next section provides a quick overview of the bias introduced by endogenous selection in the Regression Discontinuity Design. It also establishes the distinction between monotonic and non-monotonic selection. This is followed by a description of the density based validity test and the modification which enables it to detect non-monotonic selection, in Section 3. Section 4 is dedicated to selection issues in the RDD application about incumbency advantages and also provides possible explanations for the testing results. Finally, conclusions resulting from the discussed properties of the test and applications are drawn.

2 Monotonic and non-monotonic selection in the RDD

The RDD exploits discontinuous rules, or events with discontinuous effects, to estimate local average treatment effects (LATE). Treatment is assigned according to a deterministic function, which is often a policy, law, or institutional program which assigns resources or sanctions. In addition to the outcome variable and treatment status, an independent variable is observed. This is also called the running, assignment or forcing variable. Selection into treatment is determined by a function of this variable.[4] In the Sharp Regression Discontinuity Design, assignment is completely determined by this function. In the Fuzzy Regression Discontinuity Design, the value of the running variable only partly determines participation in treatment. Since the same sorting dynamics create identification issues for both Sharp and Fuzzy RDD, this section will focus on the Sharp version of the design.

Let $X \subset \mathbb{R}$ denote the assignment variable, with $x_i \in X$ the realization of this variable for individual i and $y_i \in Y$ the outcome variable. If an individual's realization of X is above a specific threshold value c, then the individual is assigned treatment. Let $I_i(x_i) \,\epsilon\, [0, 1]$ denote treatment status. This treatment assignment mechanism implies that no overlap exists between treatment and control groups in terms of the independent variable X.

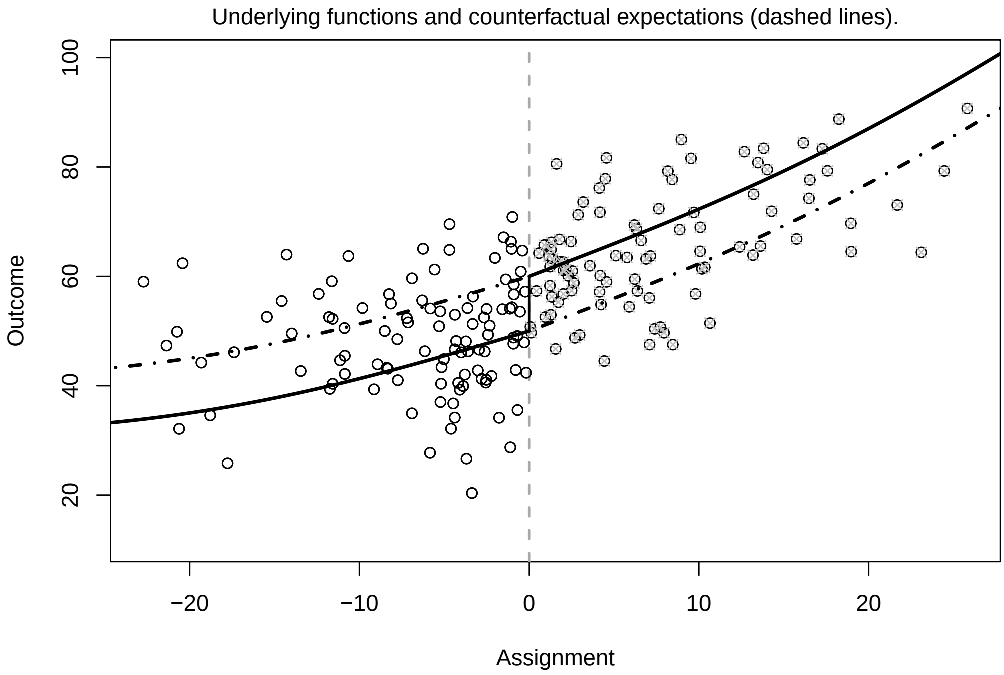
If the location of c is determined exogenously, individuals with very similar realizations of the assignment variable are likely to be similar in those characteristics which determine the outcome in the absence of treatment. In the limit, when comparing individuals directly at the threshold, the control individuals should, on average, be perfectly comparable to those receiving treatment. Identification of the LATE requires an assumption about the smoothness of counterfactual outcomes at the threshold: The conditional expectation functions $E[y_i(1)|x_i = c]$ for treated and $E[y_i(0)|x_i = c]$ for non-treated individuals must be continuous in c.

[4]For the purpose of this paper, only a single assignment variable is considered. An extension of the RDD with multiple assignment variables is discussed in Papay, Willet and Murnane (2011).

When this assumption holds, the LATE is identified as: $LATE = \lim_{x \downarrow c} E[y_i|x_i = x] - \lim_{x \uparrow c} E[y_i|x_i = x]$

Figure 1 shows an example of a fictional RDD, where the counter-factual expectations are smooth across the threshold: The above assumption is fundamental for causal inference

Figure 1: Counter-factual expectations, Sharp RDD



Note: Underlying functions in solid lines, counter-factual expectations in dashed lines.

from RDD results. Violation of this assumption skews the RDD estimates with systematic selection bias of unknown magnitude and makes LATE estimates invalid.

In an applied setting, it might not be intuitively clear what would cause the continuity condition to be plausible. In its pure form, it is empirically untestable. To remedy this problem, Lee (2008) has linked the continuity assumption to the degree of control that individuals have over their realization of the assignment variable.

In many empirical settings, observed persons have some control over their realization of the assignment variable. They will take action to influence their realization of the assignment variable in accordance to their personal motives and underlying abilities. If individuals only roughly influence the assignment variable, then this function will include a stochastic error component. Individuals have imprecise control over X, when the density of X conditional on characteristics is continuous. This characteristic enables the empirical specification test discussed in section 3.

With imprecise control, treatment assignment in an area close to the threshold is “as good as randomized”, meaning that the probabilities of having a value of X slightly above or below the cut-off are the same for an individual with given characteristics.[5] The continuity

[5] Lee (2008), p. 676.

assumption for the potential outcomes $y_i(I_i)$ is satisfied as a consequence of random assignment near the threshold. [6]

In many applications it is assumed that all individuals have the same preference regarding treatment status. There might be a clear benefit from participation or non-participation. If individuals have uniform treatment preferences and the ability to precisely manipulate the assignment variable, then they will only shift their realization of the assignment variable in one direction.

However, this is not true for all applications. What I call Non-monotonic manipulation occurs when some individuals realization of the assignment variable is shifted in one direction, while that of others is shifted towards the opposite. This can happen when the population consists of heterogeneous groups with different preferences regarding treatment assignment. A situation where this kind of manipulation was suspected was the introduction of the new German parental leave benefit (Elterngeld) on Jan. 1. 2007. The reform created incentives for some parents to postpone the birth of their child and for others to accelerate it. Birth-shifting to exploit cut-off dates is often considered unlikely, but the results of Tamm (2013) , Dickert-Conlin & Chandra (1999) and Gans & Leigh (2009) indicate that it is actively practised. This finding is of high importance, since several articles about policy evaluation use the timing of births as the cutoff for RDD analysis.[7] In this case, Tamm (2013) finds evidence of selection into the new parental leave system, but the results for the group of parents who are expected to profit from the old system are less clear.

If both groups are of comparable size, similar numbers of individuals sort themselves to each side of the threshold. Therefore, the manipulation taking place at the threshold will not result in a jump in the density of the running variable, while still leading to systematic differences between treated and control groups.

Individual treatment preferences are not the only source of non-monotonic selection. It can also occur when realizations of the assignment variable are precisely manipulated by outside forces with contrasting preferences. In the application of Section 4, the sample individuals have strictly monotonic treatment preferences and some of them appear to be able to shift their assignment variable slightly above the threshold. For some other individuals however, their assignment variable is precisely manipulated to slightly below the threshold by non-sample individuals with opposing preferences. Both mechanisms can lead to problems with internal validity of the RDD, because the commonly used specification tests have trouble detecting non-monotonic manipulation.

3 Specification Testing

Since sorting dynamics in RDDs are often not immediately apparent, data driven specification tests are commonly used to rule out selection bias.

[6] It is important to note that some forms of random components in the running variable are not sufficient for the continuity assumption to hold. If the random component is censored at the threshold, endogenous sorting may still be a threat to the validity of the RDD.

[7] See for example Dustmann & Schönberg (2011) and Lalive (2008).

Two general data-driven approaches to testing RDD assumptions are available. One is to check observed covariates for smoothness at the threshold. If clear imbalances in covariate levels exist between individuals slightly below and above the threshold, the smoothness of counterfactual outcomes is unlikely. This approach relies on the availability of high-quality covariate data relevant to potential selection dynamics, which is often not available. In those cases where the characteristics affecting selection are unobserved or mismeasured, covariate tests can not rule out sorting dynamics.

A more generally applicable method for testing the identifying assumption of an RDD is the density based test developed by McCrary (2008). Applying this test to the entire sample will not detect non-monotonic sorting at the threshold. But the method used for finding discontinuous jumps in the density is also used for testing sub-samples to identify non-monotonic sorting.

The object of analysis for this test is the density function of the assignment variable. Uncensored random components in each individual's value of X imply continuity of the cumulative distribution function, conditional on underlying characteristics of each individual. And therefore imply continuity of the conditional density of the assignment variable. Continuity of the conditional density also implies continuity of the overall density of the assignment variable across the population.

If precise sorting or other types of non-random selection into treatment take place at the threshold, the density of X will not be smoothly distributed at the cut-off. It is therefore possible to check for violation of the identifying assumption by testing for continuity of the density function of the running variable at the threshold. This is done by estimating the size of a potential discontinuity in the density at the cut-off, which, in principle, is similar to performing a RDD-analysis on the density function of the assignment variable, with the treatment effect being equivalent to the deviation from continuity. Consequently, the techniques used for the density based specification test are closely related to those used in conventional RDD settings.

Under certain conditions a variation of the density testing procedure can be used to detect non-monotonic sorting. The size of the bias introduced by this kind of sorting dynamics is in direct proportion to the share of the two subgroups with contrasting treatment prefrences in the sample. Within each treatment-preference group, a discontinuity in the density of observations would be present at the threshold. Therefore, testing separately for each sub-group would allow the researcher to discover these sorting dynamics.

Precise identification of the subgroups can be challenging, since the mechanics of manipulation and the individuals involved are rarely observable. If they were, data driven tests would not be required. If group membership can not be precisely determined for each individual, it is still possible to find evidence of non-monotonic manipulation. For this purpose it is sufficient to identify elements of the population for which the probability of belonging to one group is higher than that of belonging to the other group. As long as one of the sub-groups is over-represented in the tested sample and the sorting dynamics are sufficiently strong, the density test can detect those dynamics.

This approach requires additional information about the individuals compared to the straight density test, in order to determine group membership. It does however have two distinct advantages over simple tests of covariate smoothness. First, covariate data does not need to be of high quality and missing observations pose less of a problem. For example, categorial data can be used to determine subgroups suspected of self-selection. Second, this approach is helpful when self selection can not be identified in terms of a single covariate level but instead depends on interactions between covariates.

3.1 Estimation

For the specification test, a histogram of the density function is created, by finely binning the running variable and assigning the frequency counts to the bin midpoints. The bins are constructed in such a way that no bin contains values of X from both sides of the threshold.

Then a Local Linear density smoother is applied separately to the histogram on each side of the threshold.[8] [9] A kernel-weighted linear regression is applied to small sections of the data. Each section is defined by an evaluation point x_0 and the bandwidth h. The bin midpoints are used as regressors and the counts per bin, as regressands. [10]The kernel function that is most beneficial for RDD-applications is the triangle kernel, which shows optimal performance at boundary points.[11][12] Weights are assigned in a linear way, with the peak of the weight distribution at the evaluation point. At the boundary, the weight distribution is truncated and its peak lies at the boundary point itself.

A potential discontinuity in the density function will be found by performing separate regressions on both sides of and estimating the outcomes at the cut-off. The discontinuity would show up as the difference of the boundary estimates at the threshold being significantly different from zero.

The specification test is then performed as a Wald-Test with the null-hypothesis that the jump in the density is zero.

It is necessary to select two tuning parameters for the estimation process: The size of the histogram bins and the bandwidth for Local Linear estimation.

The binsize has only minor effects on the results. In most applications, the estimator described above is very robust to changes in binsize, under the condition that a sufficient

[8]A detailed discussion of the asymptotic properties of local linear estimation can be found in Fan and Gijbels (1996). It has been shown by Hahn, Todd and Van der Klaauw (2001) that, for the purposes of the RDD, local linear estimation is highly efficient.

[9]As discussed by Lee and Card (2008), the treatment effect is asymptotically not identified for non-parametric estimation without functional form assumptions in conventional RDD applications with discrete running variables. However, this issue is not present in the density based specification test, if the running variable has continuous support. The binned running variable is not discrete in the conventional sense, because it can be defined by the researcher and the bin width can asymptotically shrink to zero when the data density approaches infinity.

[10]A detailed description of the Local Linear estimator as described in McCrary (2008) is included in the Appendix section A.3.

[11]See Cheng, Fan and Marron (1997).

[12]See Lee and Lemieux (2010) for a discussion about the merits of different kernels in the RDD.

number of bins are covered by the bandwidth of choice.[13]I employ the binsize selection rule suggested by McCrary (2008), which is a variation of the widely used Scott's rule for binsize selection.[14]

Both more critical and more difficult is the choice of the bandwidth. It is a measure of the flexibility of the local linear model. For each evaluation point, the bandwidth determines which bins, and therefore which observations, are used for the point estimator. In RDD applications, this means that the bandwidth determines how close to the threshold the data is evaluated for the discontinuity estimate.

The choice of bandwidth is essentially a trade-off between reduced bias and precision of the estimates. A small bandwidth for Local Linear estimation will result in a better approximation of the underlying function and reduce the bias, since only observations closer to the cut-off are used for estimation purposes. However, the estimate will be based on a smaller number of observations, which will reduce the precision of the result. [15]

Bandwidth choice for non-parametric estimation has been analysed in detail in the literature and a number of solutions have been proposed.[16]When ease and speed of implementation is a priority, as it is in the case of specification testing, so called 'rule of thumb' (ROT) bandwidth selectors are commonly used. A ROT bandwidth for the special case of density estimation at boundary points has been proposed by Fan, Gijbels (2006) and by McCrary (2008). Using the suggested procedure, I fit a fourth-order polynomial model to each side of the histogram and choose the bandwidth depending on the mean squared error and the curvature of the fitted model.[17] However, since the suggestions for the best bandwidth selection technique vary wildly in the literature, I treat the ROT bandwidth as a starting point and calculate tests for a wide range of bandwidths. The results which we can be most confident in, do not depend on specific choices of the tuning parameters.

4 Empirical Applications

The applications analysed in this section are an RDD by D. Lee (2008), about the political landscape in the United States, and an RDD by Dell (2015), about the effects of the partisanship of mayors on violent crime in Mexico.

Lee determines the inherent vote share advantage which candidates for the House of Representatives receive if their party is the incumbent at the time of election. The

[13]This robustness has been formally shown by McCrary (2008) and the results found in Section 4 are in line with those conclusions.

[14]The suggested binsize is $b = 2\hat{\sigma}N^{-\frac{1}{2}}$, with $\hat{\sigma}$ being the standard deviation of the assignment variable in the sample. See Scott (1979).

[15]As part of a discussion of the asymptotic properties of local linear estimation at boundary points, it has been shown by Hahn, Todd and Van der Klaauw (2001) that the optimal bandwidth converges to zero at a rate of $N^{-\frac{1}{5}}$, when the sample size approaches infinity. Implying that the bandwidth should be proportional to $N^{-\frac{1}{5}}$.

[16]See Pagan and Ullah (1999) for practical results from subjective bandwidth choice. Cheng (1997) and Imbens, Kalyanaraman (forthcoming) for presentations of plug-in methods. Fan and Gijbels (1996) for a "rule of thumb" for bandwidth selection. And Ludwig & Miller (2005) for a cross-validation technique.

[17]See Appendix A.4 for a description of the ROT.

hypothesis is that, individual characteristics being equal, those candidates whose party is in office at the time of the election have advantages over their competitors in terms of the vote share. Specification tests for the whole sample reject the presence of sorting effects. However, when applying the sub-sample test for non-monotonic sorting, the results indicate some level of precise selection at the threshold. This is an instance where the selection preferences of sample individuals are monotonic, but where the assignment variable can be subject to manipulation by exogenous agents, in this case the opposing Republican candidates.

Dell (2015) treats the average vote share of different party candidates as a proxy for socio-economic characteristics of the mayoral district. The hypothesis is then that districts in which a party barely won are, on average, comparable to those where the party barely lost. In contrast to the article by Dell, who uses election results for the Partido Acción Nacional (PAN), I consider election results for the Partido Revolucionario Institucional (PRI), which allow for higher numbers of observations. While Mexican elections are, in principle, not immune to experiencing manipulation of vote shares, results of the specification test indicate no sorting issues in this sample. [18] Both applications showcase why we should be aware of the unexpected ways in which non-monotonic sorting can affect RDDs.

4.1 Testing the full sample of the incumbency Regression Discontinuity Design

Treatment, in the form of incumbency, is assigned when the vote share difference of a party crosses the threshold at zero percent. The vote share difference is defined as the percentage difference in vote shares between a candidate and his next closest contender. This value is positive for the winner of the election and negative for the losers. Only candidates of the Democratic Party are included in the sample. The results for Republican candidates are expected to reversely mimic those of the Democrats in the majority of cases.

The assignment variable is the vote share difference at time t. This value is centred by definition, so that the threshold value c lies at zero. All districts with Democrat vote share differences to the right of the cut-off have Democrat incumbents at the time of the next election in period $t+1$. The indicator $I_{it+1} = 1[VS_{it} \geqq 0.5]$ describes the incumbency status of the candidate's party.

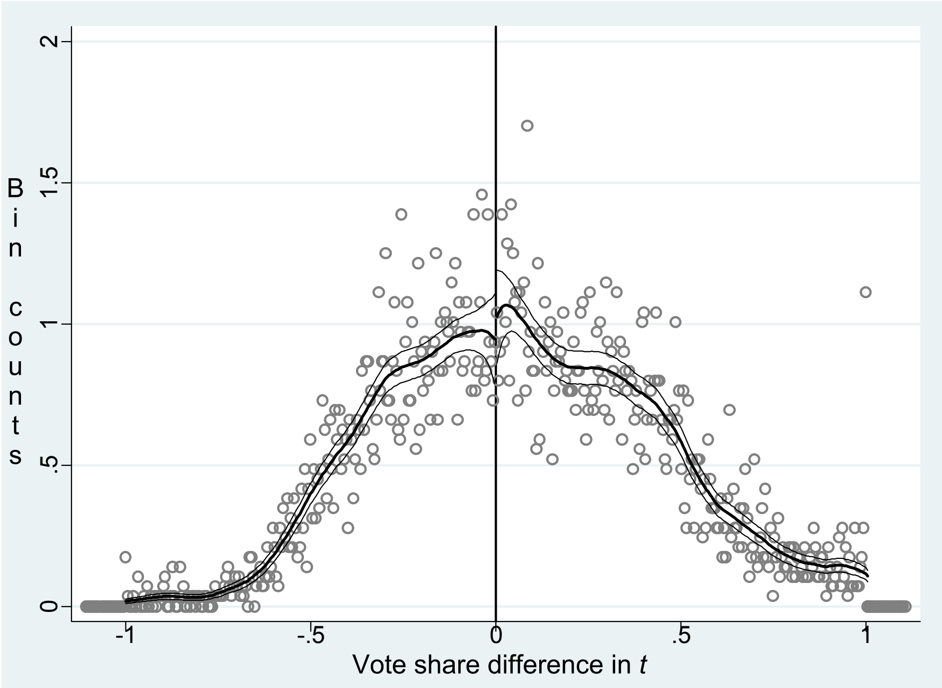
The outcome variable is the party vote share in the election at time period $t+1$. In the application, Democrat vote share in the following election (VS_{it+1}) is regressed on the Democrat vote share difference in the previous one and on a vector of candidate characteristics (w_{it+1}).

$$VS_{it+1} = \alpha_{t+1} w_{it+1} + \beta_{t+1} I_{it+1} + \gamma_{t+1} VS_{it} + e_{it+1} \tag{1}$$

with $E[e_{it+1}|w_{it+1}, VS_{it}] = 0$. The RDD is necessary because w_{it+1}, VS_{it} and I_{it+1} are all correlated with w_{it}. By performing parametric regressions separately on both sides of the

[18] Compare Grant (2012)

Figure 2: Density estimates for Democratic candidates



Note: $b = 0.0048$, $h = 0.1061$, 95% confidence bands in thin lines.

cut-off, Lee (2008) finds that there is an incumbency advantage of about 7.8 percent of the vote share in the data.

The identifying condition of $f(VS_{it}|w_{it})$ being continuous in VS depends on the assumption that election results contain a substantial random component, because many factors influencing election outcomes are beyond any candidate's control. For example, weather or traffic conditions can influence election turnouts. Limited opportunities for precise manipulation in political elections have been known however, and I discuss them in Section 4.5.

A small range of available covariates, past political experience, number of election runs, party vote share in $t-1$ and the probability of the party winning the election in $t-1$, show balanced levels within a 5 percent margin of the cutoff.

When applying the density based specification check described in Section 3 to the data, no significant sorting can be detected.[19][20] To establish the robustness of the results for different values of the tuning parameters, I performed the test with the reference binsize and bandwidth, as well as fractions and multiples of both reference values. Table 1 shows the discontinuity estimates for all combinations of tuning parameters and Figure 2 shows the fitted model. The version in this graph provides, upon visual inspection, the best approximation of the data close to the cut-off of all tested variations. [21]

The results indicate a very smooth distribution of election results at the threshold. No discontinuity estimate exceeds two standard deviations and the estimated differences in

[19] A dataset containing the information for the Lee study has been obtained from the Mostly Harmless Econometrics Data Archive: http://economics.mit.edu/faculty/angrist/data1/mhe (last visited 15.05.2014).

[20] The sample is trimmed at the extreme ends of the forcing variable to remove outliers and improve the clarity of plots without affecting local linear estimators.

[21] Related testing is performed in chapter IV of McCrary (2008).

Table 1: Estimated discontinuities in the density at the threshold, test results for the full sample of Democratic candidates

Results are estimated discontinuities in the density at the threshold, standard errors in brackets, p-values in italics.

	Quarter the reference bandwidth 0.0531	Half the reference bandwidth 0.1061	Reference bandwidth 0.2123	Twice the reference bandwidth 0.4245	Four times the reference bandwidth 0.8491
Reference binsize 0.0097	0.1359 [0.1871] *0.4676*	0.0844 [0.1247] *0.4985*	0.1258 [0.0856] *0.1417*	0.0484 [0.0605] *0.4237*	-0.0108 [0.0414] *0.7942*
Half the reference binsize 0.0048	0.1004 [0.1851] *0.5875*	0.0800 [0.1245] *0.5205*	0.1163 [0.0856] *0.1743*	0.0454 [0.0605] *0.4530*	-0.0116 [0.0414] *0.7793*
Twice the reference binsize 0.0194	0.1831 [0.1827] *0.3163*	0.0968 [0.1247] *0.4376*	0.1304 [0.0856] *0.1277*	0.0484 [0.0604] *0.4229*	-0.0104 [0.0414] *0.8017*

log-densities at the threshold range between 0.0104 and a maximum of 0.1831. For the entire range of bandwidths and binsizes, t-tests do not reject the null hypothesis of a smooth distribution. On the aggregate level, the density function is continuous at the threshold.

4.2 Testing the sub-sample of incumbent Democratic candidates

In a recent article, Caughey and Sekhon (2011) have questioned whether the outcome of close elections to the U.S. House really is as randomised as Lee (2008) assumes. They show that a number of relevant covariates are not well balanced at the threshold.[22] Covariate imbalance is greatest away from the threshold, diminishes when looking at observations closer to the threshold, and increases again for extremely close elections. They report that covariates become more balanced within shrinking margins around the threshold, down to a margin of five percent. This finding is in line with the results from Lee (2008). However, for smaller margins, especially those of less than one percent, covariates become less balanced.

As causes for this behaviour, Monotonic manipulation of the running variable is ruled out by the density based test. Non-monotonic sorting issues in the sense that equal numbers of individuals with opposing treatment preferences sort themselves to each side is also not

[22]These covariates include, among others, the political experience advantages for Republican and Democratic candidates, campaign money spent and donation funds received.

possible, because the sample of Democratic candidates have strict monotonic treatment preferences.

An explanation could be a combination of monotonic manipulation by sample individuals and external forces with different treatment preferences: From the perspective of each party, manipulation is always strictly monotonic positive, since winning the election is the primary goal of any candidate running for office. However, the aggregate situation is not as clear, since both Democrat and Republican candidates engage in manipulative activities.[23] If a Democrat party candidate was able to precisely control his vote share, he would realize a vote share margin of victory marginally above the threshold. As a direct consequence, his Republican contender would receive a vote share slightly below that of the Democrat candidate, and therefore marginally below the threshold.

If some candidates from one party have the ability and opportunity to manipulate their vote shares, we have to assume that the other party would possess the same capabilities. Consequently, a number of Republican candidates would also be able to precisely manipulate their vote share. Those candidates would win a disproportionate number of close elections, causing a similar number of Democrat candidates to barely lose the elections. This would lead to a discontinuous jump downwards in the density of Democrat vote shares.

If comparable amounts of successful precise manipulation were achieved by both Democrats and Republicans, the effects would mask each other over the entire sample and make detection by means of the density test impossible.

We can not identify in which elections which candidate might have successfully engaged in precise sorting. However, it is enough to identify sub-samples of candidates who have an above-average probability of precisely manipulating their assignment variable or having it manipulated by the opposing candidate.

One such sub-group would be those candidates whose party already was the incumbent party at the time of the election which determines assignment. This is in line with the finding that the covariate imbalances in close elections found by Caughey and Sekhon (2011) are especially pronounced between candidates running for the incumbent party and the candidates of the challenging party. The incumbent party is more deeply interwoven with the administrative institutions and therefore has potentially greater influence on the election process.

Another possible subgroup with a higher chance of successful manipulation would be those candidates who's party holds the office of secretary of state, who is in charge of the elections, or who's party provides the state governor.[24]

One might ask if the increase in the probability of winning of incumbent party candidates is not just the expected effect of the incumbency advantage from the previous election. Indeed, when looking at the aggregate of all incumbent party candidates, they have substantially higher chances of winning the next election. However, under the identifying

[23] For this argument, a strict two-party system is assumed. This assumption closely but not entirely reflects the political realities of post-war elections to the U.S. House of Representatives.

[24] In the case of the state governour, I could not detect similar evidence of sorting mechanisms.

assumption of the RDD, this should not be true for close elections. Instead, incumbent party candidates should be winning more often with higher margins of victory, since the incumbency advantage is reflected in a higher vote share.

If the incumbent party candidates are more successful at shifting their vote share precisely upwards, the mechanics leading to a discontinuity in the subgroup density would work as follows:

I take Equation 1 as starting point, with the vote share at $t+1$ as the outcome of interest for the RDD. For the specification test, the density of the democrat vote share at time t is analysed. And $t-1$ is the election which determines incumbency status for the purpose of subgroup testing. The density of the winners of the election in $t-1$, who will be the incumbents in time period t, $f(VS_{it-1}|I_{it-1}=1)$, is truncated at zero (compare Figure 3).

If no selection process is at work, then the distribution of election results in the next period, t, will appear like that of Figure 4. The results for incumbent party candidates, the winning party of the election in $t-1$, are concentrated at the upper end, because β_t, the vote share advantage from incumbency, shifts them upwards. The model determining vote share for this election follows the same concept as Equation 1:

$$VS_{it} = \alpha_t w_{it} + \beta_t I_{it} + \gamma_t VS_{it-1} + e_{it}$$

Under the no-sorting assumption, since individual characteristics w are continuously distributed, $f(VS_{it}|w_{it})$ is continuous in VS. The density of VS_t is smooth across the threshold for all groups of candidates.

If however the non-monotonic sorting dynamics described earlier is present, then we would expect densities like to those in Figure 5. Incumbent candidates of both parties have a higher chance of winning close elections. For our sample of Democrat candidates, this leads to discontinuous jumps in the density of both the winning and losing party candidates of the previous election. For the winning ones, the discontinuity of value $\delta_1 > 0$ is caused by their ability to influence close elections in their favour. For the losing party candidates, the discontinuity of value $\delta_2 < 0$ is caused by their opponents ability to win close elections.

Over the density of the entire sample of Democrat candidates, a discontinuity of size $\delta = \delta_1 + \delta_2$ is present. When both parties are very similar in terms of average political influence over time, both discontinuities cancel out and the density for the full sample does not show a gap at the threshold.

Figure 3: Determination of incumbent status in t-1

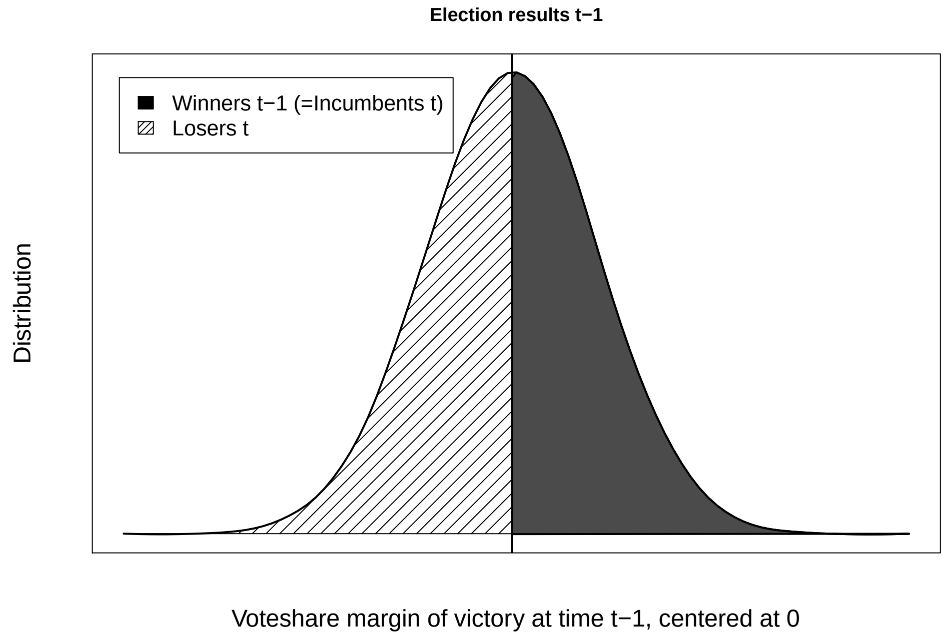
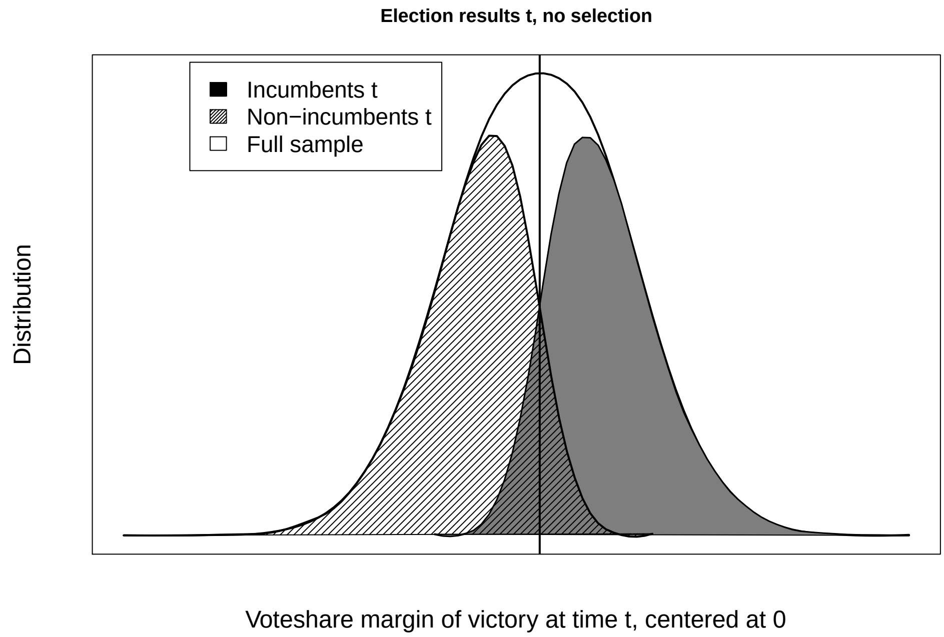


Figure 4: Assumed distribution at t without sorting



4.3 Results of sub-group testing

These results from the test on the sub-sample of incumbents at time t differ sharply from the ones in for the aggregate sample. Estimated discontinuities in the density at the cut-off are considerably larger for all bandwidths and binsizes. The estimates vary in size depending on the choice of tuning parameters, generally exceed two standard deviations and are always larger than one standard deviation. Plotting the local linear smoother over the histogram in Figures 6 and 7 shows a sharp downturn in the chance of barely loosing an election for the sub-sample. This indicates that of the incumbent Democrat candidates,

Figure 5: Distribution at t with sorting

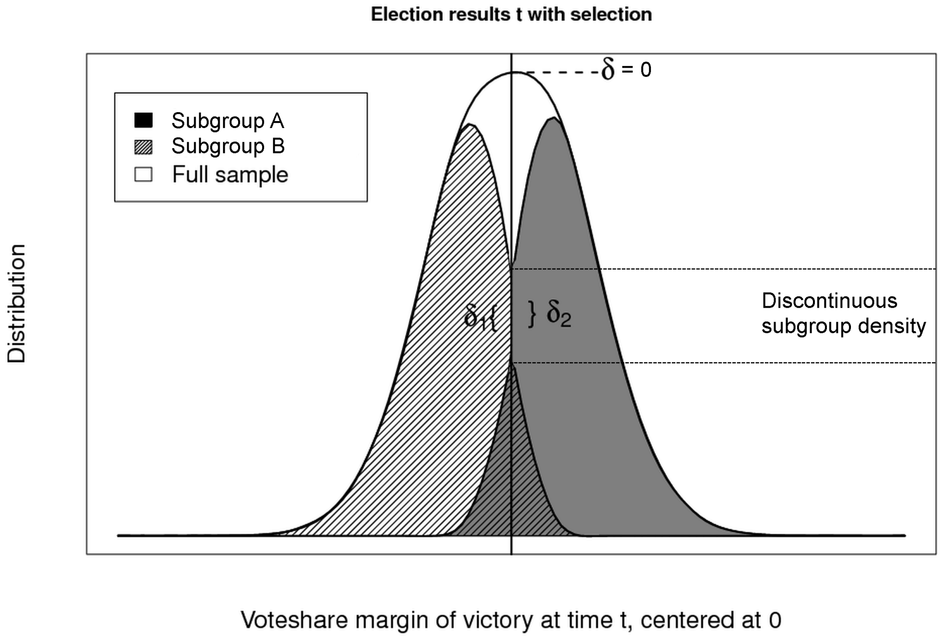


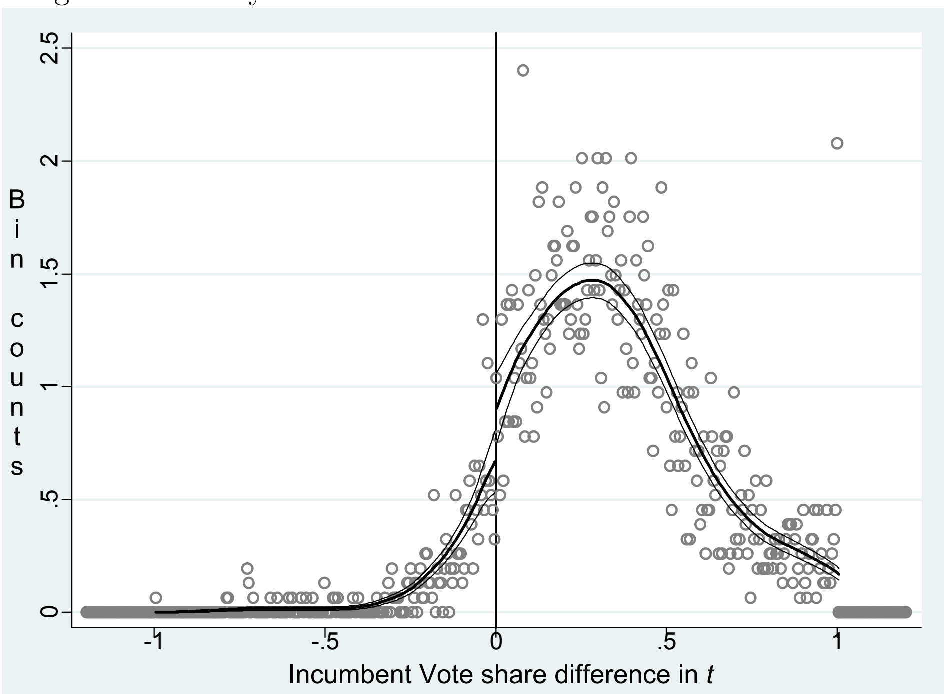
Table 2: Estimated discontinuities in the density at the threshold, test results for the sub-sample of Incumbent-party candidates

	Quarter reference bandwidth	Half reference bandwidth	Reference bandwidth	Double reference bandwidth
Discontinuity	0.882	0.344	0.310	0.542
Standard Error	0.356	0.216	0.143	0.103
P-value	0.013	0.111	0.030	0.000
Bandwidth	0.049	0.098	0.196	0.391

the vast majority wins the elections they are running in, strengthening the notion that incumbent party candidates on average possess superior means of securing election wins. Especially when restricting the analysis to very close elections, by selecting a bandwidth below one percent of the vote share difference, the average chance of winning the election is significantly higher if the candidate's own party is in power.

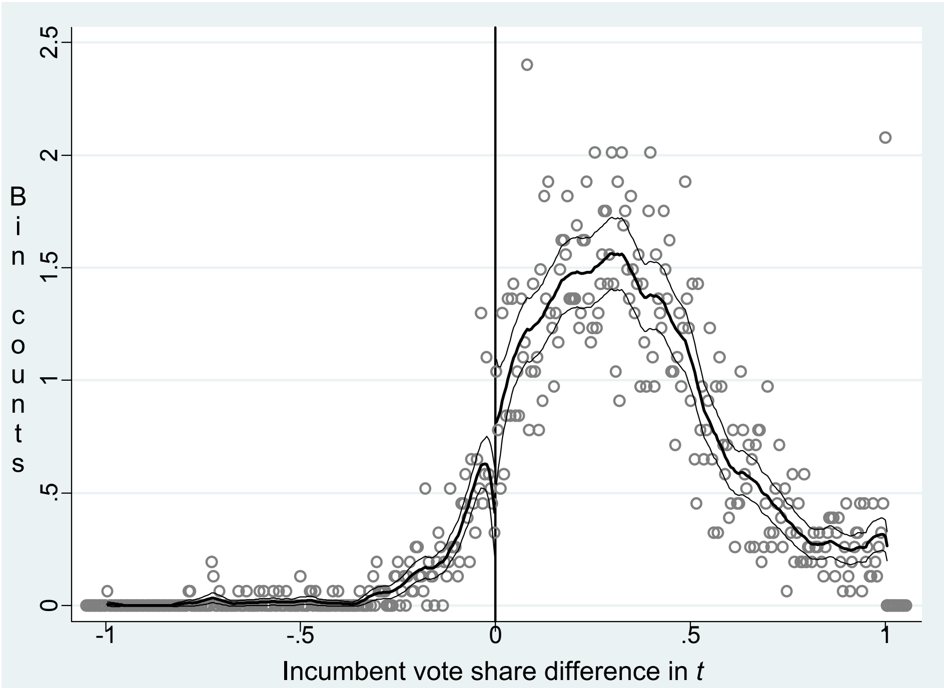
Some variation is visible in the results, depending on choice of the bandwidth. Because of this sensitivity, I performed the test for a finely gridded range of bandwidths ranging from 0.02 to 0.25, maintaining the reference binsize of 0.0098 (Figure 8). As would be expected, precision of the estimates degrades with shrinking bandwidths, due to lower observation counts available within the bandwidth. For bandwidths larger than of 2% of the vote share, as well as for bandwidths smaller than one percent, significant discontinuities are estimated. As with previous applications, the results are relatively stable under variations in binsize. The null hypothesis of continuity of the vote share difference is not rejected for half the reference bandwidth, even though it is rejected at all other bandwidths. Figure 8 shows a more detailed view of this phenomenon. Tests for the discontinuity being non-zero are significant at the 5% level for all bandwidths smaller then 0.08 and larger then 0.2. When considering 10% significance levels, the null is rejected everywhere except for a small range of bandwidths between 0.11 and 0.125. The graph of the significance level exhibits a hump in the area of the halved reference bandwidth. While the discontinuities are not always significant at very high levels for all bandwidth choices, a sharp increase in the differences between treated and control candidates at the threshold, compared to the full sample analysis, can not be denied. This strongly hints at substantial differences in the behaviour at the boundary between incumbent party candidates and candidates of the challenging party. Even more important, the estimated jump in the density at the cut-off actually increases for very small bandwidths, when only data from the closest elections is used. This result is in line with the findings of Caughey and Sekhon (2011), who report that differences in covariate values increase for extremely close elections with vote share differences of one percent or lower, after having converged before with shrinking margins.

Figure 6: Density estimates for incumbent Democratic candidates



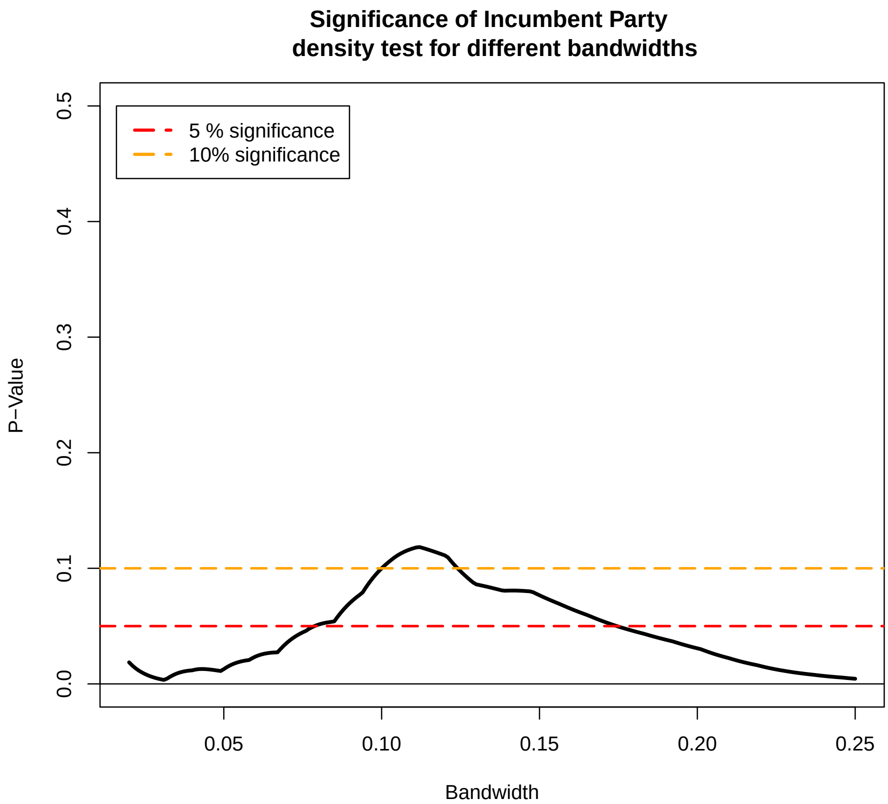
Note: $b = 0.0049$, $h = 0.2014$, 95% confidence bands in thin lines.

Figure 7: Density estimates for incumbent Democratic candidates



Note: $b = 0.0049$, $h = 0.0504$, 95% confidence bands in thin lines.

Figure 8: Significance levels depending on bandwidth



4.4 Application for Mexican mayoral elections

In Dell (2015), the author uses an RDD to analyse the causal effect of mayor partisanship on drug related homicides in Mexico from 2007 to 2010. When conservative president Felipe

Calderon came to power in 2006 his party, the Partido Acción Nacional (PAN) spearheaded the "war on drugs". Municipalities which were won by PAN mayors experienced more frequent and effective police activity against drug trafficking organisations. Dell's RDD strategy is based on the concept that, on average, municipalities in which the PAN barely won the mayor's office are comparable with municipalities in which it barely lost. Exploiting this natural experiment, the article shows an average difference in drug related homicides of 33 per 100.000 inhabitants at the vote share threshold.

As with the application by Lee (2008), sorting dynamics can not be ruled out a priori in this electoral RDD. Surrounding characteristics of the elections are different, with mayoral elections during the observed time frame not being balanced between two parties. Mexico has three prominent partys, with the PAN, the Partido Revolucionario Institucional (PRI) and the Party of the Democratic Revolution (PRD) alternating in strength across municipalities. Within the sample, the PRI wins 59% of the elections and the PAN 24%. Therefore the mechanics of the relatively strict two party system in the Lee example are no longer present. Since the three parties are, on average, not equal in political strength, sorting into treatment is less likely to be masked by equal magnitudes of sorting out of treatment, if non-monotonic selection takes place. The full sample density test has higher chances of detecting sorting behaviour.

The replication files provided in the online appendix of Dell (2015) are limited to elections within a +5% and -5% vote share interval around the threshold. I perform the density based test analogous to Sections 4.1 and 4.2 in first for the the full sample of PAN candidates, with the ROT selected bandwidth and a range of tuning parameters spanning 2% to 5% of the vote share.

Table 3: Density test results for all PAN candidates

	Refrence Bandwidth	Bandwidth 3 % vote share	Bandwidth 4 % vote share	Bandwidth 5 % vote share
Discontinuity	-0.412	-0.221	-0.133	-0.039
Standard Error	0.314	0.246	0.218	0.194
P-value	0.190	0.368	0.542	0.840
Bandwidth	0.020	0.030	0.040	0.050

The results in Table 3 show no indication of selection issues. Fitting a linear model on the 5% vote share bandwidth, the estimated discontinuity is almost zero. At all bandwidths, the discontinuity is not significant. Although it is not small at the reference bandwidth the confidence bands increase with smaller bandwidths and therefore diminishing numbers of observations. If the PAN was capable of deciding significantly more or less close elections for itself than the other two parties combined, it would show up as a discontinuity.

Table 4: Density test results for for PAN incumbent candidate sub-sample

	Reference Band-width	Bandwidth 3 % vote share	Bandwidth 4 % vote share	Bandwidth 5 % vote share
Discontinuity	-1.953	-0.711	-0.402	-0.276
Standard Error	1.395	0.600	0.469	0.401
P-value	0.161	0.236	0.392	0.491
Bandwidth	0.017	0.030	0.040	0.050

When restricting the sample to those candidates in whose municipalities the PAN was already in office at the time of election, analogous to the incumbency sub-sample of Section 4.2, Table 4 shows a similar picture to the results in Table 3. For all bandwidths, the discontinuity is not significant at any level, but it is larger across the board, compared with the full sample estimates. These results reinforce the testing done by Dell (2015), who reports no hints of sorting behaviour in Mexican mayoral elections. Even though the potential for precise manipulation of the vote share is not not lower in Mexican mayoral elections,

4.5 Discussion

The results from the sub-sample specification test performed in section 4.3 suggest the presence of non-monotonic sorting issues in this particular RDD application. Such a result poses the question why it should be possible for incumbent party candidates to influence close elections in a way which systematically increases their chances of victory. A number of possible channels of influence exist which might allow for relatively precise manipulation of election results. Three general concepts are possible. The first is precise influence on the vote count by non-democratic means. Either in the form of ex-ante activities, such as the buying of votes, or in the form of ex-post manipulation, an example of which would be miss-reporting of vote counts. While the literature reports no evidence that vote fraud would be a regular or systematic issue for elections in western democracies, the possibility can not be ruled out completely[25]. Events like the recount of the Florida presidential elections votes in 2000 occasionally make the headlines and spark scepticism about the validity of election processes[26]. Especially in the case of very close elections, where only a relatively small amount of manipulation would be necessary to turn the results. As vote count manipulation is, by nature, a clandestine activity, small scale manipulation could potentially go undetected in the majority of cases. Occasional incidents of verified vote rigging show that democratic safeguards are not always effective. One such incident of electoral fraud in the U.S. was the Texas senatorial runoff election in 1948. The election was a very close one, with both campaign offices being aware of that fact. Caro (1990) reports that the campaign staff of challenger Lyndon Johnson influenced voters by directly paying out cash, appointing sympathetic election officials and bribing influential local

[25]See Alvarez and Hall (2006).
[26]See Lott Jr. (2001).

bosses who would send their employees and dependants to vote for Johnson. Additionally, allies of Johnson later confirmed acts of ex-post manipulation, where election officials would interfere with the counting and tabulation of votes, to ensure favourable results for their party.[27] In the end, Johnson won the election by a very small margin of only 87 votes out of about one million votes total.[28]

Yet, even though media attention devoted to elections, especially close ones, has increased during the last decades, the number of confirmed incidents of vote rigging has been small and ever declining.[29] If electoral fraud was common in U.S elections, the increase in media coverage and scrutiny should have led to an increase in contested elections. But the fraction of contested elections for the U.S. House has constantly and substantially decreased during the postwar period, as Jenkins (2004) reports. Even though some cases of illegal manipulation of the voting process surely have not been discovered, it is likely that the fraction of elections which were decided by fraudulent actions, is quite small. These findings suggest that vote fraud has rarely been a deciding factor for elections in western democracies, during the time period covered by the data. In all likelihood, vote fraud alone is not sufficient for explaining the dynamics at the threshold apparent in the Lee data.

In the case of Mexican mayoral elections, we do not observe precise sorting behaviour which might result from electoral fraud, even though elections in Mexico are sometimes subject to fraudulent behaviour, as reported by McCann (1998) and Lehoucq (2003). However, the stakes in mayoral elections are not as high as those for representatives on the national level, which might reduce incentives for fraud. Another possible explanation would be that on the local levels the consequences for fraud by the incumbent are so weak, that vote-rigging results in clear wins or losses, instead of close elections.

A second mechanism, which might play an important role in deciding close elections, is the use of 'emergency'-resources. Political parties will allocate more resources to close elections, than they would to those where they expect a clear win or loss. In these elections, the marginal effect of resources spent is greatest. Even activities which are extremely costly can be deemed worthwhile, if only a minimal shift of the vote count is necessary to win the election. These resources are not necessarily of a monetary nature, but can also take the form of organisational capacities or the ability for dealing with unforeseen events. Particularly, parties and candidates in extremely close elections will perform actions which are costly in terms of political influence or long term credibility, in order to win the race. Examples of such actions would be the trading of political favours, populist promises or policies which are not in line with the party platform. They might also make use of one-time resources, like calling in favours from influential groups or individuals. It is hardly possibly to measure a candidate's emergency-resources, which prevents researchers from analysing potential imbalances in this covariate. Candidates with superior financial and

[27] Election judge Luis Salas, who was involved in the tabulating of votes, later admitted the fraudulent manipulation of election results. See Caro (1990).

[28] See Caro (1990).

[29] The work of Campbell (2005) only reports a minimal number of elections where fraudulent activities were discovered over the last decades.

organisational resources are better informed about the current state of the race and can react more quickly and effectively to problems which their supporters might encounter. Those actions combined would exacerbate existing imbalances in terms of campaign funds for very close elections. They would also lead to a situation where the candidate with access to superior 'emergency'-resources has a distinct advantage. The actual magnitude of this specific kind of resource is likely unobservable. However, it is reasonable to assume that incumbent party candidates usually do possess an advantage in that regard.

For the Mexican mayoral elections, as noted earlier, the stakes are not as high and the resources available to the candidates are orders of magnitude below those for US House representatives. It is therefore likely that candidates are unable to monitor ongoing elections as quickly and effectively, and might lack the necessary information and resources to precisely influence elections in progress. This, in turn, would lead to a larger random component in the vote share outcome, which strengthens the RDD identification strategy.

The third, channel by which close elections can be non-randomly decided is the legal influence which incumbent parties have over the political administration. Among those measures are voter-suppression tactics like restrictive voter ID laws and targeted placement of polling stations.[30] For example by increasing the density and convenience of the polling infrastructure in areas with historically strong support for the incumbent party. Selective recounting of votes is another instrument which may allow for relatively precise manipulation of the vote score. Increased influence allows a party to lobby more effectively for or against vote recounts in districts which they expect to favour their, or the opposing candidate, respectively. Nevertheless, the number of times where recounts have reversed the results in a U.S. House race has been very small. According to Caughey and Sekhon (2011), vote recounts only had a pivotal effect on the election results in less than ten percent of the sample elections in which recounts did happen. While the result was reversed in favour of the incumbent party candidate in all reported cases, the low overall percentage of pivotal recounts rules out recounting as the main factor in explaining the observed imbalances. Incumbent parties do have other means, by which they could influence the vote share in close elections. Election officials in local offices usually do have a certain amount of discretion when dealing with unclear or provisional ballots, as is analysed by Kimball et. al. (2006). The party which has endeared itself to the administrative personnel during their last term in office will have gained an advantage as a result. Also, the partisanship of election officials can play a role in circumventing adverse conditions for the own party's supporters. One example of such practices, reported by Hauser and Holusha (2006), is that officials and judges can extend voting hours in districts which favour the party they are affiliated with. This kind of manipulation is more likely in very close elections, because the marginal effect is larger.

The channels for manipulation presented here probably do not represent all avenues by which candidates can precisely sort themselves around the threshold. While no singular main reason why sorting at the the vote percentage cut-off should be possible in U.S.

[30] For a discussion of the impact of voter ID laws on election outcomes in the US, see Weiser et al. (2005).

House elections is apparent, it is likely that the described effects, possibly combined with undiscovered factors, lead to the observed sorting behaviour.

5 Conclusions

In this article, the issues associated with non monotonic endogenous sorting in the context of the RDD are presented and a testing method for the validity of the design is discussed. The merits of thoroughly checking the data for evidence of precise manipulation of the assignment variable are motivated by a description of the various ways by which individual units can influence their realization of the assignment variable. Such manipulation can be detected by a specification test which is designed to find discontinuities at the threshold in the density function of the running variable. When non-monotonic sorting is happening at the threshold, the testing procedure described in the literature can be modified to suit the challenge. By testing sub-samples of the data which contain disproportionate numbers of individuals who manipulate their assignment score in a single direction, non-monotonic sorting can be detected. While it is possible to detect this problem by means of covariate distributions, the density based test expands the arsenal of researchers with a method far less demanding of the quality and spectrum of covariate data.

As example applications, the well established RDD analysis of the US-House incumbency advantage by D. Lee (2008), and the Mexican mayoral election RDD by Dell (2015), are examined with regards to a special form of non-monotonic selection effects. For the US House elections, testing on the aggregate level of all Democrat candidates does not reject the hypothesis of continuity of the assignment variable at the threshold, which is in line with the results of McCrary (2008) and the specification testing performed by Lee himself. However, the situation is not quite as clear when considering non-monotonic selection. Within a sample of all Democrat-candidates, the strictly positive self-selection of Democrats would be masked by the strictly positive self-selection of Republican candidates, considering the predominantly two-party system. The sub-sample of democratic-party candidates whose party was the incumbent at $t-1$ displays unexpected behaviour of the density function at the threshold. This is the sub-sample of individuals who are most likely able to precisely influence their assignment variable, according to Caughey & Sekohn (2011). Results for this sub-sample reject the hypothesis of a smoothly distributed assignment variable at the threshold for a substantial range of tuning parameters, with significance actually increasing for very close elections. Therefore it appears likely that certain candidates possess the ability to precisely sort themselves to one side of the threshold. In the case of the Dell (2015) application, no evidence for non-monotonic sorting can be found.

So far, no comprehensive explanation is available which would explain why precise manipulation of the vote share difference should be possible in US House elections, although a number of factors which might contribute to the sorting dynamics were discussed. Qualitative analysis of candidate behaviour in close elections could shed more light on this issue.

These results might lead to a reinterpretation of the incumbency advantage estimated by Lee (2008), if higher probabilities of winning close elections are common perk of being the incumbent. Since manipulation of the assignment variable is performed primarily by incumbent party candidates, it is worth considering to what extent the ability for manipulation in one election translates into a vote share advantage in the next election. From this perspective, it is worth considering if the discovered imbalances between close winners and losers introduce lead to a cumulative incumbent party advantage over multiple elections. In the latter case, sorting behaviour in subsequent elections might be an integral part of this advantage.

References

Alvarez, M. and Hall, T. E. (2006). Controlling democracy: The principal-agent problems in electoral administration. *The Policy Studies Journal*, 34:491–510.

Campbell, T. (2005). *Deliver the Vote: A History of Election Fraud, an American Political Tradition - 1742-2004*. Carrol & Graf Publishers, New York.

Caro, R. A. (1990). *The Years of Lyndon Johnson: Means of Ascent*. Alfred A. Knopf, Inc., New York.

Caughey, D. and Sekhon, Jasjeet, S. (2011). Election and the regression discontinuity design: Lessons from close U.S. house races, 1942-2008. *Political Analysis*, 19:385–408.

Cheng, M.-Y. (1997). A bandwidth selector for local linear density estimators. *The Annals of Statistics*, 25(3):1001–1013.

Cheng, M.-Y., Fan, J., and Marron, J. S. (1997). On automatic boundary corrections. *The Annals of Statistics*, 25(4):1691–1708.

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

Dell, M. (2015). Trafficking networks and the mexican drug war. *American Economic Review*, 105(6):1738–1779.

Dickert-Conlin, S. and Chandra, A. (1999). Taxes and the timing of births. *Journal of Political Economy*, 107.

Dustmann, C. and Schönberg, U. (2011). Expansions in maternity leave coverage and childrens long-term outcomes. *American Economic Journal: Applied Economics*, 4:190–224.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London.

Gans, J. S. and Leigh, A. (2008). Born on the first of july: An (un)natural experiment in birth timing. *Journal of Public Economics*, 93:246–263.

Grant, W. (2012). No early end to mexico's election stand-off. BBC News. http://www.bbc.com/news/world-latin-america-18700953.

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

Hauser, C. and Holusha, J. (2006). Problems lead 8 states to extend some voting hours. *The New York Times - International Herald Tribune*. `http://www.nytimes.com/2006/11/08/world/americas/08iht-web.1108day.3441243.html?_r=1&emc=eta1`.

Imbens, G. W. and Kalyanaraman, K. (forthcoming). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*.

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

Jenkins, J. A. (2004). Partisanship and contested election cases in the house of representatives, 1789-2002. *Studies in American Political Development*, 18:112–135.

Kimball, D. C., Kropf, M., and Battles, L. (2006). Helping america vote? Election administration, partisanship, and provisional voting in the 2004 election. *Election Law Journal*, 5:447–461.

Lalive, R. (2008). How do extended benefits affect unemployment duration? a regression discontinity approach. *Journal of Econometrics*, 142:785–806.

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

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

Lehoucq, F. (2003). Electoral fraud: Causes, types, and consequences. *Annual Review of Political Science*, 6:233–256.

Lott Jr., J. R. (2001). Documenting unusual declines in republican voting rates in florida's western panhandle counties in 2000. Unpublished, American Enterprise Institute, Crime Prevention Research Center.

Ludwig, J. and Miller, D. L. (2005). Does head start improve children's life chances? Evidence from a regression discontinuity design. *National Bureau of Economic Research Working Paper*, (11702).

McCann, J. A. (1998). Mexicans react to electoral fraud and political corruption: an assessment of public opinion and voting behavior. *Electoral Studies*, 17(4):483–503.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design, a density test. *Journal of Econometrics*, 142.

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Themes in modern econometrics. Cambridge University Press, Cambridge.

Papay, J. P., Willet, J. B., and Murnane, R. J. (2011). Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161:203–207.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66:605–610.

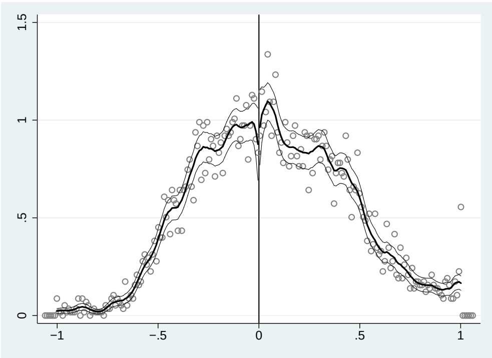
Tamm, M. (2013). The impact of a large parental leave benefit reform on the timing of birth around the day of implementation. *Oxford Bulletin of Economics and Statistics*, 7(4):585–601.

Weiser, W. R., Levitt, J., Weiss, C., and Overton, S. (2005). Response to the report of the 2005 commission on federal election reform. *Brennan Center for Justice*.

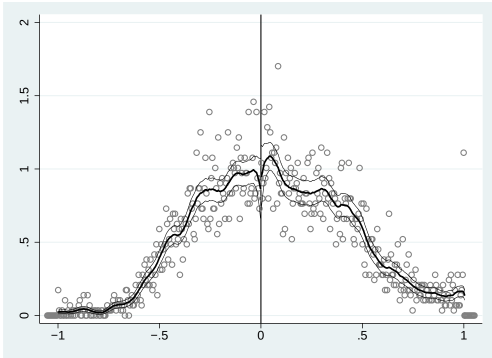
A Appendix

A.1 Additional specifications from Section 4.1

Figure 9: Additional density estimates for Democratic candidates
Horizontal axis: Democratic vote share difference in t. Vertical axis: Bin counts.
Confidence bands (95%) in thin lines.

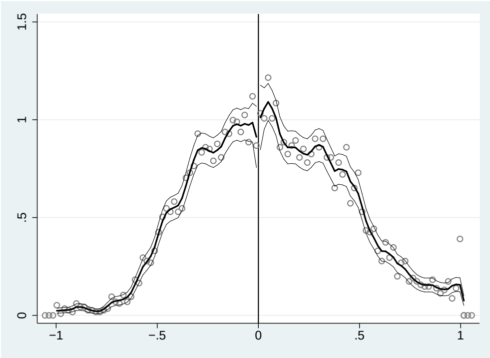


Note: Quarter the reference bandwidth, reference binsize.

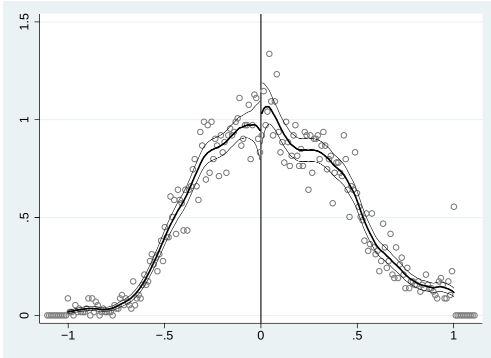


Note: Quarter the reference bandwidth, halved reference binsize.

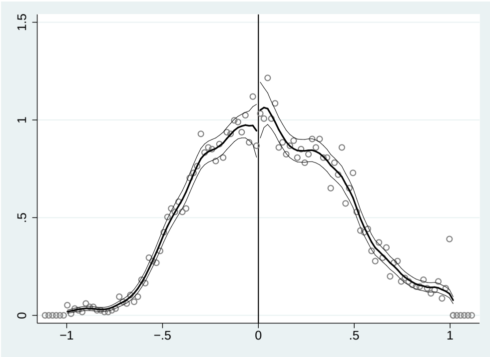
Horizontal axis: Democratic vote share difference in t. Vertical axis: Bin counts. Confidence bands (95%) in thin lines.



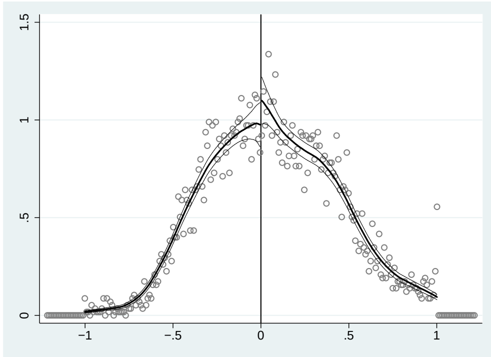
Note: Quarter the reference bandwidth, twice the reference binsize.



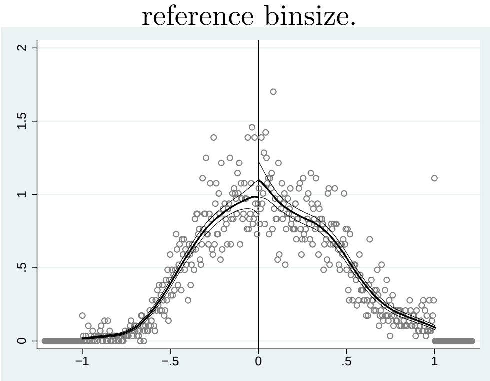
Note: Halved reference bandwidth, reference binsize.



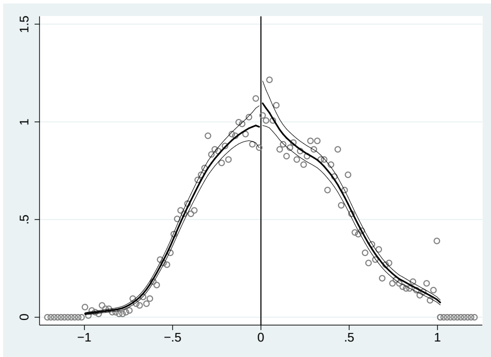
Note: Half the reference bandwidth, twice reference binsize.



Note: Reference bandwidth, reference binsize.

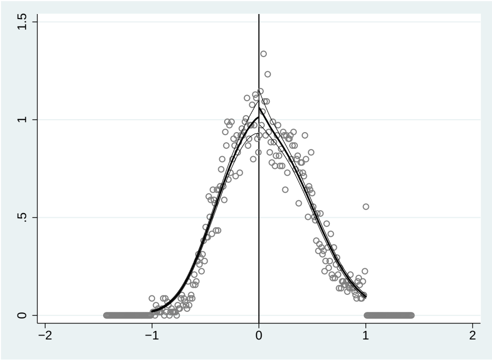


Note: Reference bandwidth, half the reference binsize.

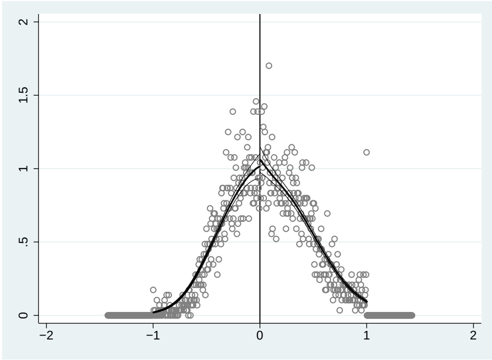


Note: Reference bandwidth, twice the reference binsize.

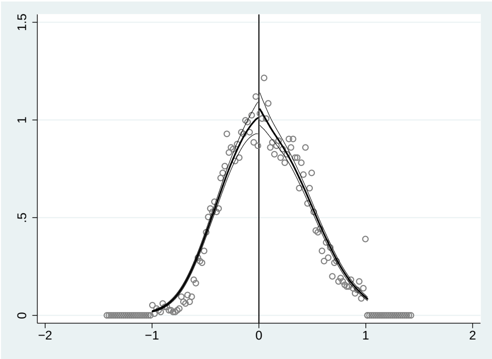
Horizontal axis: Democratic vote share difference in t. Vertical axis: Bin counts. Confidence bands (95%) in thin lines.



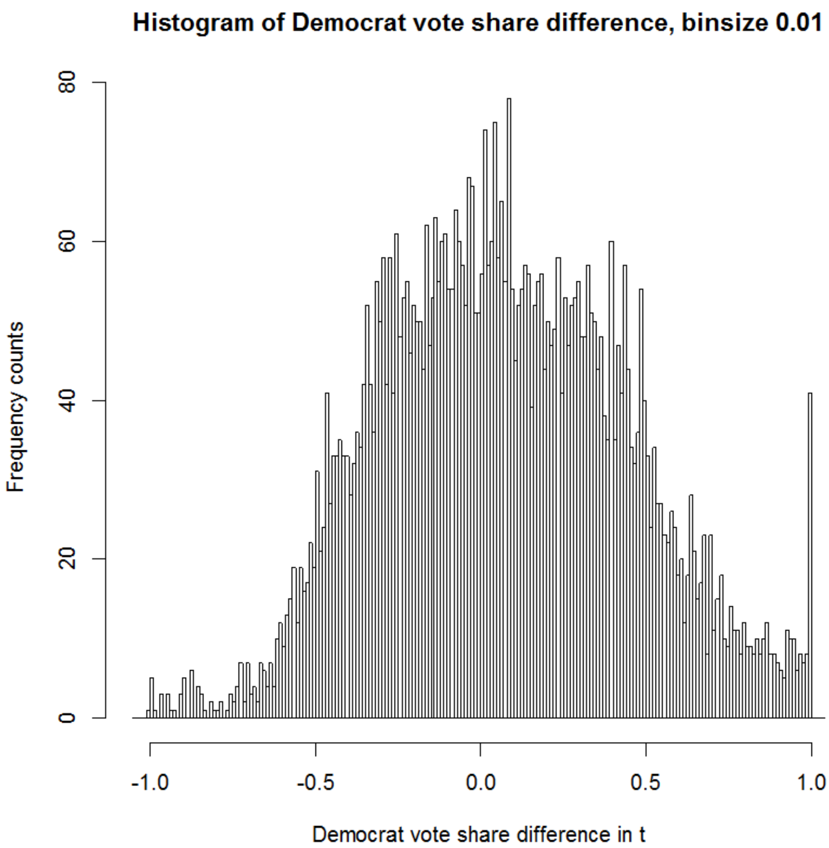
Note: Twice the reference bandwidth, reference binsize.



Note: Twice the reference bandwidth, half the reference binsize.

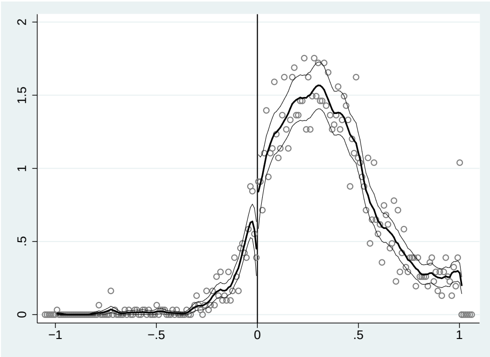


Note: Twice the reference bandwidth, twice the reference binsize.

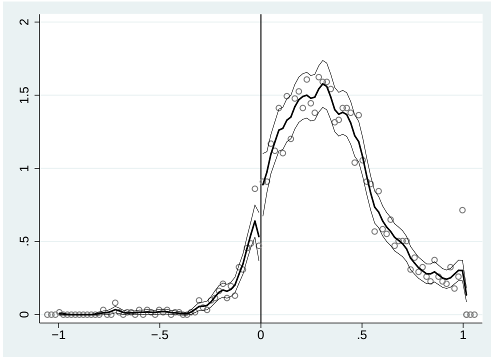


A.2 Additional specifications from Section 4.3

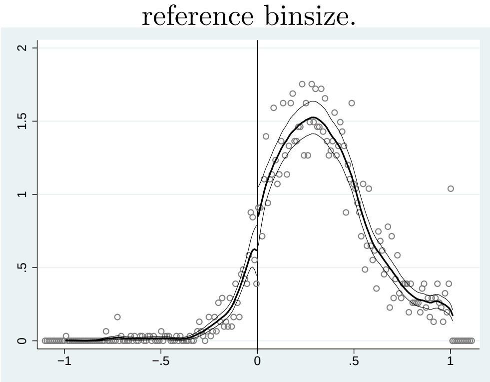
Figure 10: Additional density estimates for incumbent Democratic candidates
Horizontal: Democratic vote share difference in t. Vertical: Bin counts.
Confidence bands (95%) in thin lines.



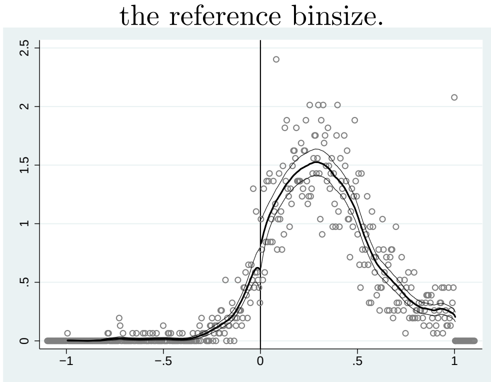
Note: One quarter reference bandwidth, reference binsize.



Note: One quarter reference bandwidth, twice the reference binsize.

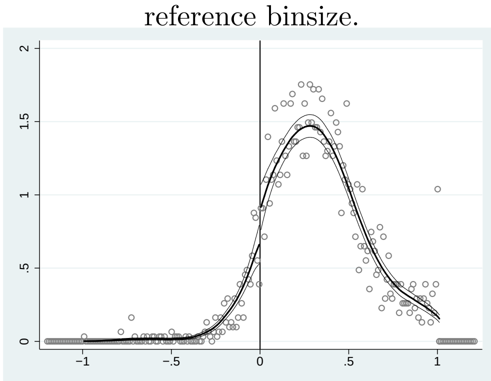


Note: Half the reference bandwidth, reference binsize.



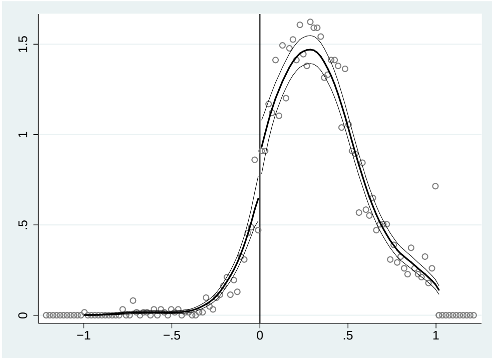
Note: Half the reference bandwidth, half the reference binsize.

Note: Half the reference bandwidth, half the reference binsize.

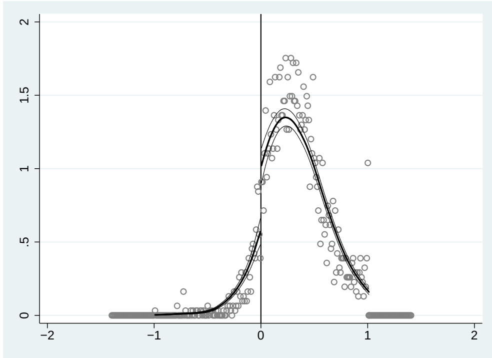


Note: Reference bandwidth, reference binsize.

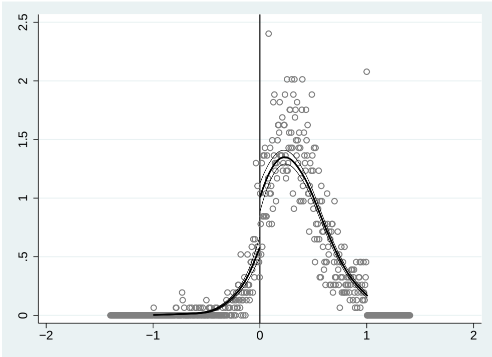
Horizontal: Democratic vote share difference in t. Vertical: Bin counts. Confidence bands (95%) in thin lines.



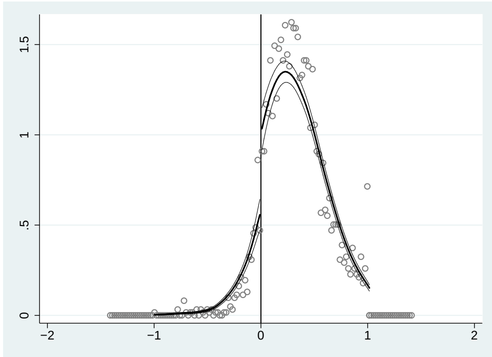
Note: Reference bandwidth, twice the reference binsize.



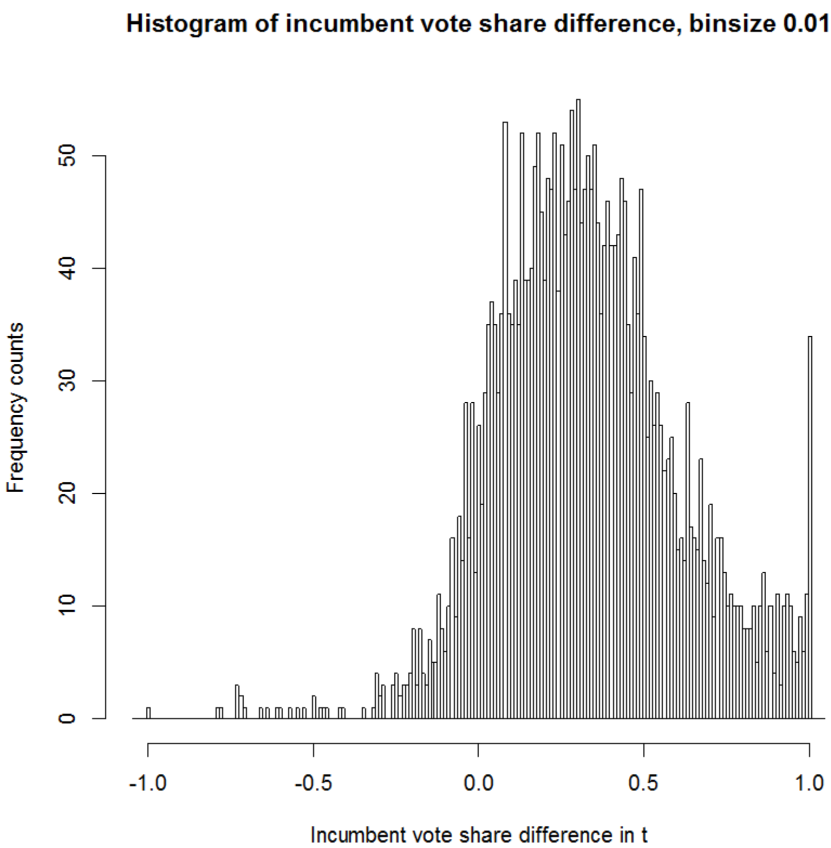
Note: Twice the reference bandwidth, reference binsize.



Note: Twice the reference bandwidth, half the reference binsize.



Note: Twice the reference bandwidth, twice reference binsize.



A.3 Description of the Local Linear density smoother

Construct J bins with $j = 1...J$ and bin width b. Let J_l and J_r denote the number of bins to the left and right of the cut-off c, respectively. The bins are defined as intervals:

$$(d_j,\, d_{j+1}] \quad with \;\; d_j = c - b(1 - j + J_l)$$

and bin midpoints X_j with $|X_j - d_j| = |X_j - d_{j+1}| = \frac{b}{2}$. Calculate the normalized observation counts per bin:

$$N_j = \frac{1}{Nb}\sum_{1}^{N} 1(d_j < x_i \leq d_{j+1}) \tag{2}$$

The histogram is then established by plotting the frequency counts $N_j$on the bin midpoints X_j.

The Local Linear estimator for a given bandwidth h and a kernel weighting function K, at $x_i = x_0$is described by:

$$\hat{y}(x_0) = \hat{\beta}_0(x_0) + \hat{\beta}_1(x_0)(x_0 - x_0)$$

with $\hat{\beta}_0(x_0)$ and $\hat{\beta}_1(x_0)$ minimizing the loss function:

$$\begin{aligned} & L\left(\hat{\beta}_0(x_0), \hat{\beta}_1(x_0)\right) \\ = & \sum_{j=1}^{J}\left(N_j - \hat{\beta}_0(x_0) - \hat{\beta}_1(x_0)(X_j - x_0)\right)^2 K\left(\frac{|X_j - x_0|}{h}\right) \\ \cdot & \;\{1(x_0 \geq c)1(X_j > c) + 1(x_0 < c)1(X_j < c)\} \end{aligned}$$

The expression in curly brackets ensures that no observations from one side of the threshold are used to calculate density estimates on the other side.

The triangular kernel is given by the expression:

$$K(x_0) = \begin{cases} 1 - |x_0| & if \; |x_0| \leq 1 \\ 0 & otherwise \end{cases}$$

The coefficients for the local linear regression are then calculated as:

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} S_0 & S_1 \\ S_1 & S_2 \end{pmatrix}^{-1} \begin{pmatrix} T_0 \\ T_1 \end{pmatrix}$$

With S_k and T_k defined as:

$$\begin{aligned} S_k &= \sum_{j=1}^{J} K\left(\frac{(X_j - x_0)}{h}\right)(X_j - x_0)^k \\ T_k &= \sum_{j=1}^{J} K\left(\frac{(X_j - x_0)}{h}\right)(X_j - x_0)^k N_j \end{aligned}$$

Consequently, the estimator at point x_0 is described by:

$$\hat{y}(x_0) = \hat{\beta}_0(x_0) = T_0 \frac{S_2 - S_1(X_j - x_0)}{S_0 S_2 - (S_1)^2}$$

The outcome of interest is then:

$$\gamma \equiv \ln \lim_{x_0 \downarrow c} y(x_0) - \ln \lim_{x_0 \uparrow c} y(x_0)$$

Define $\lim_{x_0 \downarrow c} y(x_0) = y^+$and $\lim_{x_0 \uparrow c} y(x_0) = y^-$. The estimate for a jump in the density is then:

$$\begin{aligned} \hat{\gamma} &= \ln \hat{y}^+ - \ln \hat{y}^- \\ &= \ln \left(T_0^+ \frac{S_2^+ - S_1^+(X_j - c)}{S_2^+ S_0^+ - (S_1^+)^2} \right) - \ln \left(T_0^- \frac{S_2^- - S_1^-(X_j - c)}{S_2^- S_0^- - (S_1^-)^2} \right) \end{aligned}$$

With $S_k^+ = S_k$ for $X_j > c$ and $S_k^- = S_k$ for $X_j < c$ as well as $T_k^+ = T_k$ for $X_j > c$ and $T_k^- = T_k$ for $X_j < c$.

It is shown by McCrary (2008) that the estimation bias $\sqrt{nh}\,(\hat{\gamma} - \gamma)$ is approximately normally distributed and asymptotically converges to zero under the following conditions: Everywhere except at c, the density function $y(x)$ has three continuous and bounded derivatives, $h \to 0$, $Nh \to \infty$ and $\frac{b}{h} \to 0$.[31] This leads to an approximate standard error for the estimator $\hat{\gamma}$ of:

$$\hat{\sigma}_\gamma = \sqrt{\frac{1}{hN} \frac{24}{5} \left(\frac{1}{\hat{y}^+} + \frac{1}{\hat{y}^-} \right)} \qquad (3)$$

A.4 Bandwidth selection Rule of Thumb

The histogram from the first step of the specification testing procedure is taken as a starting point. Then a separate ROT bandwidth is computed for both sides of the threshold. This is done by fitting a polynomial of the fourth order to the data on each side and calculating:

$$\begin{aligned} \bar{h}_l &= \kappa \left(\bar{\sigma}_l^2 \frac{|c - X_l|}{\sum \ddot{\hat{\lambda}}_l (X_j)^2} \right)^{\frac{1}{5}} \qquad (4) \\ \bar{h}_r &= \kappa \left(\bar{\sigma}_r^2 \frac{|c - X_r|}{\sum \ddot{\hat{\lambda}}_r (X_j)^2} \right)^{\frac{1}{5}} \end{aligned}$$

Where $X_l = X_0$, $X_r = X_J$. The index l describes the variables for the regression to the left of the cut-off, and index r describes those for the regression to the right. Let $\bar{\sigma}_l^2$ and $\bar{\sigma}_r^2$ be the mean squared error for the regressions on both sides of the cut-off. Also $\ddot{\hat{\lambda}}_l$ and $\ddot{\hat{\lambda}}_r$ are the estimated second derivatives of the fourth order polynomial model.

In order to calculate the standard errors for the local linear regression as per equation 3, the average of both ROT bandwidths is taken. This average is used for the local linear

[31] For a proof, see Appendix I of McCrary (2008).

estimator on both sides of the cut-off.