The curious case of a replication

Sumit Mishra
sumit.mishra@krea.edu.in
Krea University, Sri City *

31 August 2023

This short note is a response to Yajnavalkya (2023)[1] which argues that some of the results in Das (2023) have serious problems due to data and estimation errors. I have three main findings to report.

Density test for non-BJP ruled states

1. There is no McCrary density test failure in the case of non-BJP ruled states.

Yajnavalkya (2023) shows, using the R command `DCdensity`, that there was a discontinuous shift in BJP victory margin in non-BJP ruled states too. This procedure relies on Imbens-Kalayanaraman (IK) bandwidth. Das (2023) uses the optimal bandwidth computed using the MSERD approach (Calonico, Cattaneo, and Titiunik (2014)). Unfortunately, the IK optimal bandwidth can incorrectly indicate discontinuity even when it's not there leading to the rejection of null (Kuehnle, Oberfichtner, and Ostermann (2021))[2]. Using the standard R command for the density test (`rddensity`) as well as running `DCdensity` with half the bandwidth reported in the Medium post[3], I rule out the possibility of failure of McCrary test, for the non-BJP ruled states, in the 2019 Lok Sabha elections.

Growth rate of electorate in close elections

2. The growth rate of electorate is indeed lower in close contests where BJP won.

*I would like to thank an anonymous student at Krea University for providing excellent research assistance. The replication material is available here.

[1] The dataset and code for the analysis is available on Github.

[2] For detailed technical exposition on density estimators, please see Cattaneo and Vazquez-Bare (2017) and Cattaneo, Jansson, and Ma (2020)

[3] It is standard in the literature to see how sensitive results are to changes in the bandwidth.

The original post on Medium uses the RDestimate command in R whose results are strikingly similar to the aggregate PC level findings reported in Das (2023). The reported estimate is -0.044 (4.4%) with a p-value of 0.01. However, one should transparently report all possible results by varying two key parameters: bandwidth (the window around the cutoff) and kernel (the assignments of weights to the observations around the cutoff). The tables below report all the results produced by rdrobust[4] command in R. There are two sets of results. The first set (Table 1) reports the estimates by varying the bandwidth. The results remain stable and close to the estimates provided in Das (2023). The confidence intervals of all these estimates overlap and we can argue that in close elections won by the BJP, the growth rate of electorate is 5% lower than the seats won by non-BJP parties. One might be worried that these results are driven by the choice of kernel. I run the model with the optimal bandwidth and vary the kernel; results are reported in Table 2. I show that the effect size remains stable and closer to those reported in Das (2023).

Figure 1 plots the estimates from the nonparametric Regression Discontinuity Design (RDD) which Yajnavalkya (2023) complains is missing from the original paper. The author argues, using the average outcome on the far left of the chart, that the seats which non-BJP parties won saw slower growth of Muslim voters. This interpretation is incorrect. We must focus on the points around the cutoff. Figure 1 suggests that there is a sharp fall in growth rate of electorate in closely contested elections which BJP won.

Dealing with 'outliers'

3. We cannot arbitrarily drop observations in the neighbourhood of the cutoff while using regression discontinuity design.

The last argument in the Medium post is about handling outliers in the data. Yajnavalkya (2023) adopts a strange strategy to show that the growth rate of voters between 2014 and 2019 was no different in close contests that BJP won or lost. The author removes some of the observations around the cutoff. This is erroneous. One cannot pick and choose observations around the cutoff; RDD hinges on having all possible data points in the vicinity of cutoff. The procedure to deal with the outlier problem in RDD is to run what is known as the 'donut' RDD[5]. I perform the donut RDD by trimming the running variable at the margin of 0.5% (0.005) around the cut-off and show that the results only shift marginally. Table 3 reports these results.

To sum up, the estimates and graphs provided in Yajnavalkya (2023) are misleading and based on poor understanding of how regression discontinuity design works.

[4]While RDestimate relies on local linear regression, rdrobust deploys local polynomial regressions.

[5]The procedure drops units that are very close to the cutoff and tells us how sensitive the results are to the 'outliers'.

Table 1: Electorate growth rate in PCs barely won by BJP (Different bandwidths)

	Bandwidth = 0.127 (MSERD)	Bandwidth = 0.106 (CER-SUM)	Bandwidth = 0.094 (CER-COMB1)
Bias-Corrected	-0.051**	-0.049**	-0.047*
	(0.017)	(0.018)	(0.019)
Robust	-0.051**	-0.049*	-0.047*
	(0.019)	(0.020)	(0.020)
Number of observations	156	129	113

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The first column reports the bandwidth from the MSERD procedure, the second column has the bandwidth from the CER(coverage error rate)-optimal bandwidth selector for the sum of regression estimates, and the third column applies the bandwidth from the minimum of two CER-optimal bandwidth selectors. Robust standard errors in the parentheses.

Table 2: Electorate growth rate in PCs barely won by BJP (Different kernels)

	Kernel: Triangular	Kernel: Epanechnikov	Kernel: Uniform
Bias-Corrected	-0.051**	-0.052**	-0.043**
	(0.017)	(0.017)	(0.016)
Robust	-0.051**	-0.052**	-0.043*
	(0.019)	(0.019)	(0.018)
Number of observations	156	143	159

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
All the estimates are based on the optimal bandwidth using the MSERD procedure (Calonico et. al. 2014). Robust standard errors are reported in the parentheses.

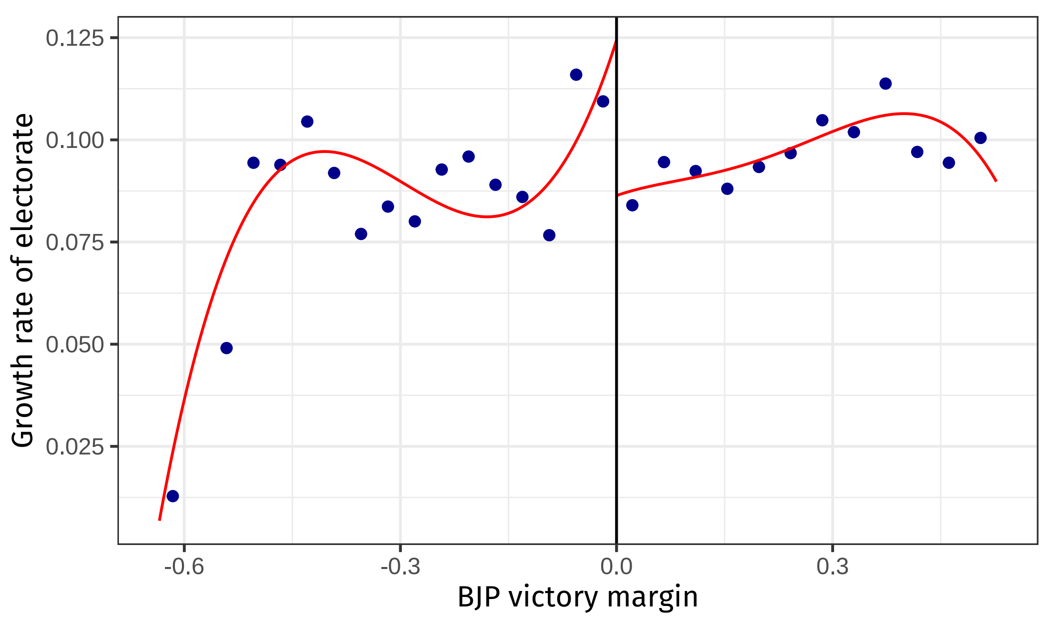

Figure 1: RD chart: Growth rate of electorate v BJP victory margin

Table 3: Donut RD estimate

Bias-Corrected	-0.042**
	(0.016)
Robust	-0.042*
	(0.019)
Number of observations	184

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
The estimates are based on the optimal bandwidth using the MSERD procedure (Calonico et. al. 2014). Robust standard errors are reported in the parentheses.

References

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

Cattaneo, Matias D, Michael Jansson, and Xinwei Ma. 2020. "Simple Local Polynomial Density Estimators." *Journal of the American Statistical Association* 115 (531): 1449–55.

Cattaneo, Matias D, and Gonzalo Vazquez-Bare. 2017. "The Choice of Neighborhood in Regression Discontinuity Designs." *Observational Studies* 3 (2): 134–46.

Das, Sabyasachi. 2023. "Democratic Backsliding in the World's Largest Democracy." *Available at SSRN 4512936*.

Kuehnle, Daniel, Michael Oberfichtner, and Kerstin Ostermann. 2021. "Revisiting Gender Identity and Relative Income Within Households: A Cautionary Tale on the Potential Pitfalls of Density Estimators." *Journal of Applied Econometrics* 36 (7): 1065–73.

Yajnavalkya. 2023. "The Ashoka Paper — a Case of Democratic Backsliding or Academic Backsliding?" 2023. https://medium.com/@yajnavalkya10/the-ashoka-paper-a-case-of-democratic-backsliding-or-academic-backsliding-37adf09b467a.