Zig Zag:
When Noisy Regression Discontinuities Yield Exaggerated Claims

Joseph T. Ornstein[*]

July 12, 2019

Abstract

In applied settings, regression discontinuity (RD) designs often suffer from noisy datasets and low power. When combined with a statistical significance filter, this tends to produce exaggerated causal effect estimates, typified by implausibly large slope and/or concavity parameters. In this paper, I illustrate the problem through a Monte Carlo simulation and two empirical illustrations from the published literature. I propose a method for estimating the magnitude of the distortion from the statistical significance filter and modified guidelines for the graphical display of RD results.

[*] Postdoctoral Research Associate, Washington University in St. Louis

1 Introduction

Regression discontinuity (RD) is an approach to causal inference that leverages a discontinuous change in treatment at a cutoff. The key identification assumption of the RD design is the continuity of other pre-treatment covariates; so long as treatment status is the only variable that changes discontinuously at the cutoff, the causal effect of treatment is identifiable at that point (Hahn, Todd and Van Der Klaauw, 2001). Because such thresholds, cutoffs, and boundaries are a common feature of political institutions, RD has proven a popular research design in political science over the past two decades (de la Cuesta and Imai, 2016).

Estimating the average treatment effect at the cutoff requires finding the limits of the outcome variable as it approaches from the left and right, and taking the difference between the two limits. An important practical consideration involves choosing a bandwidth – i.e. how much data to include in the estimation (Imbens and Kalyanaraman, 2012). This choice involves a bias-variance tradeoff. Too wide a bandwidth yields biased estimates; high-order global polynomial regression tends to overfit to observations far away from the cutoff (Gelman and Imbens, 2017). Too narrow a bandwidth reduces number of observations used to estimate the treatment effect, increasing the variance of the RD estimator.

Calonico, Cattaneo and Titiunik (2014), hereafter CCT, propose a nonparametric approach to RD estimation that selects the bandwidth (h) to minimize the mean squared error of the RD estimator. Their approach estimates the limits approaching the cutoff using local linear regression weighted by a triangular kernel and adjusting for a bias-correction term, and the authors derive robust standard errors for inference. Confidence intervals from this method produce the best empirical coverage of any method proposed to date, and its estimates perform well in a wide array of simulations.

Because RD estimates rely heavily on observations near the cutoff, datasets that are sparse or noisy in that neighborhood can yield low-powered hypothesis tests. The *power* of a frequentist statistical test is defined as the probability that it will reject the null hypothesis, conditional on a given value for the parameter being estimated. (Its complement is the Type II error rate.) Conventionally, a study is considered high-powered if power exceeds 0.8, though this convention is just as arbitrary as the convention of constraining the Type I error rate to 0.05 (Cohen, 1992).

In the experimental literature, there is a longer tradition of conducting power analyses prior to a study in order to determine the minimum necessary sample size to collect. Because experimenters can control the size of n, this emphasis is well-justified; done properly, a power analysis can ensure that effort is not wasted on experiments that are unlikely to detect the true effect. In observational statistical analysis, however, explicit power analyses are less frequently undertaken. In part, this is because researchers working with observational data are typically unable to control their sample size, and must work with the data available. When power is mentioned in observational research, it is often in the context of explaining away null results.

This is an unfortunate state of affairs. As scientists studying the replicability of studies in psychology and neuroscience have shown, low-powered experiments are pernicious not simply because they often fail to achieve statistical statistical significance, but because when they do, the estimated effects are several orders of magnitude too large (Button et al., 2013). Gelman and Carlin (2014) refer to these distortions as Type M and S errors. They show that, for low-powered experimental studies, a statistically significant result implies either an overestimate (Type M) or an estimate with the wrong sign (Type S).

To illustrate this problem in the RD case, consider a Monte Carlo simulation in which the variables X and Y are generated through the following process:

$$X \sim U(-1, 1)$$

$$Y_i = \mathbf{1}(X_i \geq 0)\tau + X_i\beta + \varepsilon_i$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

where $\mathbf{1}(\cdot)$ is the indicator function, and τ is the true causal effect of treatment. In other words, when τ and β equal zero, Y is purely Gaussian noise.

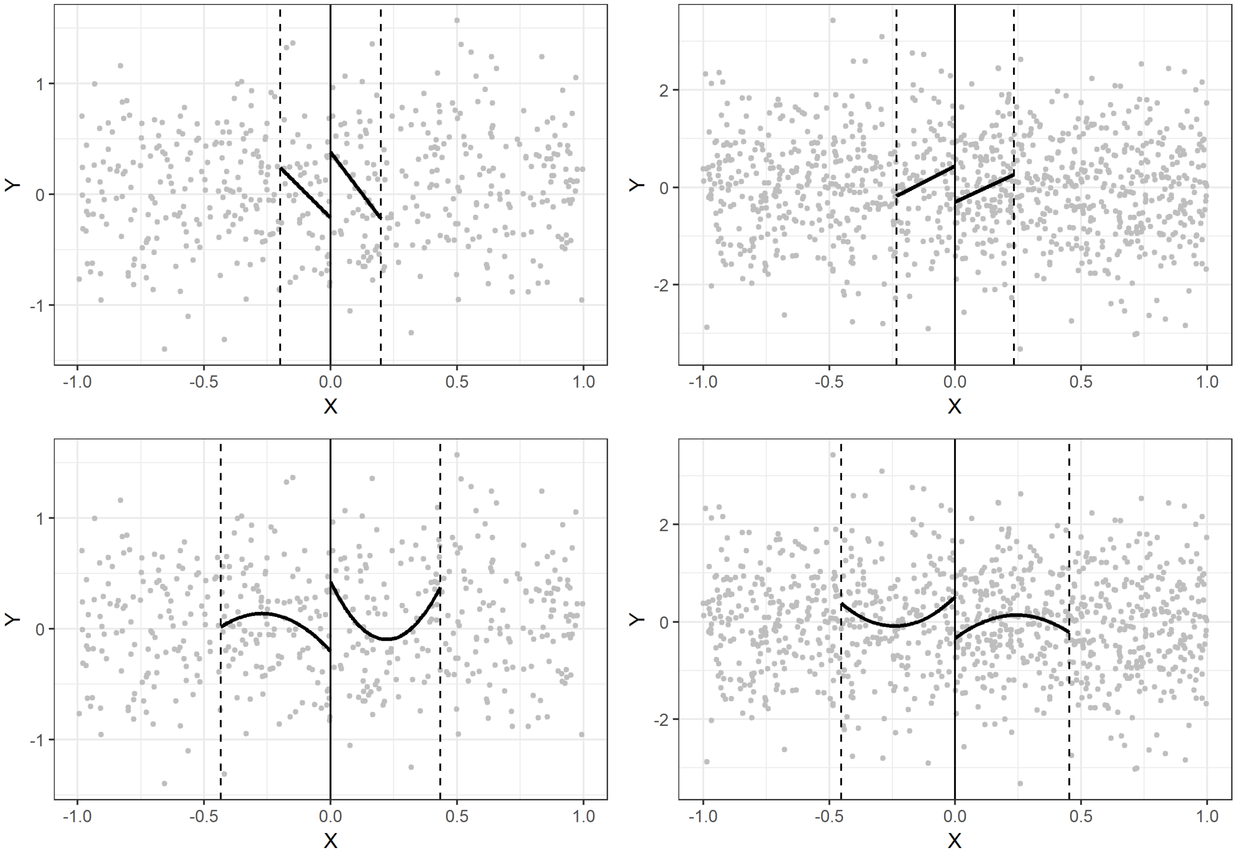

Figure 1: Zig Zag: False positive estimates from the Monte Carlo simulation. Dashed vertical lines are the CCT MSE-optimal bandwidth. Solid vertical line is the cutoff. Gray points are the raw data, and black lines are the local low-order polynomial fits. Left two figures generated with parameters $\tau = 0$, $\beta = 0$, $\sigma = \frac{1}{2}$, $n = 500$. Right two figures generated with parameters $\tau = 0$, $\beta = 0$, $\sigma = 1$, $n = 1000$.

When $\tau = 0$, CCT 95% confidence intervals reject the null hypothesis ($H_0 : \tau = 0$) roughly 7% of the time. Of these rejections, over 90% display the characteristic 'Zig Zag' pattern illustrated in Figure 1 – steep-sloped regression functions on either side of the cutoff and a treatment effect of the opposite sign.[1] This pattern is emblematic of false positive or

[1] Or, when estimated using a local quadratic regression, a concave regression function on one side of the cutoff, and a convex function on the other side.

exaggerated claims from noisy RD data, and will be present in the two empirical illustrations that follow.

In some cases, there may be a good theoretical reason to believe that the treatment effect will move in the opposite direction of the slope of the regression function. For examples, see RD studies on the incumbency disadvantage in Latin America (Klašnja and Titiunik, 2017) or the effect of property tax incentives on homeowners' mobility rates (Ferreira, 2010). But one would rarely expect the slope of the conditional expectation function to diverge sharply only in the neighborhood of the cutoff (for an exception, see Hays, Franzese and Ornstein (2019)). Absent such a theoretical motivation, large slope or concavity estimates near the cutoff are indicative of overfitting to noisy data.

Next, consider what happens as τ increases. In the top two panels of Figure 2, the value of τ is small relative to noise (σ). As a result, the test is low-powered, only rejecting the null hypothesis in 9% and 17% of iterations. And when an estimate passes the statistical significance threshold, it tends to be many multiples of the true value (Type M error) or have the wrong sign (Type S error). The largest distortions are associated with large estimated slopes on each side of the cutoff (the sum of slopes is plotted on the x-axis).

Only when power is high (bottom two panels) are statistically significant estimates roughly centered around the true value of τ. And even then, large slope parameters are indicative of a mis-estimated treatment effect.

2 Design Analysis

There exists a wide array of now-standard falsification tests in the RD literature. These include density tests for manipulation of the running variable (McCrary, 2008), covariate balance tests (Lee, 2008), and placebo tests on alternative cutoff points. Although these tests are invaluable for assessing the validity of the RD design, none of them guard against

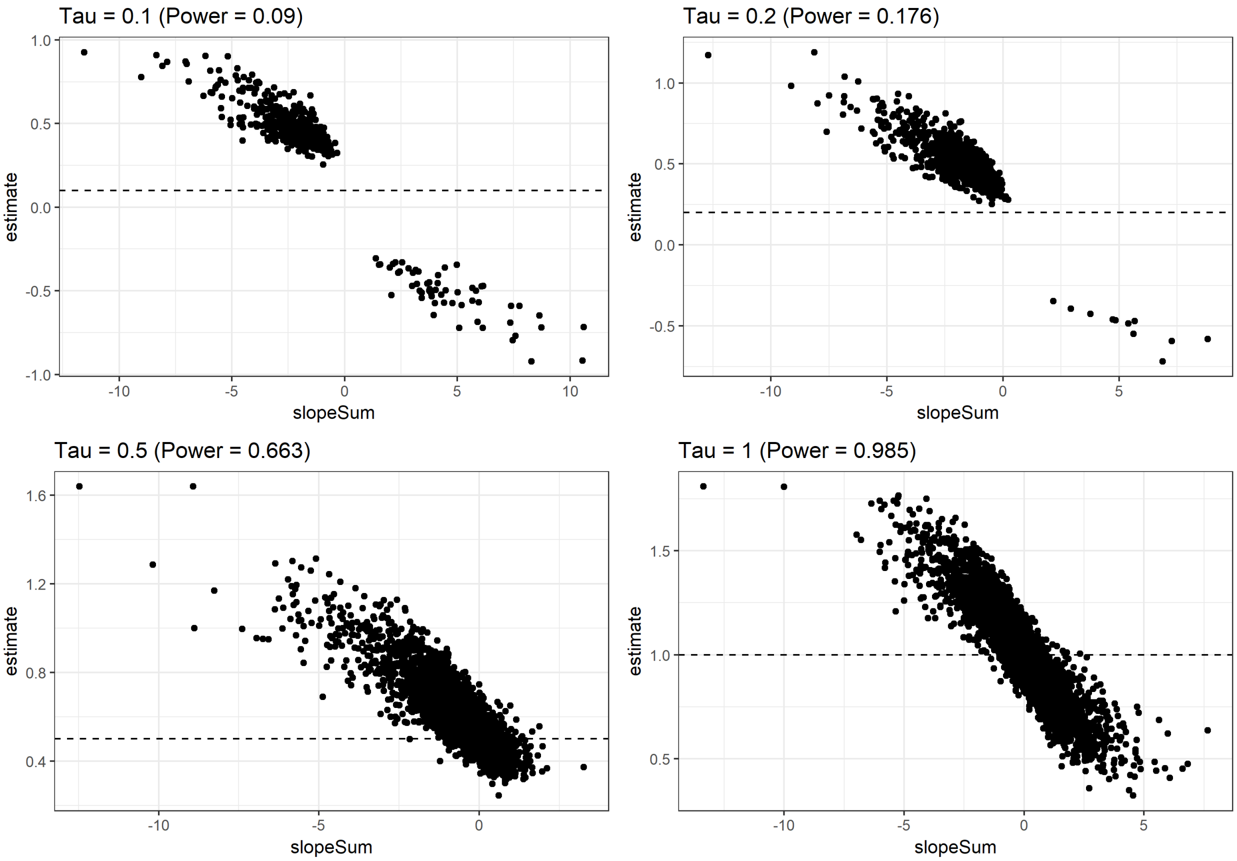

Figure 2: Observed treatment effect and slope estimates from Monte Carlo after applying statistical significance filter $\alpha = 0.05$, for varying values of τ. More extreme slope estimates yield more distorted treatment effect estimates. When the true value of τ is low, the estimated treatment effects from cases that pass statistical significance are greatly exaggerated (Type M error) and many have the wrong sign (Type S error).

the problems of low power described above. In fact, as Hartman and Hidalgo (2016) note, low power *increases* the likelihood of failing to reject the null hypothesis of no difference in conventional balance and placebo tests.

In light of these issues, I propose two solutions: a method for formal power analysis in RD designs and improved guidelines for data visualization.

In experimental studies, the quantity of interest from a power analysis is typically the minimum sample size necessary to achieve 80% power. Because sample size is typically less manipulable in observational studies, this quantity is less useful for RD analysis. Instead, I propose computing the *Minimum Discernible Effect* (MDE). This is the value of τ that, if true, would be detected with high power by the RD design. We can derive this quantity as follows. Calonico, Cattaneo and Titiunik (2014) show that

$$\frac{\hat{\tau}^{bc} - \tau}{\sqrt{\mathbf{V}^{bc}}} \xrightarrow{d} \mathcal{N}(0,1)$$

where $\hat{\tau}^{bc}$ is the bias-corrected local polynomial regression estimate of τ, and $\sqrt{\mathbf{V}^{bc}}$ are robust standard errors. Under this approach, $(1-\alpha)$ confidence intervals are given by $\left[\hat{\tau}^{bc} \pm \sqrt{\mathbf{V}^{bc}}\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right]$.

Power is defined as one minus the Type II error rate – the probability that, conditional on $H_1 : \tau = \theta$, the test will reject $H_0 : \tau = 0$ at the α level. Given the confidence intervals defined above, power $(1-\beta)$ can be estimated as follows:

$$1-\beta = \Phi\left(\frac{\tau}{\sqrt{\mathbf{V}^{bc}}} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) + \Phi\left(-\frac{\tau}{\sqrt{\mathbf{V}^{bc}}} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) \quad (1)$$

The two terms on the righthand side of equation (1) denote the probabilities of estimating a value of $\hat{\tau}^{bc}$ in the left or right tails of the null distribution, conditional on the true effect being equal to τ. From this, we can compute the MDE by solving for τ:[2]

$$MDE = \sqrt{\mathbf{V}^{bc}}\left(\Phi^{-1}\left(1-\frac{\alpha}{2}\right) + \Phi^{-1}(1-\beta)\right) \quad (2)$$

This quantity has a straightforward interpretation: *given the variation and sample size near the cutoff, what is the smallest value of τ that would be reliably detected by the research design?* The advantage MDE as an estimate of power is that it is a function of only the standard error, and is not sensitive to one's choice of a prior on τ (unlike conventional power analyses, or those proposed by Gelman and Carlin (2014)). The downside is one common to post-hoc power analyses: the estimate of the noisiness of the data ($\sqrt{\mathbf{V}^{bc}}$) comes from the

[2] Without loss of generality, suppose that $\tau > 0$. The second term in equation (1) denotes the probability of rejecting the null hypothesis on the opposite side (i.e. a Type S error). In a high-powered test, this term shrinks to zero, so when solving for equation (2) it drops out.

sample itself, rather than a prior estimate of variability.

2.1 RD Plots

One great strength of the RD approach is that its results can be displayed graphically, and it is common practice for researchers to include data visualizations when reporting their results. Calonico, Cattaneo and Titiunik (2015) suggest using binned scatter plots for such visualizations, and propose a number of data-driven bandwidth selectors to automate the process of choosing a bin size. The advantage of this approach is that it is easier to "see" a regression discontinuity with a binned scatter plot than in a scatter plot of raw data, particularly when working with large datasets.

The disadvantage of the binning approach is that it obscures issues of power by reducing the apparent variation of the dependent variable. To illustrate, consider Figure 3. The top two panels display raw data, generated by the Monte Carlo simulation described above.[3] In both cases, the true value of τ is equal to 1, but in the right panel the standard deviation of the outcome variable (σ) is reduced by half. The middle two panels display the binned scatter plot approach to RD visualization. Surprisingly, these two middle panels are nearly identical, despite the fact that the righthand dataset has much greater power.

Clearly, the evidence for a discontinuity is stronger in the righthand dataset, and a visualization of the RD analysis should make that clear. The final row presents a solution that combines the strengths of the first two: it includes both the raw and binned scatter plots, and rather than a high-order global polynomial, includes the selected bandwidth and local polynomial fits used to estimate the effect.

The advantages of this final approach to visualization are threefold. First, it does not discard the raw data, so it is easier to observe the underlying variation in both the X and Y variables. At a glance, one can detect problems with both power in the outcome variable

[3] Left panel parameters: $\tau = 1$, $\beta = 1$, $\sigma = 1$. Right panel parameters: $\tau = 1$, $\beta = 1$, $\sigma = \frac{1}{2}$

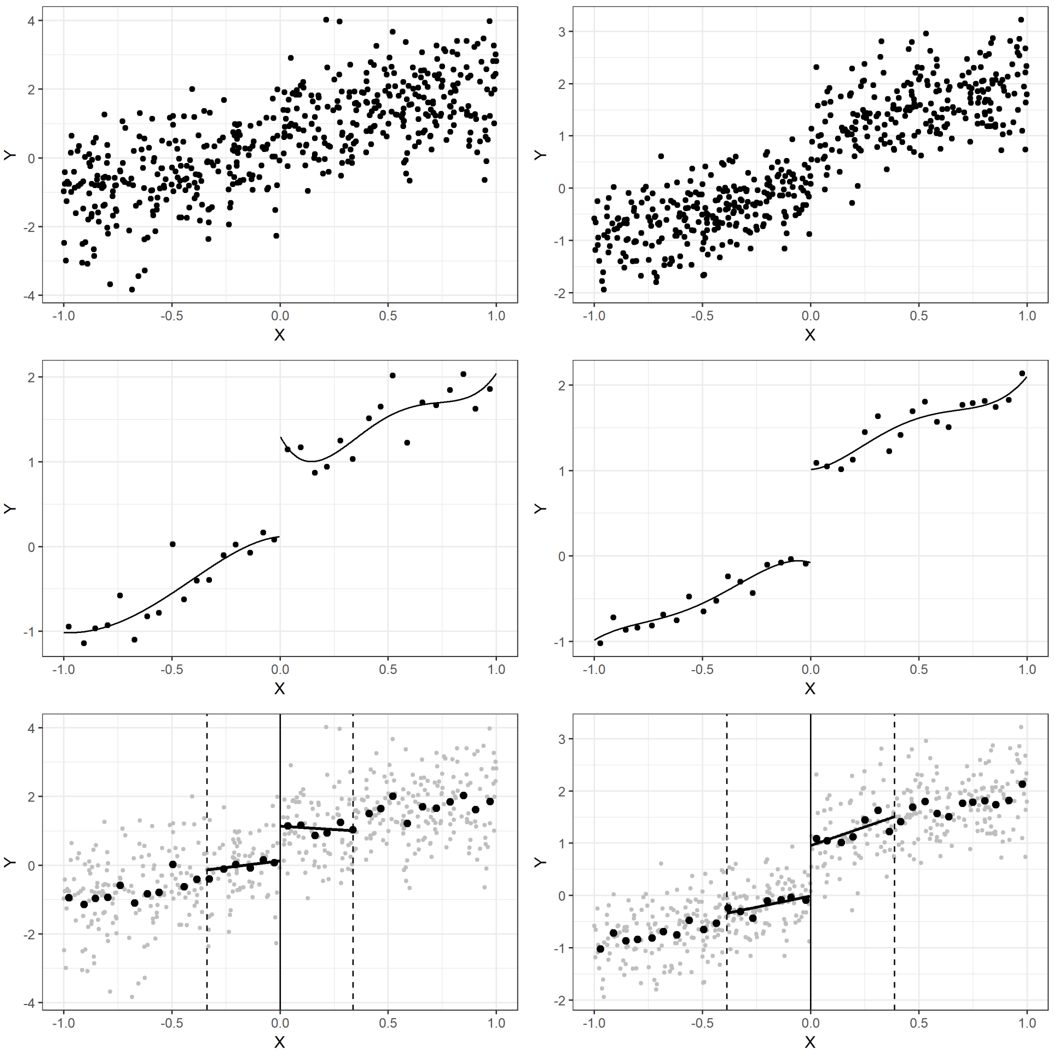

Figure 3: Raw, binned, and combined scatter plot approaches to visualizing a regression discontinuity.

and the density of the running variable. Second, it maintains the binned scatter plot, so that even in studies with high power and large datasets, the discontinuity is apparent in the visualization. Third, it includes details of the RD estimation, including the selected bandwidth and local polynomial fit. This makes it easier to observe when a large magnitude result may be driven by implausible slope estimates in the neighborhood of the cutoff.

3 Empirical Illustrations

The two empirical illustrations that follow share several common features. First, they both use electoral thresholds to study the effect of certain representatives and parties holding power on observed political outcomes. Because the control of political power varies discontinuously at electoral thresholds, the RD design is appropriate in each case, and the causal inference itself is not problematic. Second, they both employ state-of-the-art methods in RD analysis, using local polynomial approaches with automated bandwidth selectors and robust standard errors. Third, they each conduct extensive falsification and robustness tests to ensure that their result is not sensitive to choice of bandwidth and that pre-treatment covariates are balanced near the cutoff. Their estimated treatment effects are large and highly statistically significant.

Nevertheless, in both cases the data they employ is either too noisy or small to reliably detect plausible treatment effects, and conventional binned scatter plot visualizations obscure rather than illuminate problems with low power. I will show how the type of design analysis and visualizations I propose above can help shed light on these issues.

3.1 The Radical Right and Party Manifestos

Abou-Chadi and Krause (2018) study how the presence of radical right parties influence the platforms of mainstream parties. Their causal identification strategy is based on electoral thresholds in parliamentary systems; typically it is required that a party clear some percentage of the total vote before gaining representation in parliament, where the particular threshold varies by country. The authors compare elections where radical right parties barely exceeded the threshold (gaining representation) and barely missed the threshold, observing how mainstream political parties respond.

The dependent variable is change in a measure of Cultural Protection in the party's manifesto during the following election. These data are compiled by the Comparative Manifestos Project (Volkens, Pola Lehmann, Theres Matthieß, Nicolas Merz and Werner, 2015), and the outcome variable is a function of the difference between the number of favorable mentions of cultural diversity and encouragement of integration and cultural homogeneity in the party's platform (Lowe et al., 2011). In their paper, Abou-Chadi and Krause (2018) present estimates from a diverse range of specifications, ranging from 3.1 to 4.9. Replicating these results using CCT bias-corrected standard errors and robust confidence intervals yields an estimate of 3.96, with 95% confidence interval $[1.7, 6.2]$.

To get a sense of the relative magnitude of this estimated effect, consider Figure 4. This plots the average value of the Cultural Protection score, by country, for each family of political party since 1980. The Cultural Protection score is very noisy from election year to election year, but averaging across years yields predictable patterns. Right-leaning parties tend to score higher on this measure than left-leaning parties, and Nationalist parties – where they exist – typically score 2 to 3 points higher than the average mainstream party.

In this context, an estimated effect size of 3.9 is enormous. If true, it suggests that not only do mainstream parties respond to Radical Right representation by moving their platforms to the right, but they do so in such a way that their rhetoric completely closes or even overtakes the average gap between mainstream and rightwing nationalist party positions.

Figures 5 and 6 present two graphical representations of the RD analysis. The first is a binned scatterplot with global, high-order polynomial fits on each side, as recommended by Calonico, Cattaneo and Titiunik (2015), and the second is the modified visualization as proposed above. The second visualization makes clear that the estimated effect is based on a small number of observations to the right of the cutoff, and an implausibly large slope estimate to the left of the cutoff. This is more transparent representation of the analysis, because it displays the local linear regressions used to estimate the average treatment effect. The MDE, as computed by equation (2), is 3.2.

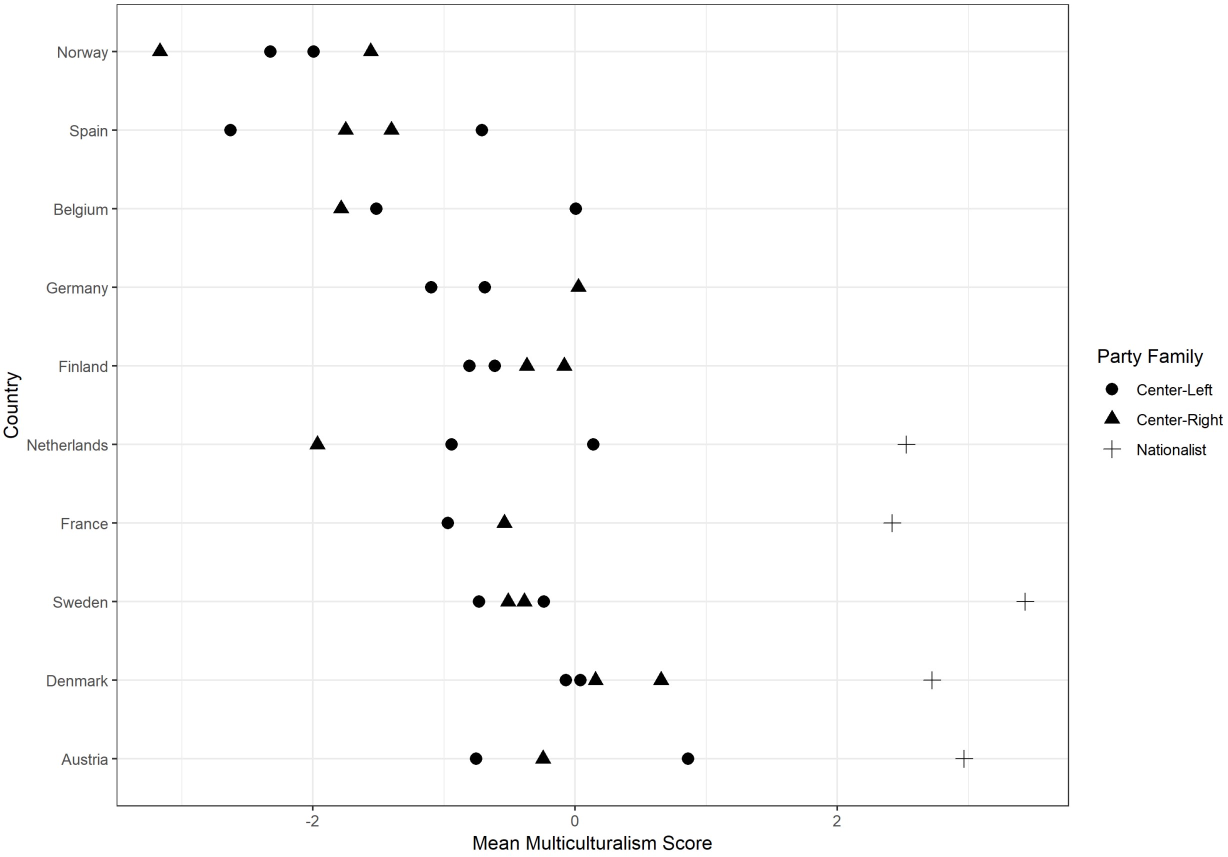

Figure 4: Mean Cultural Protection score by country and party family (all elections post-1980; Center-Left includes Social Democrats and Liberals, Center-Right includes Christian Democrats and Conservatives).

To be clear, none of the foregoing analysis suggests that the underlying claim in Abou-Chadi and Krause (2018) is *wrong.* It may yet be the case that parties change their platforms in response to representation of the Radical Right. But the study as published makes a very strong claim – an average treatment effect of 3.9 – on the basis of tenuous evidence. A much larger sample of elections would be required before such a study could reliably detect a more plausible causal effect.

3.2 Diversity and Municipal Public Spending

The second empirical illustration comes from Beach and Jones (2017), who study the effect that diverse representation on city councils has on municipal-level public goods spending.

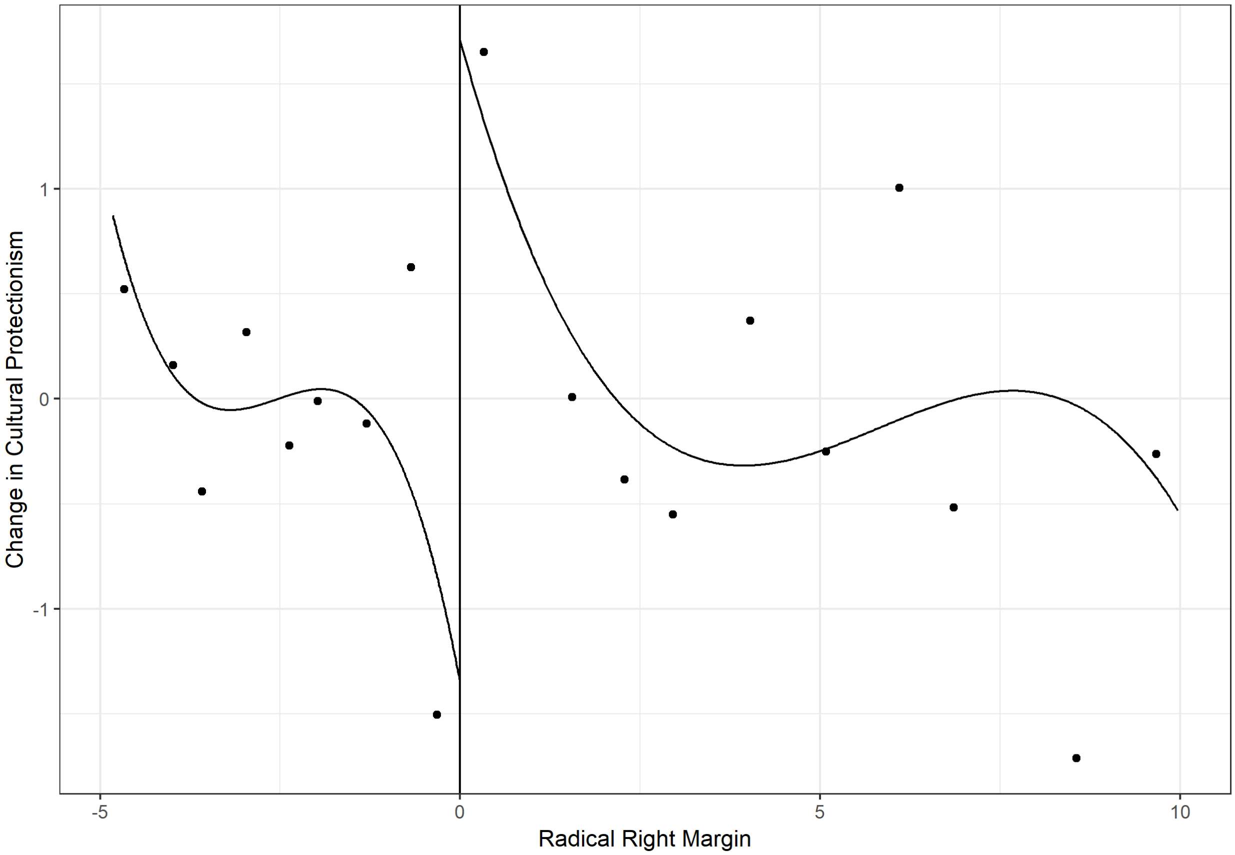

Figure 5: The Radical Right Regression Discontinuity (Abou-Chadi and Krause, 2018), binned scatterplot with third-order polynomial fit.

There is a large literature on this topic, dating back to Alesina, Baqir and Easterly (1999) and continuing through Hopkins (2011) and Trounstine (2015), finding that in ethnically diverse and segregated cities, municipal governments devote less spending to public goods than in ethnically homogeneous cities. All of these studies rely on cross-sectional or longitudinal regression analysis of public finance data, and so lack a credible causal identification. Beach and Jones (2017) approach the problem through an RD analysis of city council elections in California. Because narrowly elected city councilors from a 'non-modal' ethnicity should increase the diversity of a city council, the discontinuity at the plurality margin allows for the causal identification of an effect of council diversity on public spending.

The primary analysis regresses per capita public goods spending on the vote margin of 'non-modal' candidates for races where a non-modal candidate faced a modal candidate.

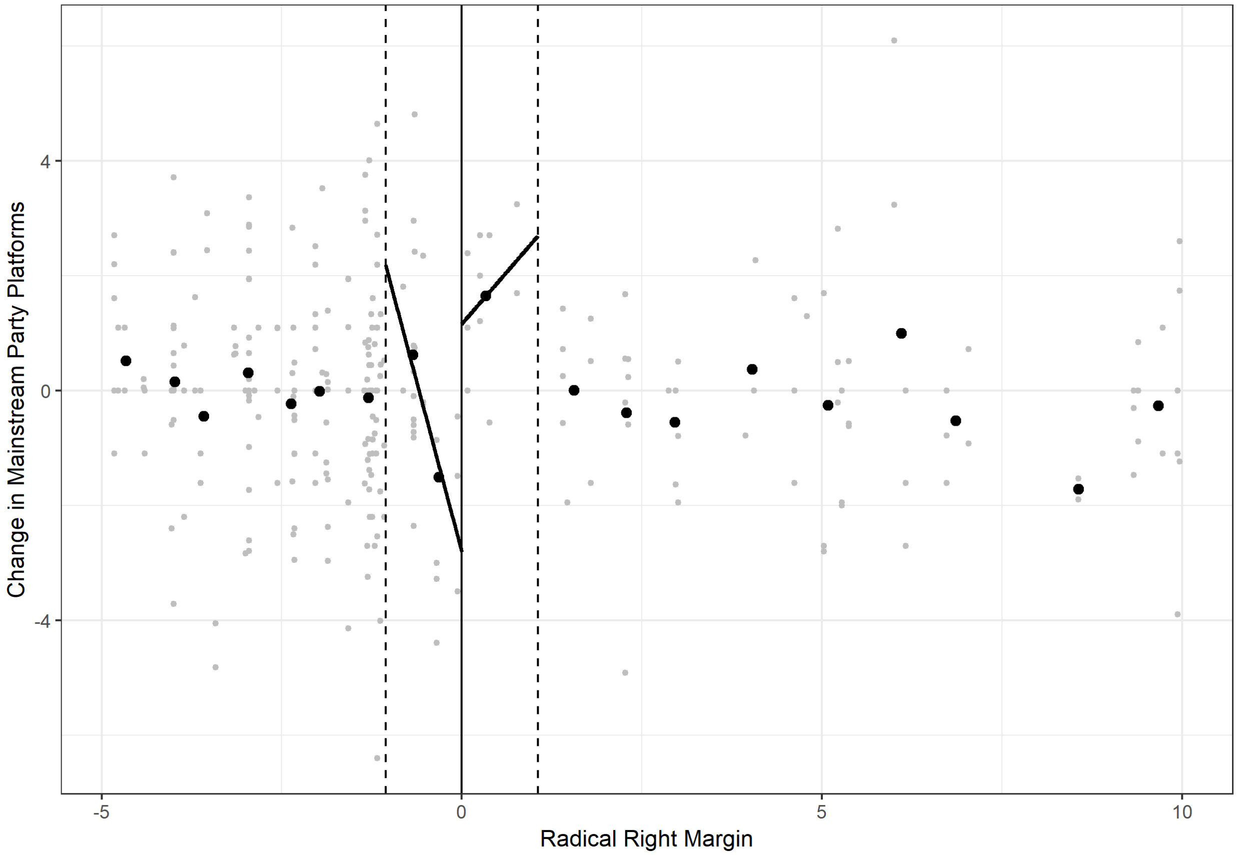

Figure 6: The Radical Right Regression Discontinuity (Abou-Chadi and Krause, 2018). Small gray points are the raw data and large black points are binned. The solid vertical line is the (normalized) electoral representation threshold and dashed vertical lines are the CCT MSE-optimal bandwidths. Black lines are the estimated local linear regression functions.

Figure 7 visualizes the results from this analysis as a binned scatter plot. The estimated effect is -0.31 on a log scale, implying that the election of a councilor that increases diversity causes a staggering 31% drop in public goods spending.

To consider this estimate in context, a 31% cut in public goods spending amounts to roughly $463 per capita for the average city. This is equivalent to eliminating all spending on police (13% in the average city), fire protection (9%), and roads (9%). By comparison, similar studies suggest much more modest effects. In an RD analyis of Democratic vs. Republican mayoral candidates, de Benedictis-Kessner and Warshaw (2016) find that narrowly-elected Democratic mayors increase public spending by 5% on average. In their original study of US cities, Alesina, Baqir and Easterly (1999) find that ethnically diverse cities spend 6% to

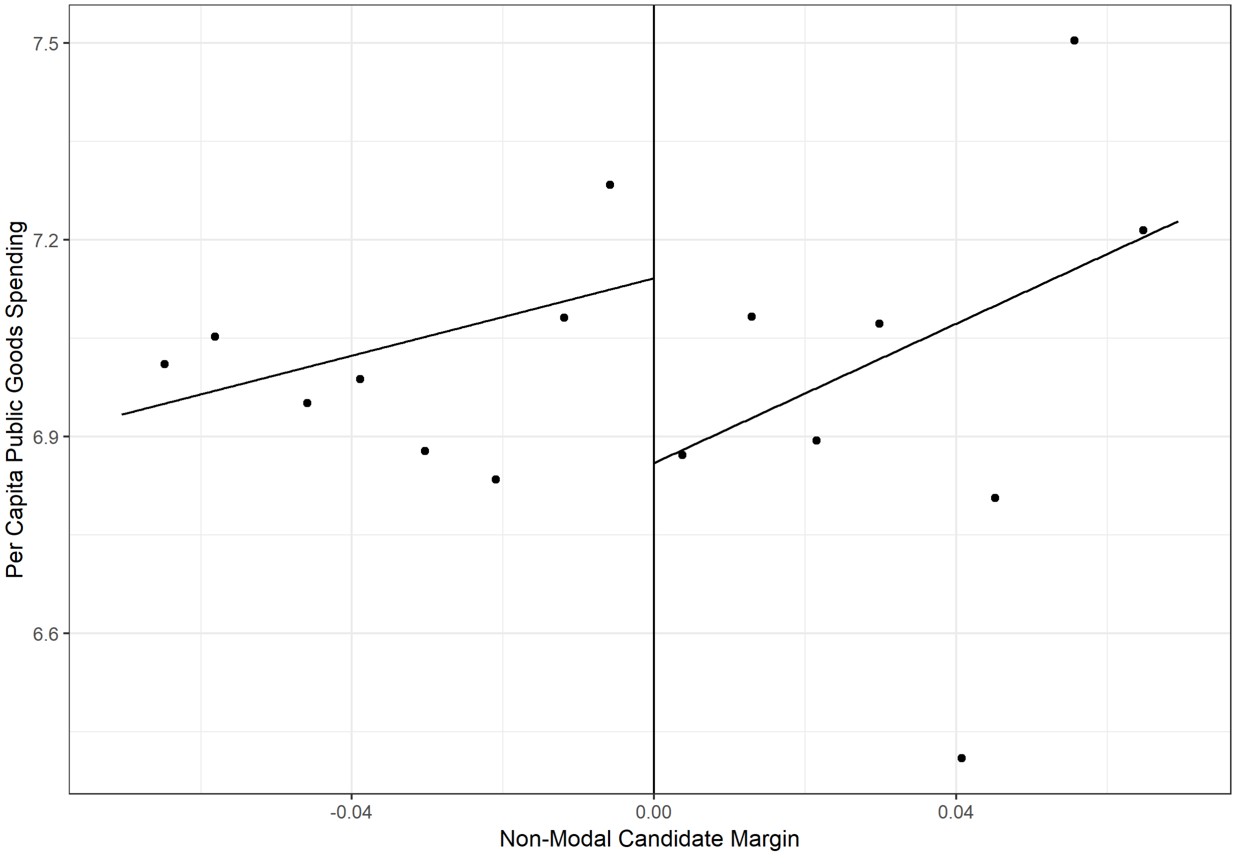

Figure 7: Diverse City Councils and Public Spending RD (Beach and Jones, 2017), Binned Scatter Plot

9% lower shares on "productive" public goods than ethnically homogeneous cities.[4] All of these estimates are several orders of magnitude smaller than 30%.

Even assuming that the effect of a non-modal city councilmember is comparable to that of a Democratic mayor,[5] the RD analysis would be too underpowered to reliably detect a more plausible effect. The MDE, as defined by equation (2), is equal to 0.32. The modified RD plot in Figure 8 makes the problem clear: although a substantial fraction of the sample lies within the bandwidth, the variation in public goods spending is very large relative to any plausible effect size. A sufficiently high powered study to detect an effect size of 5%

[4] Note that this is the estimated effect of changing the Herfindahl ethnic fractionalization index from 0 to 1 (i.e. from perfect homogeneity to perfect heterogeneity).

[5] Not an unreasonable assumption, since in many municipalities the mayor does not wield strong executive power, and is essentially the chair of the city council.

would require many times more observations within the neighborhood of the cutoff.

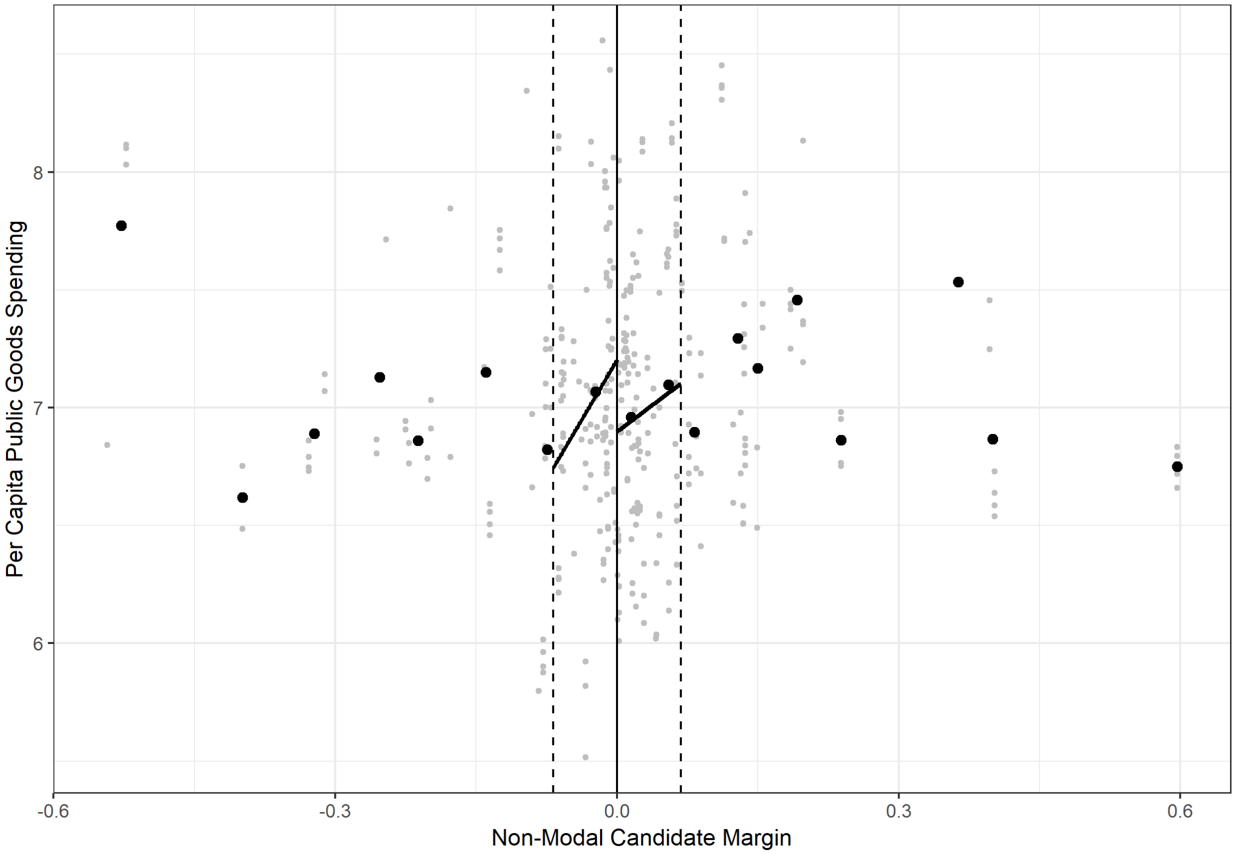

Figure 8: Diverse City Councils and Public Spending RD (Beach and Jones, 2017). Visualization with raw data.

4 Conclusion

In this paper, I have demonstrated how low-powered regression discontinuity analyses can yield misleading and exaggerated causal effect estimates, characterized by an implausibly large divergence in the slope of the conditional expectation function in the neighborhood of the cutoff. Given that standard robustness tests do not offer a fail-safe against such problems, I propose the more routine use of power analyses and data visualizations as described above.

Conventional frequentist hypothesis tests encourage researchers to view statistical inference as fundamentally binary: reject the null hypothesis or do not. But arguably a more important task is to accurately assess the size and strength of effects. This requires more rigorous standards of evidence, modesty about what inferences can be drawn from small, noisy datasets, and a perspective that does not view each new study in isolation, but as a part of a larger whole. The suggestions here offer a promising step in that direction.

References

Abou-Chadi, Tarik and Werner Krause. 2018. "The Causal Effect of Radical Right Success on Mainstream Parties' Policy Positions: A Regression Discontinuity Approach." *British Journal of Political Science* pp. 1–19.

Alesina, Alberto, Reza Baqir and William Easterly. 1999. "Public Goods and Ethnic Divisions." *The Quarterly Journal of Economics* 114(4):1243–1284.

Beach, Brian and Daniel B. Jones. 2017. "Gridlock: Ethnic Diversity in Government and the Provision of Public Goods." *American Economic Journal: Economic Policy* 9(1):112–136.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson and Marcus R. Munafò. 2013. "Power failure: why small sample size undermines the reliability of neuroscience." *Nature Reviews Neuroscience* 14(5):365–376.

Calonico, Sebastian, Matias D. Cattaneo and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82(6):2295–2326.

Calonico, Sebastian, Matias D. Cattaneo and Rocío Titiunik. 2015. "Optimal Data-Driven Regression Discontinuity Plots." *Journal of the American Statistical Association* 110(512):1753–1769.

Cohen, Jacob. 1992. "A Power Primer." *Psychological Bulletin* 112(1):155–159.

de Benedictis-Kessner, Justin and Christopher Warshaw. 2016. “Mayoral Partisanship and Municipal Fiscal Policy.” *The Journal of Politics* 78(4):1124–1138.

de la Cuesta, Brandon and Kosuke Imai. 2016. “Misunderstandings About the Regression Discontinuity Design in the Study of Close Elections.” *Annual Review of Political Science* 19:375–396.

Ferreira, Fernando. 2010. “You can take it with you: Proposition 13 tax benefits, residential mobility, and willingness to pay for housing amenities.” *Journal of Public Economics* 94(9-10):661–673.

Gelman, Andrew and Guido Imbens. 2017. “Why high-order polynomials should not be used in regression discontinuity designs.” *Journal of Business & Economic Statistics* .

Gelman, Andrew and John Carlin. 2014. “Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors.” *Perspectives on Psychological Science* pp. 1–11.

Hahn, Jinyong, Petra Todd and Wilbert Van Der Klaauw. 2001. “Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design.” *Econometrica* 69(1):201–209.

Hartman, Erin and F Daniel Hidalgo. 2016. “What’s the Alternative?: An Equivalence Approach to Balance and Placebo Tests.” *Working Paper* .

Hays, Jude C., Robert J. Franzese and Joseph T. Ornstein. 2019. “Estimating the Interest-Premium Cost of Left Government by Regression Discontinuity Analysis of Close Elections.” *Working Paper* .

Hopkins, Daniel J. 2011. “The limited local impacts of ethnic and racial diversity.” *American Politics Research* 39(2):344–379.

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

Klašnja, Marko and Rocio Titiunik. 2017. "The Incumbency Curse: Weak Parties, Term Limits, and Unfulfilled Accountability." *The American Political Science Review* 111(1):129–148.

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

Lowe, Will, Kenneth Benoit, Slava Mikhaylov and Michael Laver. 2011. "Scaling policy preferences from coded political texts." *Legislative Studies Quarterly* 36(1):123–155.

McCrary, Justin. 2008. "Manipulation of the running variable in the regression discontinuity design: A density test." *Journal of Econometrics* 142(2):698–714.

Trounstine, Jessica. 2015. "Segregation and Inequality in Public Goods." *American Journal of Political Science* 60(3):709–725.

Volkens, Andrea, Sven Regel Pola Lehmann, Theres Matthieß, Nicolas Merz and Annika Werner. 2015. *The Manifesto Data Collection: Manifesto Project (MRG/CMP/MARPOR). Version 2015a.* Berlin: .