Can Close Elections Ever Be "As-Good-As" Random? A Modified Matching Technique for Imbalanced Regression Discontinuity Designs

Ben Gruenbaum and Chris Celaya

May 21, 2013

Abstract: *We adapt and apply recently developed matching methodologies to improve covariate balance across treatment groups for the purpose of estimating the causal effect of incumbency in the US House of Representatives. In doing so we assess a recent finding (Caughey and Sekhon 2011) that the Regression Discontinuity (RD) design is an invalid approach for estimating this quantity because of unobserved differences in the ability of parties or candidates to effect or manipulate the outcome of extremely close elections. We also discuss more generally how covariate matching and RD designs can be synthesized in future work which RD practitioners may find helpful.* [1]

[1] The original authors should be strongly commended for making their full data set and executable code and a helpful appendix available and easily accessible to the public. The data for the current replication project has been made available pursuant to the instructions of this replication project. We initially obtained all data used from Professor Sekhon's webpage: http://sekhon.berkeley.edu/rep/RDReplication.zip

I - Two Approaches to Covariate Imbalance: Regression Discontinuity and Matching

Covariate Balance in the RD Set-up

Covariate balance is essential for valid causal inference in most settings since imbalance on covariates across treated and control groups will bias estimates if the imbalanced covariate is correlated with treatment or the outcome under investigation. As a result, achieving balance across treated and control groups is a central aim of many of the most popular and promising causal analysis strategies. In this paper we examine how two approaches to covariate balancing -Regression Discontinuity (RD) and matching - can be synthesized to study a particular causal quantity that has long been of interest to political scientists: the incumbency advantage in the US House of Representatives. [2]. We find that applying a simple matching strategy can significantly decrease imbalance and reduce model dependence, without reducing statistical power below the threshold required for meaningful inference. This finding is a significant departure from the most recent literature on this subject (which we discuss extensively) that asserts that the RD design for this particular data context will only allow for estimates that rely on the very parametric assumptions that quasi-experimental designs are intended to avoid.

In an RD design, treatment is assigned as a known and deterministic function of an observed variable (which is referred to as the "forcing variable". Under certain assumptions this creates sharp and observable differences in treatment between units that we would otherwise expect to be identical or nearly identical on other relevant covariates(Imbens and Lemieux 2008). The archetypal example of this situation is the (Thistlethwaite and Campbell 1960) analysis of the effect of merit awards on future academic success, when these merit awards are awarded to students who score above a certain discrete cut-off on a test. The key and in this case plausible assumption that allows for the causal effect of the award to be estimated is that those students who barely made the cutoff for the award are probably similar on their covariates as those students that barely failed to receive the award. This is the key idea motivating RD designs - units with observed forcing variable values close enough to a known treatment-determining cutpoint (that is to say within a specified "bandwidth") will actually receive treatment (in this case, the merit award) *as if by* random assignment as long as these units are unable to precisely control their own values of the forcing variable in the region around the cut-point Under these conditions, then there is some subset of units who are are essentially similar on background covariates and -crucially in the context of causal inference - similar in terms of their potential outcomes under treatment and control Lemieux and Lee (2009). Note that this implies a key feature of the RD design is that it systematically *prunes* observations with "forcing variable" values outside the immediate proximity of the cutpoint.

As with any method of statistical inference, certain assumptions must be true in order for RD to provide valid estimates. The assumption that has received the most attention in the literature -and which we focus on presently - is that while treatment probability is discontinuous at the cutpoint (going from 0 to 1) the values of the relevant covariates should be continuous through the cutpoint. In the theoretical limit as the forcing variable approaches the cutpoint from either side, covariate values will be perfectly balanced at the cutpoint in an ideal RD setting. In practice we expect them to be as good as balanced around some non infinitesimal but small bandwidth around the cutpoint.

If, on the other hand, covariates are discontinuous in the vicinity of the cutpoint, then it isn't possible to determine whether any observed variation in outcomes are due to the treatment assign-

[2] Here we follow the literature we seek to engage with and define as the causal effect of the Democratic party winning in period t on the probability that the part wins in *t+1*

ment or due to the change in the value of the covariates, and there is no way to define a region of observations containing units that are essentially similar and thus valid candidates for inclusion in the causal estimation process. Following the literature we refer to the situation where covariates and outcomes both appear discontinuous around the cutpoint as "sorting". While such sorting generally invalidates the use of the sharp RD design, in many contexts there is are strong reasons to suspect that this dynamic is occurring. [3]

For a more thorough discussion of RD considerations, see Lee (2008), Imbens and Lemieux (2008),Lemieux and Lee (2009), Angrist and Pischke (2008), and Imbens and Kalyanaraman (2012). [4]

Similarities and Differences in How Matching and RD Approach Covariate Balance

The importance of balance across treated and control units is of course more general than just the application of the RD design. Indeed this balance is central to any (or at least most) causal inference for the same reasons that it is important in RD settings -without identifying units of observation that are essentially similar on all variables that relate to treatment probability and potential outcomes, one is left making parametric assumptions about the data generating process which is rarely known or even knowable. As scholars have recognized this general problem of covariate imbalance in observational studies and the importance of addressing it for making causal inference, substantial energy has been devoted to achieving or improving balance through a process called matching. Matching is essentially a pruning process where certain observations are selected on the basis of the explanatory variables (either the covariates or the treatment variable) for inclusion or exclusion in the estimation of the quantity of interest. Generally, scholars seeking to apply matching for the purpose of making valid causal estimates with observational data will drop observations to the extent that these observations do not have comparable units that received the alternative treatment condition (Ho et al. 2007). [5]

While we leave the details of the theory and practice of matching to the interested reader, we note here one central property about matching processes that allows for all that follows: Matching - and the data pruning inherent to that process - does not introduce bias into causal estimates as long the matching and pruning is not carried out as a function of the dependent variable. This point is extremely general and powerful and largely makes the achievement of balance self-justifying - as long as one's matched data set is *not* produced via a function of the outcome variable, then any method that produces balance on the covariates that effect the probability of treatment and outcomes is a good one and will lead to estimates that are less biased and model dependent than estimates based on the original imbalanced data (Ho et al. 2007), (Iacus et al. 2011). (Ho et al. 2007), (Imai et al. 2008).

Clearly there are fundamental similarities between the RD approach and matching -specifically a shared focus on deriving causal estimates by comparing units with comparable potential outcomes.

[3] A classic example of sorting is bills in legislative bodies. Those bills that receive a bare majority are very different then those bills that barely fail to receive a majority, and this is normally due to the majority parties ability to "sort" which bills receive votes.

[4] We forego an extended discussion of the theory and general practice of RD and instead recommend to the interested reader the above articles which are considered to provide the current best practices and theoretical underpinnings of the design.

[5] For example, in assessing the causal effect of a job training program, dropping units who did not receive job training and have earnings in the top 1 percent of all wage earners would be justified if no one receiving job training had earnings in the same range (as we would expect).

Indeed the econometrics literature finds that the former is essentially a particular version of the latter, specifically “a limit form of matching at one point”(Lemieux and Lee (2009) quoting from Heckman, LaLonde, and Smith (1999)).

Of course there are important differences as well. RD assumes that covariates an potential outcomes will essentially be balanced as a function of the convergence of a single forcing variable. Matching in general however can be applied over any set of covariates - *except* for those that are forcing variables. As a result when treatment is determined in an RD type environment (according to some sharp criterion) it is in fact not *possible* to balance on the forcing variable, however RD estimation is designed to take advantage of the information encoded in this variable by looking at behavior as similarlities on this variable approach the limit for units in the two distinct treatment groups. The flip side of this is that since probability of treatment as a function of the forcing variable is either 1 or 0 it isn’t possible to use matching techniques to match on this covariate. Thus given an RD setting, we are forced to assess not just the balance on the covariates that can be balanced on but how these covariates vary with the forcing variable in the bandwidth nearest to the cutpoint.

Both RD and matching are the subject of an extensive and rapidly developing literature in statistics, econometrics and political science. We will have more to say about features of both that are relevant to our particular questions in the coming sections but readers should consult the references for the foundational articles that define the theory and practice of these techniques.

Recent Regression Discontinuity studies of the Incumbency Advantage

The sections to come focus on recent findings (Caughey and Sekhon 2011) regarding the validity of the RD approach for estimating the effects of the incumbency advantage in elections to the US House of Representatives. Specifically, Caughey and Sekhon (2011) (which we refer to as CS-2011 for convenience) finds that a previous application of RD by Lee (2008) produces biased estimates of this quantity because within the data set covariate imbalance worsened as observations approached the cutpoint - as opposed to the key assumption of the RD design that covariates will converge over this region. As election margins get tighter, according to CS-2011, certain types of candidates appear to become more likely to win, thus suggesting that the treatment assignment is not random for these elections.

The authors go considerably further than merely concluding that standard RD assumptions are not valid for close US House races. They also assert that “[t]he outcomes of close elections are so predictable that it is impossible to obtain covariate balance between matched treated and control observations. Thus, strong and unverifiable assumptions about the functional form of the relationship between treatment, covariates, and outcome must be made. Although covariate adjustment may be plausible in other applications of RD, it is not in the case of U.S. House elections except under strong assumptions.”

This stronger claim is what we test in sections 2 and 3 by attempting to use matching to create a data set that is balanced on the important covariates and which retains enough observations to allow for valid inferences in reasonably defined bandwidths around the cutpoint. We find that fairly straight forward matching on only two covariates eliminates the vast majority of imbalance in the regions around the cutpoint that are considered problematic n the CS-2011 analysis. In sections 4 and 5 we use this matching technique to derive point and uncertainty estimates which we compare to analagous estimates derived from the raw data.

First though, let us end this introductory discussion by motivating the analysis to come by placing the use of RD for the study of elections in a slightly broader context. Towards that end we note that the robustness and validity of the claims regarding the non-randomness of close elections in CS-2011 are important for future research not just on the incumbency effect but on elections and representation more broadly. [6] If election outcomes are as good as random in some situations then representational outcomes are as good as random as well - and thus such elections offer leverage on many questions about the causal effect of party representation and electorate-representative relationships. Additionally, the extreme imbalance noted by CS-2011 remains largely a mystery in terms of its origins and mechanisms. By examining the outcome of our matching process we hope that scholars can gain some additional insight into what precisely is driving the observed imbalance noted in CS-2011 and confirmed by our own replication of their results, and whether this imbalance applies generally or is for some reason specific to the observations contained in the current data set.

II: Analytic Approach and Matching Strategy

As a starting point for our own analysis, we replicated Caughey and Sekhon's work and confirmed their initial findings that obviously relevant covariates are in fact imbalanced for any reasonably defined region around the cutpoint. Put simply, for the elections in the data, covariates are not continuous through the cutpoint as they are required to be in a valid RD design.

To develop a strategy aimed at addressing this imbalance we have to think about what assumptions are required to hold for valid and efficient analysis in the RD setting, and how matching might be applied to achieve those goals. In other words we ask, "in order for RD to be useful in analyzing the causal effect of current interest, what would the US House elections data look like and is it possible to use matching in a way that doesn't bias our results or leave us with too few observations to generate sufficient inferential power?"

The answer to the latter question is entirely an empirical one that we turn to in the next section, the answer to the former is fairly straightforward, though certainly a departure from traditional matching routines as we now explain.

To begin to develop our matching routine, we ask what goals matching should achieve in the context of an RD design. Two main considerations guide our approach:

First, we note that the requirements for obtaining unbiased causal estimates in RD settings do not require that treatment assignment is as-good-as random across the entire range of observations. RD essentially imposes no assumptions on covariate behavior outside of some finite, typically small bandwidth on either side of the cutpoint. Within this cutpoint covariate continuity through the cutpoint is required (and overall balance on the range of relevant covariates is ideal). [7] Thus we focus our matching routine only on defined subsets of observations within symmetric and close margins around the cutpoint. Specifically we will define two matched data sets from which we hope to obtain estimates that are less biased and model dependent than those obtained using the raw data. The same matching specifications were employed to create both data sets with the only difference being that in one case the matching routine was applied to observations where the margin of electoral victory was less than 1% and in the other case this routine was applied to observations where this margin was less than 0.5%. The latter specification matches up closest with the perspective of the original authors who determined that imbalance within this tight region is particularly problematic and worthy of attention. Thus we primarily will use this data set in evaluating our results and we refer to it as "matched (0.5%)". However the results obtained by matching on observations within a 1% margin are presented in the appendix and referred to periodically as "matched (1%)"

[6] See Table 1 of CS-2011 for a list of recent articles employing RD to study election related phenomena.

[7] Determining the precise size of this region is a complicated and important question that we briefly discuss later.

to demonstrate that our findings are not due to some idiosyncratic irregularity in the 0.5% margin.

The second consideration reflected in our matching strategy is that in an RD setting we do not need to and indeed are not able to achieve balance on the forcing variable and thus we should not evaluate our results according to it nor attempt to match on it. The RD design extracts causal estimates as units' forcing variable values approach the cut point. Taking the basic premise of RD as a given then implies that balance on the forcing variable is not a requirement that our matching process must meet as long as our matched data set meets the criteria for valid RD studies generally.

Finally, we inject some analytic discretion into our strategy by choosing to specifically define our matching strategy according to the variables that are most commonly thought to affect the observed covariate imbalance. We focus in particular on achieving balance on the covariates that are most central to the literature's current explanations for why the imbalance observed in CS-2011 might occur. These explanations all center on the capacity of incumbent candidates or parties to control the mechanisms that determine close elections with greater precision than challengers to these incumbent parties and candidates. Thus we make achieving balance on incumbent party status as our top priority, and so our matching routine will be designed to ensure we have precisely the same number of incumbents winning and losing in the given window of observations over which we are matching. Additionally, we would not expect that all incumbency strength effecting the outcome of close elections is created equally across districts and time - and it is reasonable to believe that this strength is proxied by the closeness of the previous election in terms of the difference in two party vote proportion. Thus after matching exactly on incumbency status, we then match on the margin of victory in the previous election using “nearest neighbor propensity score” matching techniques that have been employed in various contexts in the past (Ho et al. 2007).

This strategy of course leaves out a large number of variables that are known or suspected to effect both the probability of treatment and the outcomes in the future that we are interested in assessing (the probability of a Democrat winning in t and t+1.) This could have posed a problem to our analysis if balancing on only these variables left us with a data set still fundamentally imbalanced on others and by balancing on these others we ended up with a data set so small that meaningful inference is rendered impossible. However, as it turns out, balancing on these two (first-order) variables essentially eliminates almost all of the other imbalances noted in CS-2011 (and the ones that remain may remain as a result of being post-treatment in nature or because they are mismeasured, as we discuss later).

In the next section we demonstrate our balancing method and its results in more detail and then discuss the quality of estimates that can be gleaned from the data set it produces.

III Assessing the Validity of RD Design with Covariate Balance Metrics

While we define our matching routine in terms of these variables alone, we note that the success of the routine in terms of creating balance requires that balance is achieved on *all* of the covariates that effect treatment and potential outcomes. Thus we will evaluate the results of our matching using all of the covariates identified as imbalanced or potentially of interest by CS-2011. In fact, we exactly apply to our matched data sets all of the primary tests and analyses of covariate balance that CS-2011 conduct on the original data paper. For the sake of making comparisons straightforward we therefore follow the specifications of the original authors in terms of which tests to run and which variables to use in these tests. We also supplement and improve their choice of evaluation metrics with additional tests that we feel are more suited to the task at hand, and we also produce graphical summaries that are easier to understand and provide greater insight into the dynamics determining covariate balance in this setting.

The accepted practice for assessing whether a treatment discontinuity implies "as good as" random assignment is to examine whether the covariates that effect either the probability of receiving treatment or the outcomes under treatment or control are continuous on either side of the cutpoint. If they are then it is taken as a fair signal that for a small enough region about the cutpoint units are essentially similar in terms of their potential outcomes since they are essentially similar over the covariates that effect these outcomes (except for the treatment of course). Essentially then, assessing the validity of a regression discontinuity design is about assessing the balance of relevant covariates for units around the cutpoint.

CS-2011's primary assertion is that this require balance in the immediate vicinity of the cutpoint does not exist, that this imbalance biases the estimates obtained by Lee (2008), and makes the obtained inferences sensitive to the modeling assumptions that RD designs are intended to obviate. Their evidence for this assertion focuses primarily on the covariate imbalance they find for elections decided by less than 2.5 percent and less than 0.5 percent depending on the balance metric employed.

In evaluating whether matching can alleviate these concerns and allow for the causal validity of the RD design to be recovered, we take the validity of the tests employed by CS-2011 as a given and see whether our methods can do better on these same tests where the raw data falls short. For the most part these tests are in fact consistent with what the literature suggests is the best way to assess covariate balance. However, we do find several areas where the tests employed by CS-2011 are suboptimal and where possible we provide better alternatives.

Graphical Assessment 1: Histograms

The first method that CS-2011 use to present evidence of imbalance is with histograms demonstrating that far more incumbents just barely win elections than just barely losing them. The particular matching strategy we used is by definition going to ameliorate this imbalance completely (or almost completely). For a defined range on either side of the cutpoint we are constricting our matched sample to contain precisely the same number of incumbents. But for completeness we still present these histograms and discuss them briefly.

As expected the matched data exhibit virtually no discontinuity at the cutpoint. In addition to presenting a graph with the identical specifications as used by the original authors, we also follow the current best practices in assessing RD design validity by subjecting these histograms to robustness checks by varying the bin size. The size of the bins is essentially arbitrary (though one can chose bins so as to minimize their bias as estimators of the observed probability distribution function, or to reduce the bias of local linear regression as we discuss below). If continuity or discontinuity is based on the position or size of the bins then this (dis)continuity is dependent on modeling assumptions that will normally lack reasonable justification. This is related to the issue of bandwidth selection that we discuss below. Presently, we show that different bin sizes and our different matching specifications all yield similar results in the area of the cutpoint - imbalance in the original data, balance for the matched data. For the sake of conserving space, we present two such examples here and two more in the appendix.

Figure 1: Incumbent Wins at Various Margins of Victory (Bin Size=0.5 %)

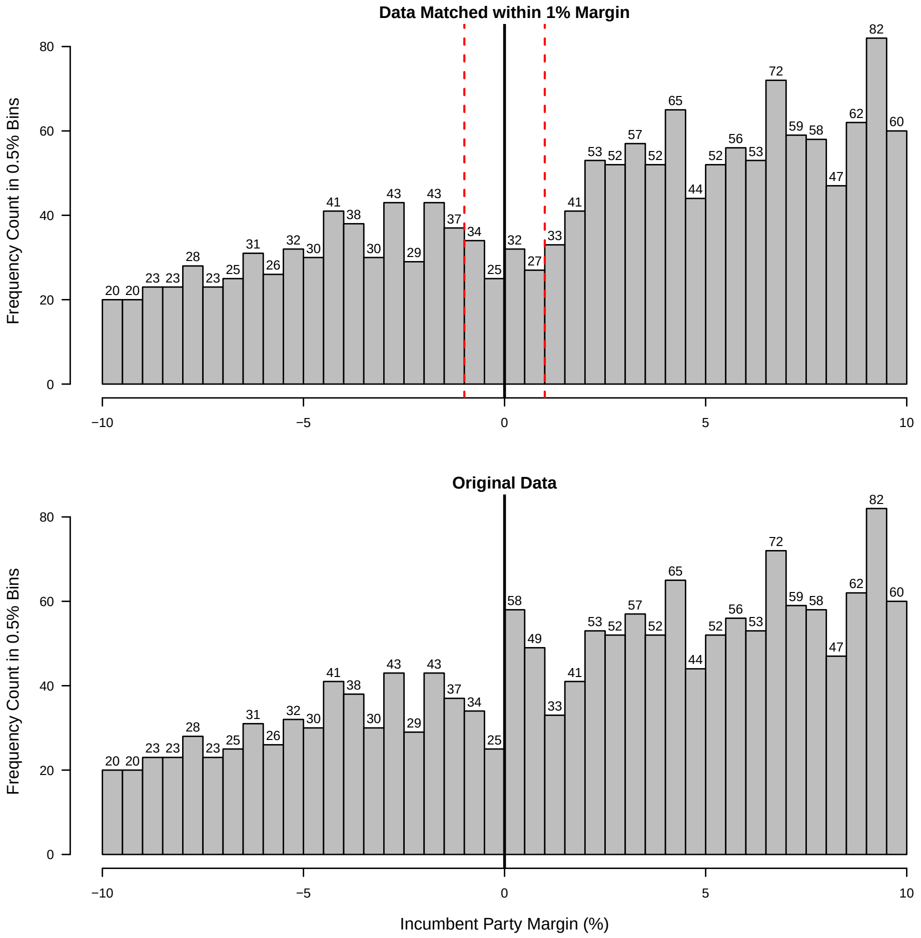
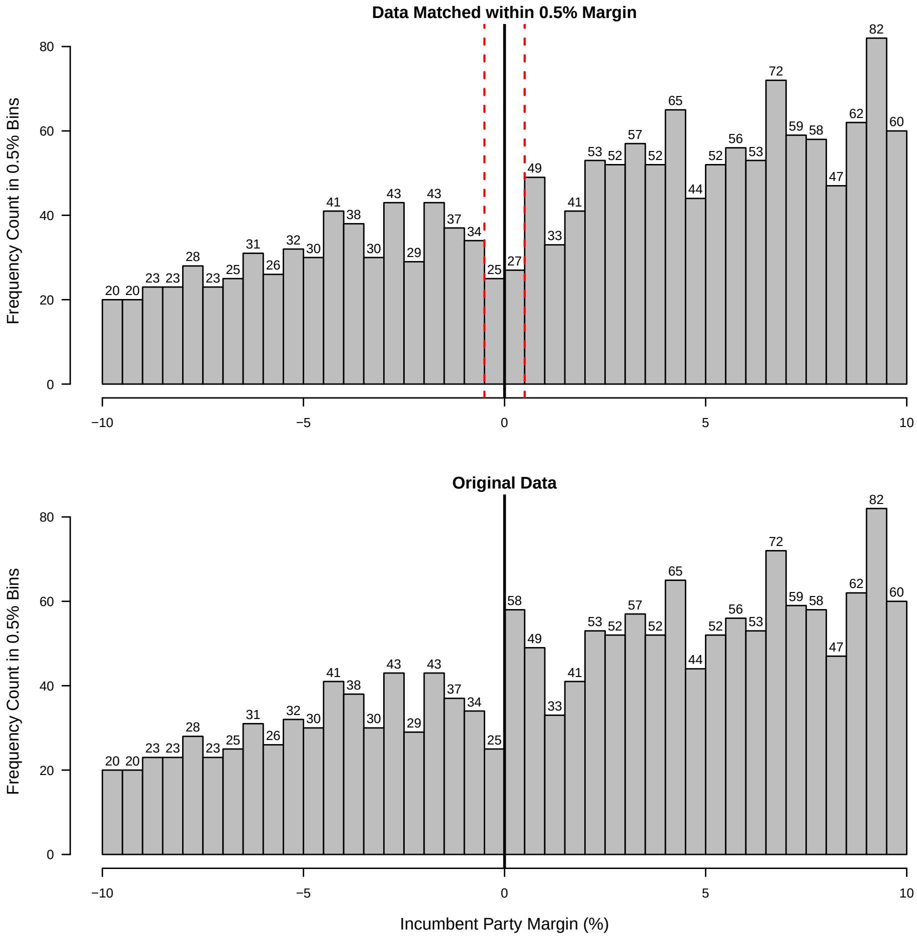


Figure 2: Incumbent Wins at Various Margins of Victory (Bin Size=0.5 %)



Additionally, we also present the tabular counterpart to these histograms in line with the presentation of the original authors. Table 1 reproduces Table 2 from CS-2011 and shows that for various electoral margins the Democratic losers of close elections are for any margin more likely to be drawn from districts where Democrats lost in t-1 and similarly for the winners in t and t-1. Table 2 presents the same information but now with our matched (0.5%) data. Not surprisingly, our matched data shows that Democratic winners and loser in time *t* are much more evenly distributed across Democratic winners and losers in *t-1*.

Tabular Assesments of Covariate Balance

Table 1: Cross Tabulation of Current and Lagged Democratic Victory (Original Data)

	Dem. Loss t-1	Dem.Win t-1	Dem % Held
Dem. Loss t(<2%)	107	65	0.38
Dem. Win t (<2%)	73	73	0.50
Dem. Loss t (<1%)	62	23	0.27
Dem. Win t (<1%)	36	44	0.55
Dem. Loss t (<0.5%)	34	8	0.19
Dem. Win t (<0.5%)	17	24	0.59

Table 2: Cross Tabulation of Current and Lagged Democratic Victory (Matched in 0.5% margin)

	Dem. Loss t-1	Dem.Win t-1	Dem % Held
Dem. Loss t(<2%)	91	65	0.42
Dem. Win t (<2%)	73	58	0.44
Dem. Loss t (<1%)	46	23	0.33
Dem. Win t (<1%)	36	29	0.45
Dem. Loss t (<0.5%)	18	8	0.31
Dem. Win t (<0.5%)	17	9	0.35

A final important note about these histograms is that the graphs of the matched data do not adjust for the fact that we have pruned observations from within a given bandwidth of the cutpoint and so there is a natural jump as one crosses the threshold into the electoral margins ignored by our matching process. Such thresholds are denoted by the red lines. Also by definition outside of these thresholds the original data and the matched data are identical and so the histograms are identical as well. While obviously creating an artificial appearance of a discontinuity at these points, we note that the relationship between the "bins" on either side of the cutpoint is not artificial or biased and since these bins are the ones that determine the estimates of our causal quantities in an RD design the spurious discontinuity at the margins of our matching region will not effect our primary estimates. In this particular case, the fact that there is not a major discontinuity on the edges of our matching margin for the matched (1%) data suggests that most of the imbalance that we are adjusting for in the matching process is within the region concentrated near the cutpoint. This might raise flags about about our ability to use the matched data sets for RD style inferences however we continue to retain a fair number of observations even after pruning within this particularly close electoral margin, and as we discuss more later, the appropriate region around the cutpoint in which to evaluate covariate balance is distinct from the region from which it is appropriate to derive point estimates.

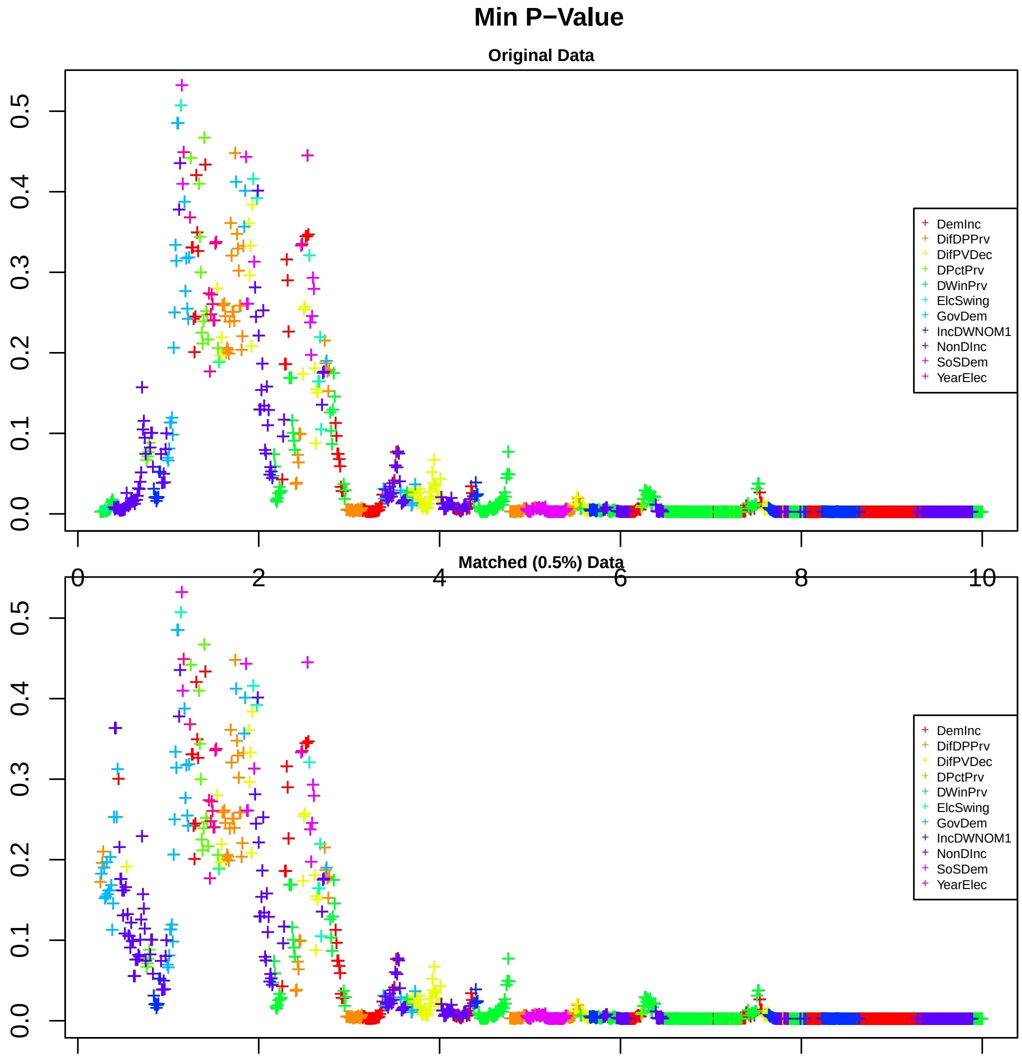
Graphical Assessment 2: Balance Metrics as Functions of Distance From Cutpoints

As alluded to earlier valid RD designs do not require that covariates be balanced over the entire range of observations. Instead in an ideal RD setting, covariates must become more balanced as they approach the cutpoint from either side. Such a trend indicates that units are more and more similar as values of the forcing variable get closer and closer to the cut point.

One primary way that CS-2011 assesses whether such a trend is evident is by presenting a graph of p-values measuring the statistical significance of covariate differences for treated and control units observed within intervals defined by electoral margins. The p-values they present are the minimum observed p-value for their selected group of covariates for all elections that were decided by that particular margin. For instance, for all elections where the margin of victory for the winning candidate was between 1.75 and 2.25 percentage points, the covariate that was most imbalanced between winners and losers of such elections was that corresponding to Republican incumbents - in other words, these incumbents were the most likely to be clustered into one treatment group (in this case the control group). The p-value for the difference in proportion of Republican incumbents receiving control compared to those receiving active treatment was about 0.021 and this is the value on the y-axis that is plotted for this particular margin. Generally then, the "higher" the value on the y-axis a point is, the more balanced we expect the data to be for this region of observations.

Under an ideal RD design, the covariate p-values should therefore increase montonicall as values of the forcing variable approach the cut-off point from either side. [8]

We do this first in the same manner as Caughey and Sekhon, using minimum p-values.



Note that the differences between these two graphs occur only in the regions where our matching routine was applied - that is elections decided by a 0.5% margin (we present similar graphs for data matched in a 1% margin in the appendix). Focusing only looking at the data in this matched range (between 0 and 0.5% on the x-axi)s, it is clear that balance improves substantially by this measure when using the matched data. This is the first graphical demonstration of the fact that while we balance on just two variables, balance improves considerably across all of the variables measured and reported by the original authors.

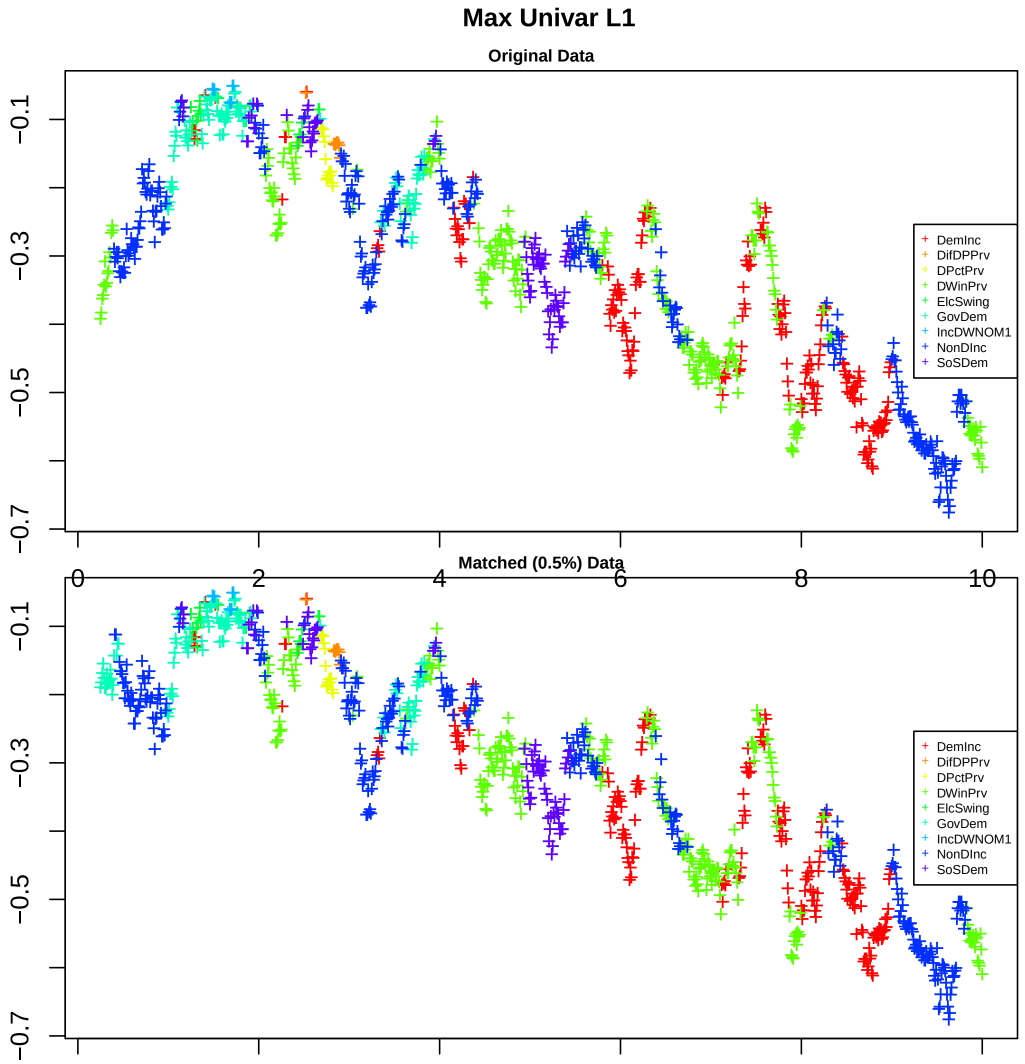
While the effect of our matching routine is discernible in this the above chart, this graphical approach is an imprecise way of presenting the imbalance along the range of observations on the forcing variable. For starters, several papers regarding covariate balance and RD design have noted that merely comparing difference in means is not a sufficient measure of balance or imbalance. While

[8] Note that not all the p-values in these figures are derived from exactly the same test statistic. Following the original authors we use exact Fisher tests for binary variables and Wilcoxon Rank Sum tests for variables with more than three values.

certainly an extreme difference in means is a powerful argument that imbalance exists, in taking the most imbalanced of such differences amongst 11 different covariates at a time, there is a strong possibility that the difference plotted at any particular point is actually the result of some spurious correlation rather than a real underlying population difference. Overall, the CS-2011 plot presents 976 points that are the most extreme observation out of 11 covariates at each defined margin of electoral victory. Thus we should consider what we'd expect to observe as the minimum p-values if we regressed 11 random variables on an unrelated random binary variable 976 times - certainly, many of the smallest p-values from the 976 regression would appear to be statistically indicative of a real imbalance between the values of the "covarates" and the values of the binary variable. We do not mean to suggest here that the relationship evinced in the original plot is substantially due to such spurious correlations but instead that presenting one reason why such a plot in the first place is not the optimal way to summarize discontinuity since at least some of the information it conveys is almost certainly not related to the underlying reality.

Additionally, by only using a difference in means this graph assumes "that any remaining imbalance in the matched sample is strictly unrelated to the treatment". (Ho et al. 2007) That is the distribution of the covariate not captured by its mean is assumed to be irrelevant. There is no realistic justification for this assumption in this context. Another issue is that p-values are generally functions not just of the observed values that go into the function but also of the sample size itself. As Imai (2008) demonstrates this means that when sample sizes decrease, t-tests become less significant and balance appears to improve even if the underlying true balance stays the same. While sample size is an important consideration in assessing the validity of any point estimates, ideally we would want to measure balance in a way that is independent of the sample size to separate the two issues. (Ho et al 2007 and Iacus et al. (2011)

We present one such way to achieve this goal below in the form of an alternative balance assessment defined by L1 distance scores. This points plotted here are now measures of L1 imbalance across the entire distribution of the covariate that is most imbalanced for the given margin on the x-axis. Thus to the extent that one can summarize overall imbalance with the imbalance measured on a single covariate this graph is superior to the minimum p=values since it conveys more information with the same amount of graphical complexity.



Interpreting, this figure, again the only variation between the two graphs occurs for values of the x-axis less than 0.5 (corresponding to observed elections decided by less than 0.5%) . In this range we see a near complete divergence in terms of the trend. The original data displays increasing imbalance whereas the matched data stays near the maximum amount of balanced achieve according to this metric. (Note that L1 scores are multiplied by negative 1 so that the trends evinced by the L1 graphs can be conveniently compared to those of the p-value plot.)

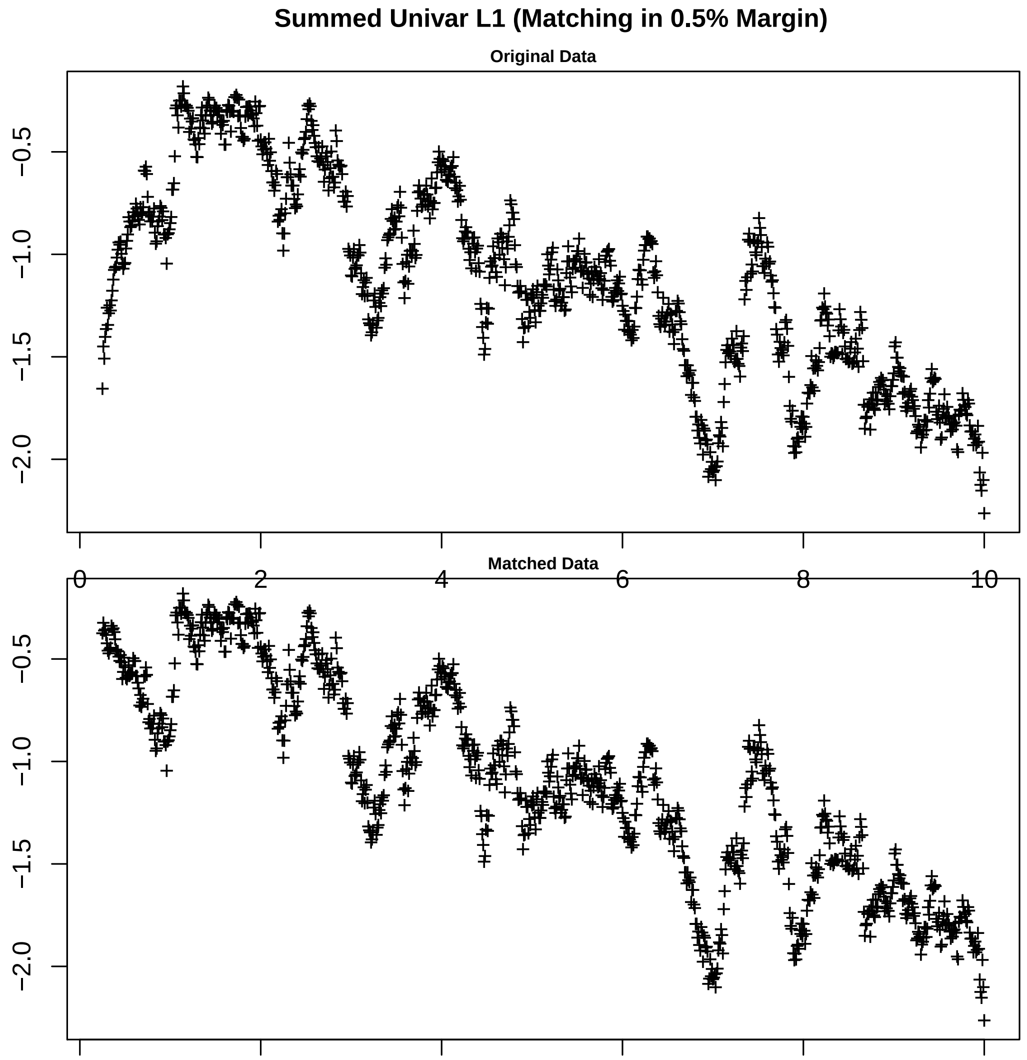
Still, this specification continues leaves something to be desired in that it only gives us information about one covariate at any given electoral margin. Ideally we would be able to summarize overall imbalance at each margin plotted along the x-axis for this graphical display. Thus, we present a third graphical specification that summarizes information about all the included covariates. In order to do so, we sum *every* each univariate L1 score to get a sense of *total* imbalance for all the covariates for *every* point along the margins. [9]

We also created a less easily deciphered plot of multivariate L1 scores analogous to those presented above. Multivariate L1 is valuable in that it is a comprehensive summary of the imbalance across all covariates and, crucially the interactions and squared terms of those covariates. As a result, the multivariate L1 score will overstate the level of imbalance when any of the covariates are

[9] Note that this is equivalent to the total distributional imbalance for all covariates if there are no interactions amongst the covariates

orthogonal to one of the other covariates. In the extreme, when all the covariates are orthogonal to all other ones, the multivariate L1 score overstates the imbalance by the difference between it and the average of the univariate L1 scores. In the current case, many of the covariates considered in calculating the multivariate L1 do not *a priori* seem theoretically related to the probability of treatment and thus speculating about the importance of their interaction terms with other terms is unlikely to lead to much more than that. For example, in our data set, there is little reason to believe that a legislative candidate's personal political experience should be interacted with a dummy for whether the current governor is a democrat, but a multivariate L1 measures matching on this potential interaction nonetheless. That being said, it still makes sense to compare the relative multivariate L1 scores within each data set to determine which specification leads to the best balance. We leave the presentation of this data to the appendix as we feel little sense can or should be made of this graphic for the above stated reasons.

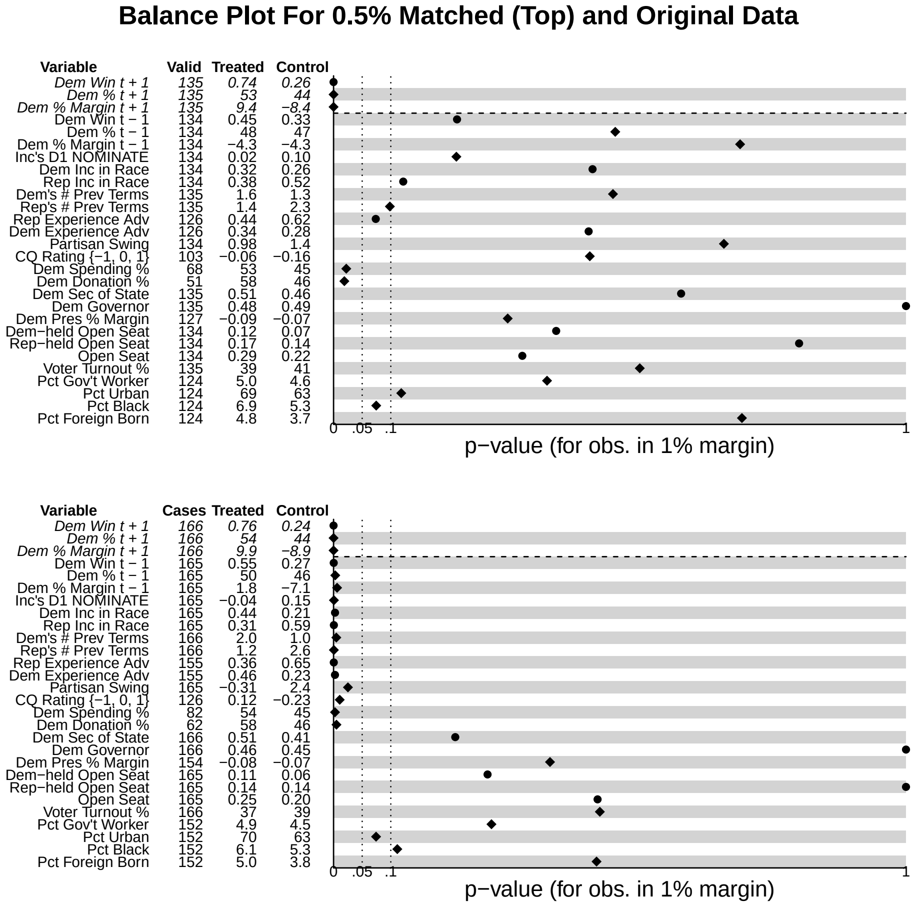
Given the extreme and seemingly unsystematic variation in the multivariate L1 score, we will assume that improved balance on other measures provided above represent improved balance for the purposes of using RD for causal estimates. We present the multivariate L1 score in the appendix.

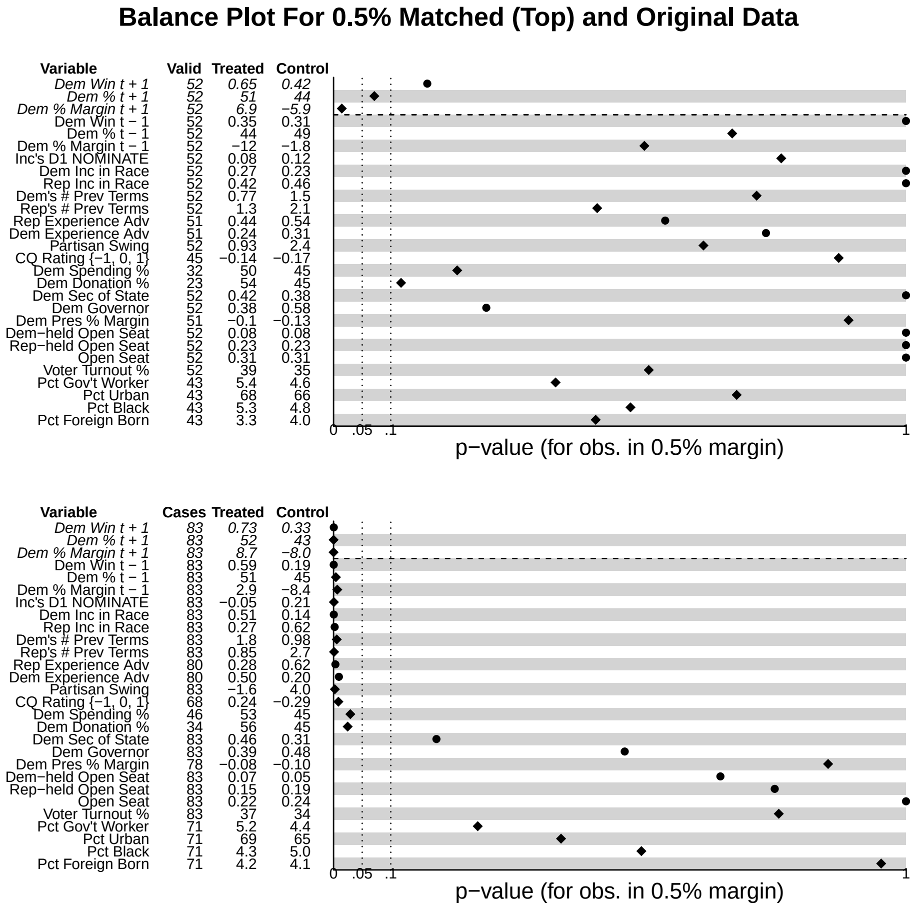


Here if we look at the margins between 0 and 0.5%, we see the original data showing increasing imbalance as observations approach the cutpoint on values of the forcing variable, whereas the matched data seems to move in the opposite direction. Clearly this remains consistent with our contention that matching significantly improves balance according to our specifications.

Graphical Assessment 3: Dot Plots

Another frequently employed method for examining covariate imbalance is to present dot or balance plots. These plot the imbalacne of each included covariate for a selected set of observations. Below we present pairs of such plots where in each pair, the upper plot is the balance obtained when matching whereas the lower plot is the balance obtained within the selected margin. The first pair shows balance for elections decided by less than a 0.5% margin and our matched (1%) data. The second pair presents the same plot using our matched (0.5%) data and looking at elections decided by 0.5%. Alternative specifications are presented in the appendix.





Here the evidence of improved balance is much more dramatic than in the previous section. In fact, in many ways, one can almost read a valid non-parametric causal estimate from these charts alone. The three first rows are various parameterizations of the outcome variable - and their proximity to the y-axis indicates the significnace of the difference in these outcomes for treated and control groups. The other rows represent the balance of the covariates considered by the original authors to most completely capture the differences between candidates who won and candidates who lost elections within the 0.5 percent margin. In the original data, most of the covariates are like the outcomes, close to the y-axis indicating significant differences between treated and control groups on these covariates, and thus real challenges to the validity of the unadjusted RD design. For the matched data however, virtually none of the covariates are close to the y-axis indicating that they are, post-matching, essentially randomly distributed over treated and control units within this range.

The exceptions to this characterization - those covariates that remain imbalanced after matching - require further study before definitive explanations can be given however there are a few observations to be made here. First, missing data is a bigger concern for these variables than any of the other observations. No analysis has been performed (either by us or the original authors) to try and determine the character of this "missingness" (at-random, completely-at random etc.) of this data however, it certainly bears noticing that the only variables still showing imbalance after matching are the only variables where there is significant missing information for observations where we have valid values for the treatment and outcome variables.

Second, while the other variables are clearly pre-treatment in nature, we are unable to definitively determine whether the campaign finance variables fall into this category. Given the complexity and intricacy of campaign disclosure rules and regulations, there is ample possibility for post-treatment effects to creep in. Most basically, the authors' analysis does not explain precisely what is being measured in these variables. If annual sums are being counted than candidates who win close elections may be getting "credit" for donations received *after* winning such elections (though given the quality of the data in the original paper we would doubt such an error occurred).

Harder to determine but more likely by our guess, candidates losing close elections have less of an incentive to accurately report all of the donations received - they have much less to pay in terms of political cost from the uncovering of any infraction. Additionally candidates who win close elections have incentives to *over* count their donations since this projects strength as one goes to serve in Congress. Since campaigns always file final financial statements after the election - this could partially explain the persistent imbalance on this metric despite the fact that the others seems to improve so dramatically.

Balance Implications

Overall, what do we learn from these plots? First, in terms of the graphical display modifications we make to the presentation in CS-2011, it appears that despite the theoretical concerns regarding the plotting of the p-values, the general trend they evinced in the original paper holds for more robust methods. That is, using L1 scores, we also observe that imbalance does not go away for observations near the cutpoint in the *original data*, as one would expect to observe if random assignment were occurring at or approximately at the cutpoint. However, the use of L1 scores also demonstrates that in both the matched and original data there is significant variation in balance over small windows of electoral margins both near and far away from the cutpoint. This is not not visible from the original minimum p-value plot because for all observations at margins greater than approximately 5%, there are covariates that are always extremely imbalanced. In the region beyond 5% the balance observed is just what one would expect in a valid RD study - at least one or some covariates are extremely dissimilar across treated and control units at these ranges. The L1 scores however help to show that the extent of this imbalance appears to vary widely, and this variation occurs even over observations which are not near the cut-point by any reasonable standard.

Finally we close this important section by noting again what we consider to be its most important finding: That by matching on just *two* covariates we essentially eliminate the imbalance on *all* the other covariates that are the subject of the original CS-2011 analysis. This means that the number of observations that we have to discard is fairly small - while of course non-zero - and so we lose little in the way of statistical power and gain much in terms of our ability to be confident that the effects we are measuring are unbiased by sorting at the cutpoint.

IV - Estimating Causal Quantities of Interest

Quantities of Interest, Covariate Balance and Optimal Bandwidths in the RD Design

Given the discussion thus far, it is important to remind readers that assessing covariate balance is not an end in itself. There is an ultimate quantity of interest that we and researchers are interested in. In the present context this is the effect of a Democrat winning in t on Democrat's probability of winning in t+1.

One of the main advantages of matching according to the approach we have used is that it is intended and defined to be a "pre-processing" strategy (Ho et al. 2007). That is, the matching process produces a data set that can then be analyzed according to other existing methods - including traditional RD designs.

While there is no one set way to extract point and uncertainty estimates from an RD design, one common, particularly straightforward method that is frequently employed is to use local linear regressions in the bandwidths nearest to the cutpoint on both sides. This is the method we discuss below. A key feature of this method is its sensitivity to the chosen bandwidth and thus a lot of attention in the RD literature is given to the choice of bandwidth and how to minimize bias resulting from this choice. At the same time, econometricians are in agreement that regardless of the optimal bandwidth, point estimates should be robust to a range of reasonable choices (Imbens and Kalyanaraman 2012).

This brings up a critical point that is not treated in the CS-2011 work. Specifically, it is important to separate bandwidth considerations for purposes of making unbiased estimates from bandwidth considerations for purposes of assessing covariate balance. As CS-2011 points out, the optimal bandwidth from the point of view of making unbiased point estimates in the current case - at least according to current best practices - is "an order of magnitude" larger than the bandwidth over which the covariate imbalance shows up most problematically and thus if one only looks at balance within the optimal bandwidth then one will miss the imbalance that CS-2011 flags as problematic.

On the other hand, the original authors appear to make the reciprocal error in evaluating the sensitivity of point estimates in (Lee 2008) to the imbalance they identify. Specifically, while it is apparent why imbalance should be checked through the entirety of a bandwidth especially over observations that define the closest subset to the cutpoint, the authors do not offer an explanation as to why we would discard the most recent and best prescriptions in the literature for defining the optimal bandwidth for producing unbiased point estimates, *conditional* on having balance throughout that bandwidth, including at the closest ranges. This decision by the authors becomes more important in the context of the sensitivity analysis discussed in the net section.

Presently however, we turn our attention to the tables below where we present various point estimates and associated uncertainty statistics for a range of bandwidths, associated both with the bandwidths where CS-2011 finds significant imbalance and for the optimal bandwidth as defined by the Imbens Kalyanaraman optimal bandwidth algorithm. Consistent with what CS-2011 report, the optimal bandwidth for making point estimates is far larger than the bandwidth we have been considering in the context of assessing covariate imbalance.

In the most conservative specifications - where point estimates are computed using matched data and the smallest intervals - we find that the lower bound on the 95% confidence intervals are in fact less than zero. However, we don't interpret this as evidence that the true causal effect may be zero since using such narrow intervals makes little sense from the point of view of computing these quantities. As just discussed, looking at covariate balance in this small window makes sense but deriving point estimates from it at best suboptimal and inefficient in the sense that it doesn't use all available, valid and relevant information (which of course would contribute to wider confidence intervals then is appropriate), and at worst introduces unnecessary bias into the estimates, which here is a real concern given that the bandwidth is so far away from the bandwidth derived specifically to reduce the bias of estimates in this context.

As a result we think that the first set of tables below - which show point estimates derived at the Imebens-Kalyanaraman optimal bandwidth, double that bandwidth and half of that bandwidth, for both the original data and the data set produced by matching on observations within a 0.5% margin - is more informative and a more valid representation of the true causal effect we are considering. Both for these estimates and all the other point estimates shown here the matched data shows slightly smaller effects than the original data, but they are essentially similar in terms of robustness to specification of different bandwidths. This robustness is of course misleading if, as is the worry in the case of the original data, it is potentially the result of significant imbalance on covariates across treatment categories. However as we showed in the previous sections this imbalance is much less of a concern after applying the matching techniques outlined above. Thus, we feel that the table is strong evidence that the treatment effects estimated using RD after matching are robust to RD's primary modeling assumption (the size of the bandwidth) used in generating point estimates.

Table 3: Point Estimates and 95% Intervals for Matched (0.5%) Data and Original Data

	Point Estimates	
Bandwidth	Matched Data	Original Data
IK-Optimal Bandwidth	14.74	15.83
Half-BW	15.69	17.61
Double-BW	16.65	16.95

	Matched Data		Original Data	
	Lower Bound	Upper Bound	Lower Bound	Upper Bound
IK-Optimal Bandwidth	10.65	18.83	11.83	19.83
Half-BW	10.12	21.27	12.26	22.97
Double-BW	13.61	19.70	13.93	19.97

Table 4: Point and Uncertainty Estimates of RD Using Matched (0.5%) Data for Different Bandwidths

	Band Width=0.5%	Bandwidth=IK-Optimal%	Bandwidth=.25%
KI-BW Original Data	0.50	8.55	0.25
Matched	0.50	8.89	0.25
Lower Bound CI Original Data	8.24	11.83	0.28
Matched	-1.29	10.65	-7.18
Point Estimate Original Data	21.94	15.83	18.52
Matched	15.68	14.74	12.70
Treated Observations Original Data	84.00	1443.00	44.00
Matched	53.00	1472.00	30.00
Robust Standard Errors Original Data	6.99	2.04	9.31
Matched	8.66	2.08	10.14

Bandwidth Considerations and Defining the Estimand

Here we offer a slightly more detailed discussion of bandwidth considerations and the definition of the effect we are estimating prior to our final section that evaluates the sensitivity of our findings.

Recently, significant progress on the question of optimal bandwidth for RD analysis has been made thanks to Kalyanaraman and Imbens who derive an optimal bandwidth equation that is specifically designed for the RD setting [10]. In the traditional RD setting this has simplified the pro-

[10] Previous work on optimal bandwidth took as the main goal estimating unbiased local linear regression or probability distributions over an entire range of data which is obviously less relevant in the RD setting where estimates are made using data from only a small region

cess of choosing the appropriate bandwidth considerably. Simply plug in the appropriate observed features of your data set into your software program and the optimal bandwidth is returned.

One could imagine this would make decisions about matching in the RD context easier as well. As with the selection of a bandwidth, the determination of any matching process also faces the classic bias-variance tradeoff. If one knew the optimal bandwidth for the *post* matched data set ahead of time, then one could match on the observations in that margin and leave untouched others. However, the particular features of RD and the inherent tendency of matching to alter the size of the data set means that one can't hold constant the optimal bandwidth of a data set for the purposes of matching in that bandwidth. The act of matching creates a new data set which in turn has a new optimal bandwidth and so on. Whether this process converges to a theoretical limit is one which we commend to future researchers.

However, this is not entirely pertinent to our immediate question because while the optimal bandwidth is important for making optimal estimates, any estimate should be robust to the different bin width specifications. Thus we can side step this theoretical concern at least partially, by testing our estimates for different bandwidths.

To examine the question of sensitivity to bandwidth specification, we derive average treatment effects and 95% confidence intervals for the original and matched (1%) data for bandwidths ranging from 0.25% up to 8.5 which is approximately the bandwidth indicated by the Imbens-Kalyanaraman optimal bandwidth algorithm, in 0.25% intervals. The results are presented below in both graphical and a tabular format. They indicate that both the matched data and the original data show similar and fairly consistent results for reasonably specified bandwidths. However, given the sorting at the cutpoint observed by CS-2011, the results found using the the original data are likely biased. The matched data on the other hand, while estimating a different quantity of interest, does so in a way that is far less susceptible to bias or modeling assumptions.

The graph shows that the consistency of the estimates across different bandwidths is fairly striking. Given that we have considerably less concern about the biasedness of those from the matched data, there does appear to be a a fairly well defined range for the causal quantity we are considering - and that quantity appears almost certainly to be greater than 0 - which is a conclusion that CS-2011 fails to reach. Overall, there appears to be fairly strong evidence that incumbent party candidates are likely to have an advantage around 15% relative to what that party would receive if they weren't the incumbent party - at least in districts where it is reasonably possible for either of the two parties to win.

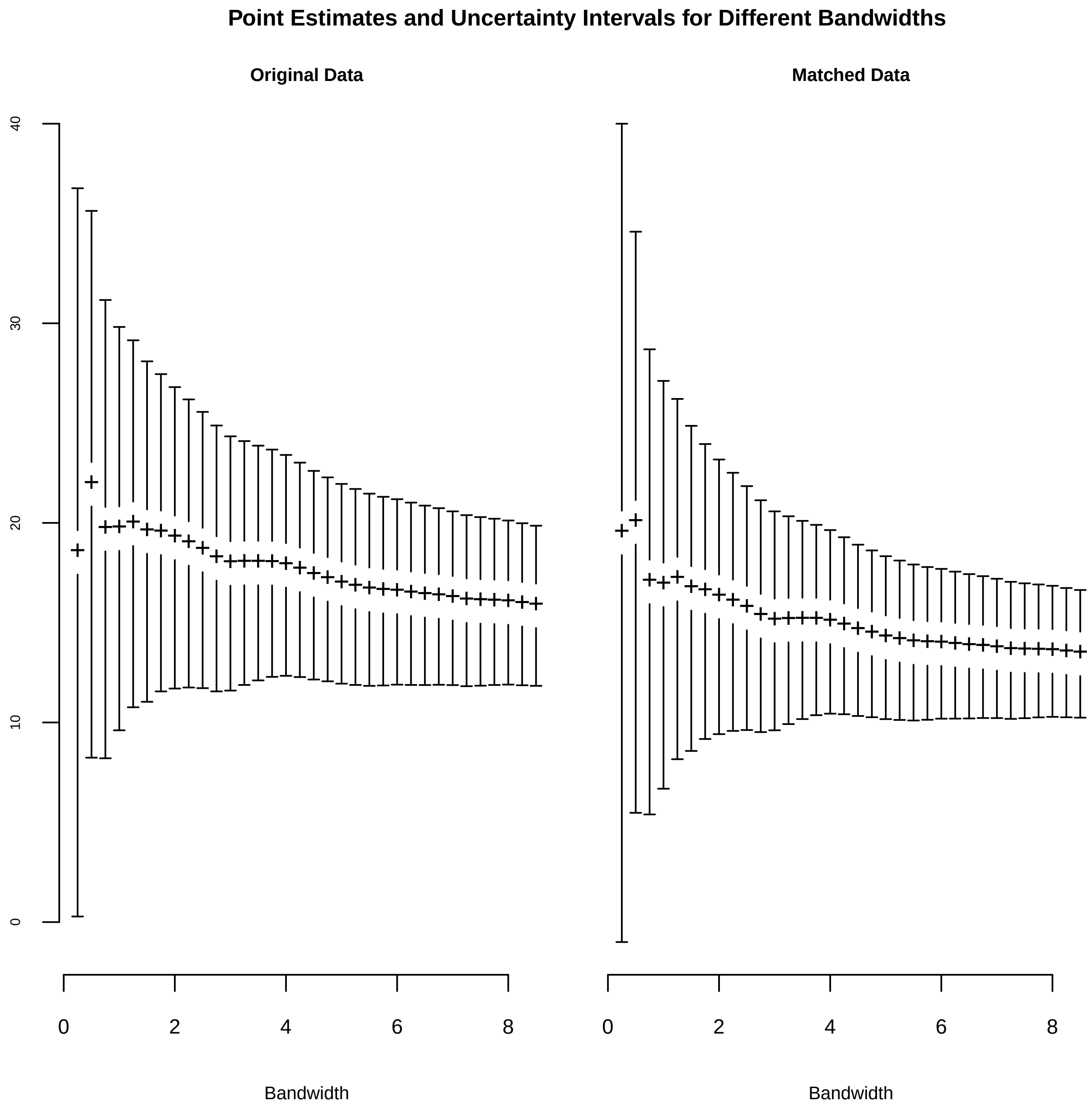

Point Estimates and Uncertainty Intervals for Different Bandwidths
Original Data
Matched Data
0
10
20
30
40
0
2
4
6
8
0
2
4
6
8
Bandwidth
Bandwidth

Table 5: Local ATE and Confidence Intervals for Matched and Original Data at different bandwidths

	BW	LATE.Full	Full.LB	Full.UB	Full.pvalue	LATE.Match	Match.LB	Match.UB	Match.pvalue
1	0.250	18.523	0.279	36.767	0.047	23.446	-3.532	50.424	0.088
2	0.500	21.935	8.238	35.633	0.002	24.145	4.987	43.303	0.014
3	0.750	19.687	8.205	31.168	0.001	20.215	4.881	35.548	0.010
4	1.000	19.713	9.610	29.817	0.000	20.022	6.578	33.466	0.004
5	1.250	19.955	10.762	29.148	0.000	20.401	8.523	32.279	0.001
6	1.500	19.565	11.037	28.094	0.000	19.785	9.068	30.502	0.000
7	1.750	19.505	11.556	27.454	0.000	19.580	9.857	29.303	0.000
8	2.000	19.252	11.701	26.802	0.000	19.229	10.176	28.283	0.000
9	2.250	18.969	11.752	26.186	0.000	18.899	10.389	27.408	0.000
10	2.500	18.641	11.721	25.561	0.000	18.488	10.446	26.529	0.000
11	2.750	18.218	11.558	24.879	0.000	17.955	10.312	25.597	0.000
12	3.000	17.968	11.599	24.337	0.000	17.646	10.429	24.864	0.000
13	3.250	17.989	11.881	24.098	0.000	17.689	10.838	24.540	0.000
14	3.500	17.989	12.108	23.870	0.000	17.704	11.170	24.237	0.000
15	3.750	17.980	12.286	23.674	0.000	17.699	11.424	23.974	0.000
16	4.000	17.871	12.338	23.404	0.000	17.577	11.524	23.630	0.000
17	4.250	17.647	12.276	23.017	0.000	17.322	11.487	23.158	0.000
18	4.500	17.379	12.153	22.605	0.000	17.021	11.373	22.669	0.000
19	4.750	17.172	12.064	22.280	0.000	16.790	11.291	22.289	0.000
20	5.000	16.949	11.945	21.952	0.000	16.536	11.166	21.906	0.000
21	5.250	16.791	11.882	21.699	0.000	16.366	11.111	21.621	0.000
22	5.500	16.650	11.836	21.465	0.000	16.218	11.076	21.359	0.000
23	5.750	16.581	11.855	21.307	0.000	16.158	11.123	21.193	0.000
24	6.000	16.541	11.895	21.187	0.000	16.133	11.195	21.072	0.000
25	6.250	16.449	11.881	21.017	0.000	16.044	11.198	20.889	0.000
26	6.500	16.371	11.878	20.865	0.000	15.968	11.211	20.726	0.000
27	6.750	16.314	11.889	20.738	0.000	15.916	11.239	20.592	0.000
28	7.000	16.226	11.874	20.578	0.000	15.828	11.236	20.421	0.000
29	7.250	16.103	11.818	20.388	0.000	15.700	11.183	20.217	0.000
30	7.500	16.067	11.844	20.290	0.000	15.674	11.227	20.120	0.000
31	7.750	16.044	11.878	20.209	0.000	15.662	11.280	20.043	0.000
32	8.000	16.009	11.897	20.120	0.000	15.635	11.315	19.956	0.000
33	8.250	15.923	11.864	19.983	0.000	15.549	11.287	19.811	0.000
34	8.500	15.847	11.836	19.857	0.000	15.472	11.264	19.679	0.000

Returning to the issue of what precise quantity is actually being estimated by RD or RD with matching within the bandwidth. This is a bit of a technically complicated issue. There is some disagreement in the literature about the appropriate characterization of the external validity of the point estimates obtained above. On the one hand Imbens and Lemieux (2008) characterizes the estimand as at best "the average effect for a subpopulation, namely the subpopulation where $X_i = c$" where c is the cutpoint and X_i represents the forcing variable. On the other hand, Lemieux and Lee (2009) suggests that the effect can be characterized as "a particular kind of average treatment effect across all individuals, a weighted average treatment effect where the weights are directly proportional to the ex ante likelihood that an individual's realization of X will be close to the threshold."

However regardless of the interpretation of this particular debate - in any RD design, the majority of treated and control observations will be dropped and this leaves researcher with an altered quantity of interest. Matching on the subset of units that is left after dropping the units outside the immediate vicinity of the cutpoint therefore does not seem to fundamentally change the quantity being estimated - that is it will still be a local average treatment effect - just one that is unbiased by sorting. This is also consistent with the philosophical framework of the matching strategy we use - that it is appropriate to "pre-process" and then proceed with an analysis strategy as one would have without matching. This is an area where the authors would benefit from the input of of better informed readers.

V - Sensitivity Analysis

A primary purpose of the original paper is to demonstrate that there are potentially problematic covariate discontinuities in very close elections that indicate sorting around the cut-point. The preceding charts and tables are the authors' primary means of making their point on this score. An equally important claim is that the discontinuity in the covariates is extreme enough such that researchers might have reason to doubt that any causally related discontinuity exists at all. In other words, not only are there significant discontinuities, but previous findings are in fact sensitive to these discontinuities. To conduct this sensitivity analysis, the authors undertake some fairly simple tests. The general framework for these tests is to measure to what extent treatment assignment departs from randomized treatment at the cut point. We do not go into great detail about these tests because we track exactly the steps taken in the CS-2011 appendix. For more detail readers can consult that appendix or the text that first defined these calculations as they are used here, (Rosenbaum 2002).

The specific test statistic employed is derived by considering two observationally equivalent units, one receiving treatment and one not, and then determining by how much the odds ratio of treatment would have to depart from random assignment in order to render a result statistically insignifcant at some given confidence level.

This value is denoted as Γ and is an upper bound on how great the odds ratio of receiving treatment can be amongst two observationally equal units, while statistical significance is still retained. It is defined as follows:

$$\frac{1}{\Gamma} < \frac{\pi_j(1-\pi_k)}{\pi_k(1-\pi_j)} \leq \Gamma$$

Where the π's represent the probability of treatment for two different units.

Unfortunately we are constrained in using the original authors' specifications to test our matched data set in at least one way. As we previously note, in the matched (0.5%) data set, the causal effect estimated is not significant at a 95 percent confidence interval. We have noted why this does not cause us to doubt that the true causal incumbency affect is greater than zero. Most importantly though and worthy of repeating is the issue of bandwidth. Looking at covariate balance within very close bandwidths is a statistically rigorous practice that the original authors should be commended for. However, estimating causal effects with this bandwidth makes little sense when every suggestion in the literature is that a much larger bandwidth is likely to be unbiased. (For example, the optimal bandwidth according to the standard Imbens-Kalyanaraman method for our matched(0.5%) data is a margin of about 8.5 percent.) If the optimal bandwidth from the point of view of obtaining unbiased point estimates is at least several times larger than the one used to test the sensitivity of results to imbalance, the sensitivity will tend to be overstated. Also as alluded to above optimal bandwidth estimates presented previous are derived from an algorithm specifically designed to eliminate bias (Imbens and Kalyanaraman 2012), and so using a bandwidth to calculate effects that is much different than this optimal bandwidth is likely to lead to avoidable bias being introduced. Thus we would suggest that sensitivity tests over such a narrow interval are not an ideal sensitivity test.

As a happy medium between the recommended bandwidth for computing causal estimates and the bandwidth investigated by the original authors we will examine the sensitivity of our results using the tests in the original article for a 4% margin around the cutpoint. Note that we use our matched (1%) data and so we are leaving much of the margin in which we are testing sensitivity unchanged from the original data.

Under this assumption we find that the matched (1%) data would have to be confounded at a level of Γ equal to 5.75 in order for the estimated effect of treatment to have an insignificant effect on the outcome - where we can think of Γ as "an upper bound on the coefficient Œ $\geq$ relating an unobserved covariate unit to i's log-odds" of being treated given an unobserved covariate that is almost perfectly correlated with the outcome. So in the current case this corresponds to the difference in the log-odds of candidates winning close elections based on unobserved variables.

The way that CS-2011 utilize this test statistic is to use an imbalanced covariate as a proxy for unobserved confounding, assume that the unobserved confounding is perfectly correlated with the outcome and see if this would produce levels of Γ approaching that of the test statistic. They use the Following their procedure, we find that the Democratic party was the incumbent party in 262 of the 572 observations in this range (0.458) but they made up 0.513 of the treated units (that is Democratic incumbents won more than half of the seats won by Democrats despite making up less than 46 percent of the sample.

The probability that a Democrat wins a Republican district was 132/(132+178)=0.42 while the probability that a Democrat won a Democrat district was 139/(139+123) =0.53. The difference in these conditional probabilities is .11, and the natural log of their odds ratio is

$\log\left(\frac{0.53X0.58}{0.47X0.42}\right) = 0.4429177$, which under the methods outlined in the CS-2011 appendix gives a maximum value of Γ equal to $e^{.443}$=1.526382.

This calculation of a possible Γ proxy, is by the original authors' standard, an unrealistic over-estimate which they then adjust downward according to the actual observed difference in treatment probability between incumbent parties. Clearly however for these observations the effects that we estimate are not sensitive to any realistic estimate of confounding based on the process outlined in the original paper.

This is not surprising given the balance that we have shown we achieved and the strength of the causal effect that remains even after this balancing process. Thus it is consistent with the overall thrust of the preceding sections - that achieving a level of covariate balance that allows for valid causal inferences is not beyond the reach of current matching methods, and that by doing so we can generate more robust causal estimates.

VI - Conclusion

Overall, we find that significant covariate imbalance is indeed present in close races for the US House of Representatives in the post-war period. However, we also find that, contrary to the most recent prominent work on this subject, simple matching techniques can pave the way for valid RD estimation. By applying these techniques ultimately we end up with causal estimates that are marginally lower than those we would estimate with the raw original data, but which are still robust to any reasonable specification check.

We also present additional graphical and analytic methods of assessing the covariate imbalance which we believe scholars will find easier to make sense of and which we also believe convey a more accurate and more complete summary of the covariate balance across the range of observations that researchers are interested in.

Possible future extensions of work in this area are numerous. First, one can obviously test the results obtained here and by previous authors with the wealth of data on elections to other legislative bodies. Indeed we believe that such work is likely forthcoming in the near future. Another area for future research relates to the theoretical issues associated with matching RD and optimal bandwidth. As far as we know, there has not been any significant effort to lay a theoretical framework for the joint application of matching techniques and RD. Specifically a way to identify the appropriate bandwidth within which one should test for covariate continuity and the relationship between this bandwidth and the bandwidth appropriate for the calculation of causal quantities would help when a regression discontinuity design might be fruitful but is potentially confounded by self-selection around the cutpoint. Without more theoretical scaffolding, there is likely to remain significant difficulty in parsing out whether observed measures of differences are the function of sample size changes as one approaches the cutpoint or legitimate changes in the behavior of subjects.

Generally, we think our findings should be taken as good news for scholars of elections and representation. With important adjustments and rigorous checking of those adjustments it appears that outcomes from close elections can in fact be taken to be as good as random and thus the window such elections provide into politically central outcomes remains open.

References

Joshua D Angrist and Jörn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press, 2008.

Devin Caughey and Jasjeet S Sekhon. Elections and the regression discontinuity design: Lessons from close us house races, 1942–2008. *Political Analysis*, 19(4):385–408, 2011.

James J Heckman, Robert J LaLonde, and Jeffrey A Smith. The economics and econometrics of active labor market programs. *Handbook of labor economics*, 3:1865–2097, 1999.

Daniel E Ho, Kosuke Imai, Gary King, and Elizabeth A Stuart. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, 15 (3):199–236, 2007.

Stefano M Iacus, Gary King, and Giuseppe Porro. Multivariate matching methods that are monotonic imbalance bounding. *Journal of the American Statistical Association*, 106(493):345–361, 2011.

Kosuke Imai, Gary King, and Elizabeth A Stuart. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the royal statistical society: series A (statistics in society)*, 171(2):481–502, 2008.

Guido Imbens and Karthik Kalyanaraman. Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, 79(3):933–959, 2012.

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

David S Lee. Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697, 2008.

Thomas Lemieux and David S Lee. *Regression discontinuity designs in economics*. National Bureau of Economic Research, 2009.

Paul R Rosenbaum. *Observational studies*. Springer, 2002.

Donald L Thistlethwaite and Donald T Campbell. Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309, 1960.