BrazilCVRepProj

Felix Deemer

5/4/2021

Replication Project - Brazil Compulsory Voting Policy Analysis

Original Paper: Compulsory Voting Can Increase Political Inequality: Evidence from Brazil

By Gabriel Cepaluni and F. Daniel Hidalgo

Introduction

The original paper I attempted to replicate is a study of Compulsory Voting (CV) policies in Brazil, determining whether or not these policies achieve their desired effects. Compulsory voting policies are those that aim to increase voter turnout by imposing financial or other penalties on those who fail to do so.

Cepaluni and Hidalgo's paper seeks to demonstrate that in the Brazilian case, where every person between the ages of 18 and 69 is legally required to register and vote, such policies can actually be counterproductive in many ways. One aim of Brazil's policy was to reduce the extent of 'Political Inequality', a phenomenon in which certain groups of people (wealthier, better educated, and more privileged groups) tend to vote at higher rates, acquiring disproportionate political influence. However, the punishment for not voting in Brazil involves restricted access to state services, which survey data collected by Cepaluni and Hidalgo indicated are disproportionately used by more educated groups (education levels were used as a proxy for wealth and privilege). This might mean, they thought, that the CV policy would provide a greater incentive for more highly educated individuals to vote, in fact increasing political inequality.

The researchers sought to investigate how large an effect CV policies have on turnout, and how the size of this effect varies according to education levels. To analyze this effect, they used a Regression Discontinuity (RD) model, in which levels of a variable of interest are modeled on either side of a sharp threshold. This is well suited to CV policy analyses, as the CV policy has a strict age cutoff. Therefore, by comparing turnout between 17-yr-olds (not required to vote) and 18-yr-olds (required to vote), and on the other end of the range, between 69-yr-olds (required to vote) and 70-yr-olds (not required to vote), this can reveal the causal effect of the policy. This is based off the assumption that individuals just above and below the threshold will be very similar in most respects, and that there are not any systematic differences that would make a comparison invalid.

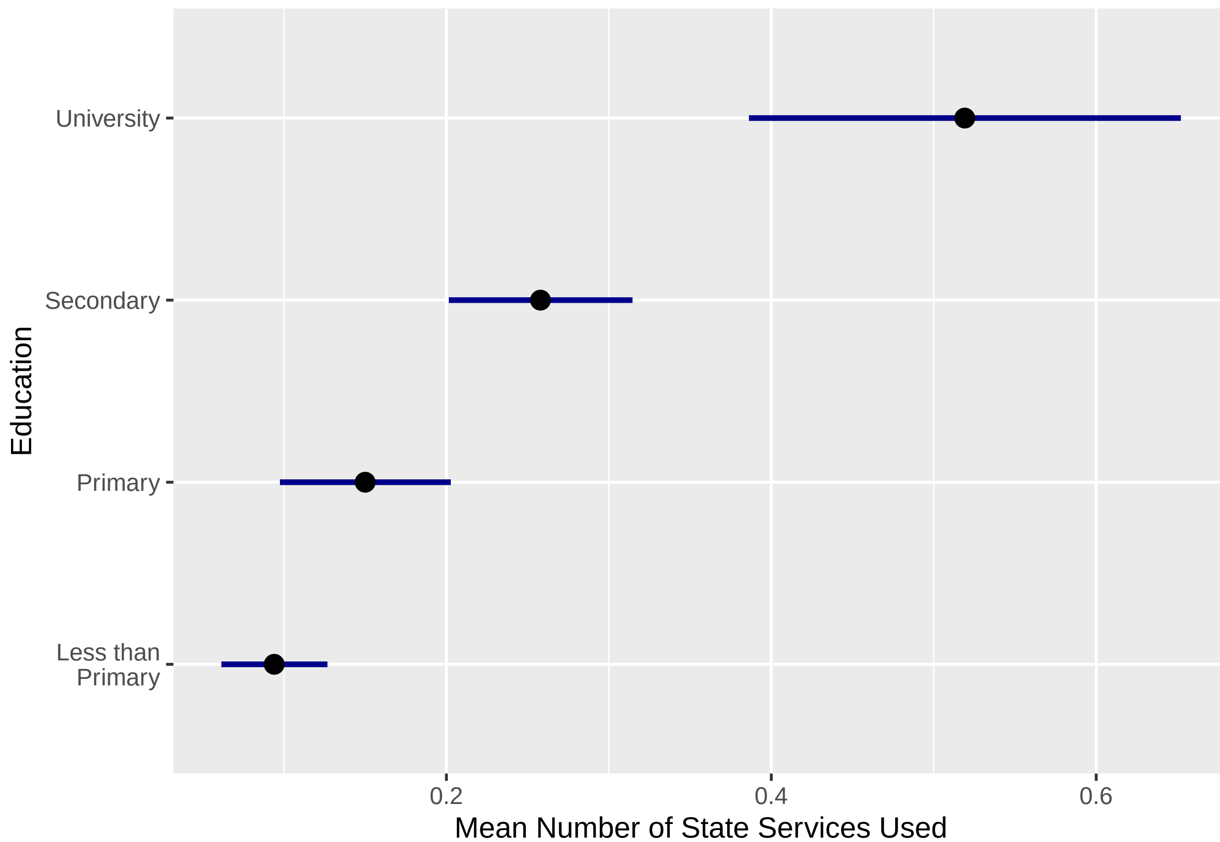
To perform this analysis, they ran local linear models on either side of the threshold (explained in detail below), calculating the difference between the two values in order to estimate the causal effect of the CV policy.

They concluded that the effect of the policy was somewhat larger for groups with higher levels of education, with this difference being most pronounced between those who had completed at least primary education and those that had not. On this basis, they concluded that far from reducing political inequality between the educated and uneducated, Compulsory Voting policies in fact increased it. The implications of this conclusion are that Compulsory Voting policies, when they are implemented, should be more carefully designed to not backfire.

To replicate their model, I performed a Regression Discontinuity model as well, although without the complex system of weighting used by theirs, instead using a slightly simpler linear model. Although their data seemed designed to investigate the effect of CV on multiple different groups, so to extent their analysis slightly I performed the same analysis based on sex instead of education. This allowed me to determine that the pattern seen in political inequality due to education (did/did not) extend to sex as well.

State Service Usage

The research question originally carried out by Cepaluni and Hidalgo was largely based on the survey finding that more educated individuals tend to make use of Brazilian state services to a far greater extent. The survey results are shown in the graph below. I used a t-test to determine the confidence intervals of the mean number of state services used. The means indicate that as education increases, the mean number of state services used also tends to increase, to a greater degree with each increased level of education.



Data Cleaning

Three main data sources were used in the paper - a survey conducted by the researchers themselves, a general population survey, and registered voter data from the Brazilian government. The survey was used to determine the number of state services used by education level, while the registered voter data was used to measure turnout of registered voters.

To attempt to replicate the results obtained by Cepaluni and Hidalgo, I followed a fairly similar data-cleaning process. I began by filtering out those who are illiterate (they cannot vote and so should not be counted for the purposes of this analysis). I also created a new variable 'primary_edu', which was a 0/1 variable type showing whether or not the individual has completed primary school. This was because in the original data, education was a categorical variable with multiple possible string values.

Discontinuity Graph

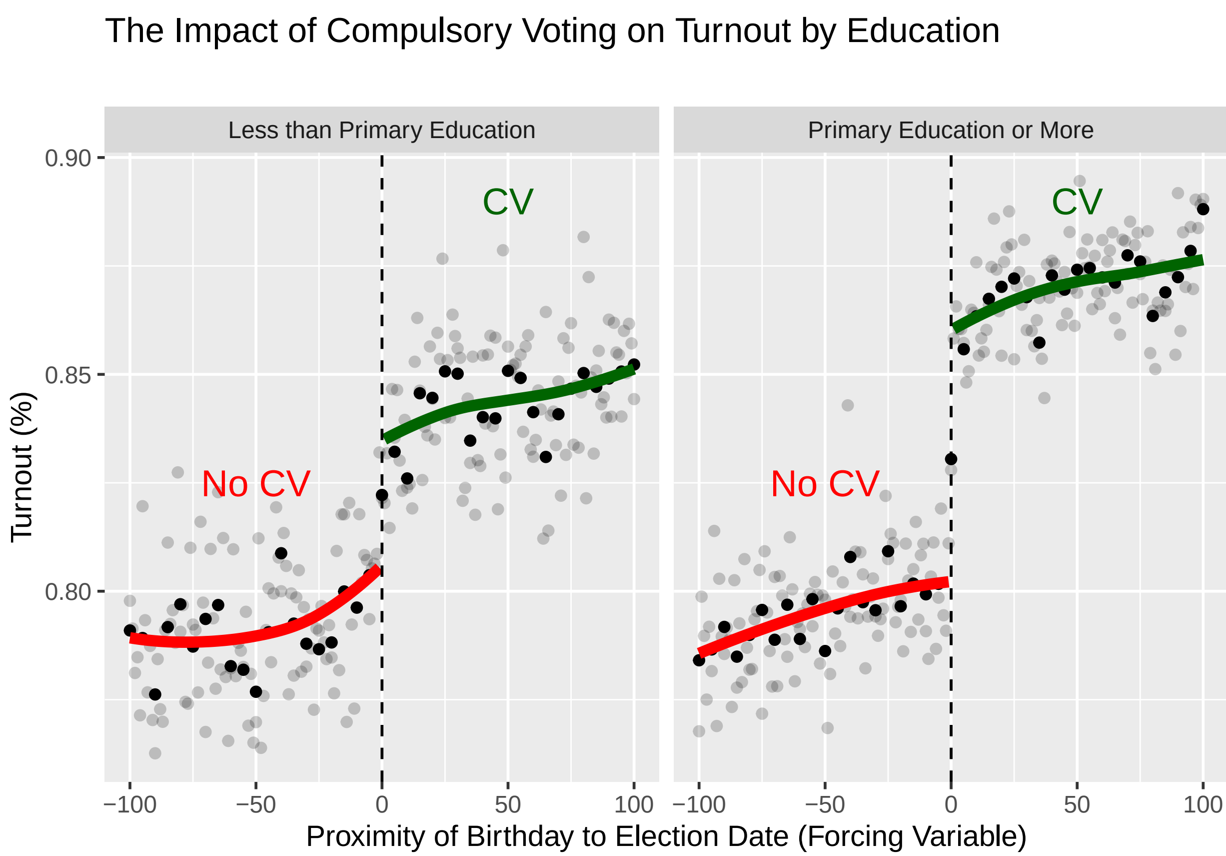
To begin the analysis, I attempted to recreate one of the author's graphs, which showed visually the difference in turnout on either side of the CV threshold. To do so, I summarized the election register data for the age 69-70 group (I began with this group for reasons discussed later) by their date of birth.

The proximity of their date of birth to the CV threshold is known as the 'forcing variable', because depending on its value, an individual has a greater or lesser chance of undergoing treatment. Those with negative values of the forcing variable have their 70th birthday before the election, and thus are not obligated to vote as they fall outside the CV age range, while those with positive values are compelled to vote.

A key assumption of the RD model is that individuals with relatively small values of the forcing variable will tend to be highly similar apart from having the treatment applied or not, so the causal effect can be much more clearly observed. As shown by the graph below, turnout experiences a significant jump when the forcing variable crosses the threshold (the black dashed line, at fv = 0).

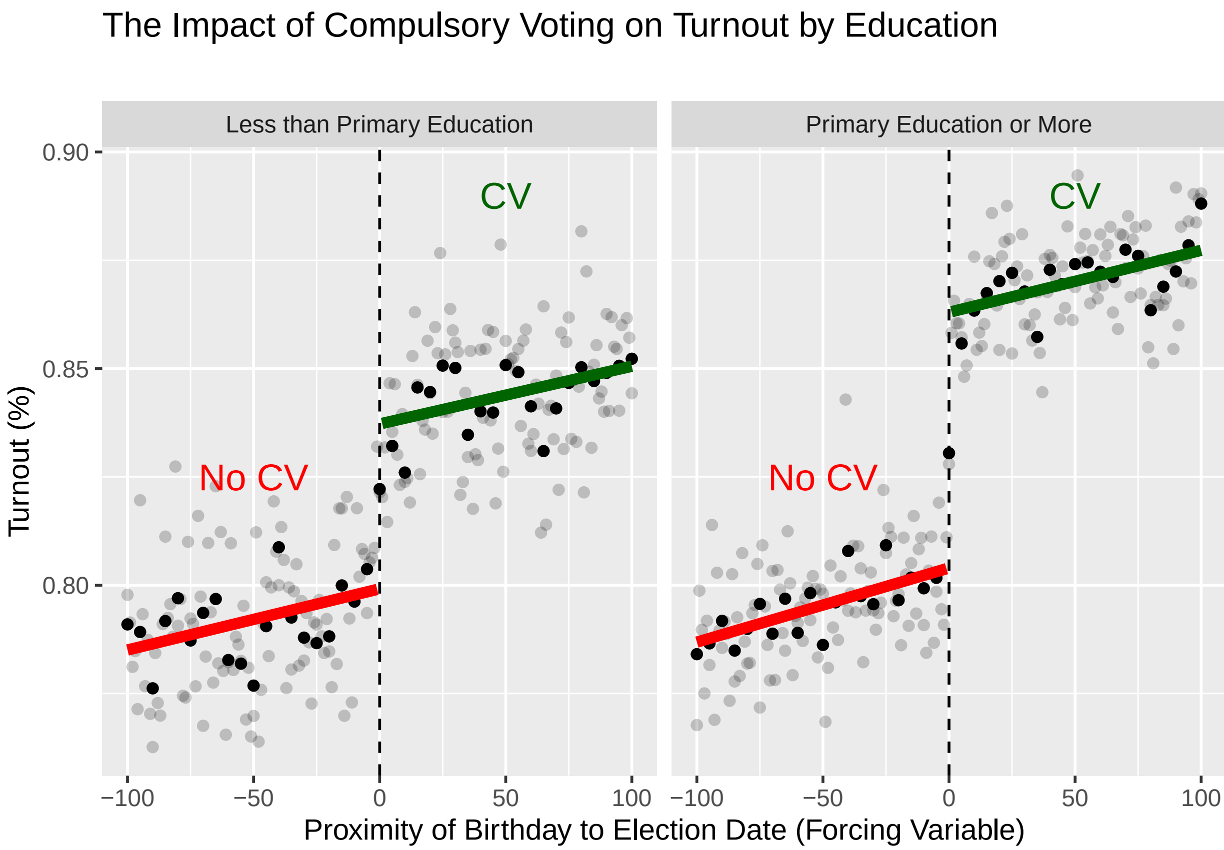
The key trend to observe here, however, is that while CV increases turnout for both the less educated and more educated groups, the effect is much greater for the more educated group. Although both groups start off with a turnout of around 80%, the impact of CV makes highly educated turnout visibly higher than less educated group turnout.

The black dots indicate the average turnout within bins of 5 days, while the lines are loess smoothing models used to visualize the trends of turnout on each side of the threshold. This is just an exploratory graph - in later sections, I quantify the exact size of the difference between the two effects.



Discontinuity Model

For the model I fitted, I used a method closer to that demonstrated by the graph below. For the data on either side of the threshold, I fitted a separate linear regression model, of turnout regressed on the forcing variable. For this model, the intercept reflects the predicted value of turnout when the forcing variable equals 0 (at the threshold). These are the points at which the lines below cross the dashed line. By calculating the difference between the two predicted values, one from the regression on the treated group, and one from the untreated group, one can obtain an estimate of the causal effect of CV on turnout.



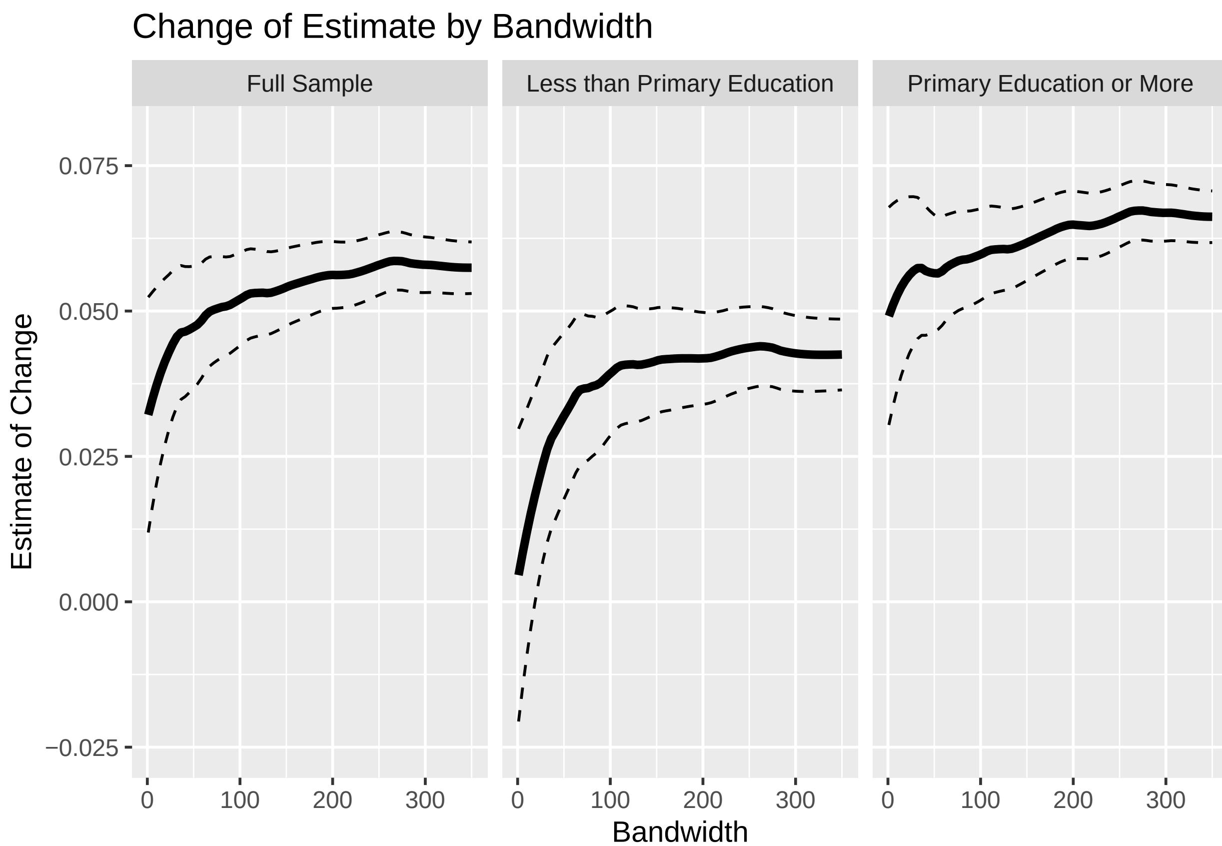
Bandwidth Selection and Graph

When performing an RD analysis, however, another decision that must be made is that of bandwidth. How far away from the threshold do you look, data-wise, before it is no longer entirely indicative of the trends you wish to capture? In this case, the paper's data only contained information for individuals with birthdays up to a year away from the election.

To decide what bandwidth to use, I analyzed how the final estimate of the CV causal effect varied by bandwidth for each group. I created a function to calculate and return the estimate of the CV causal effect, as well as the 95% confidence interval (calculated using the standard errors of the intercept coefficients from the individual models). I then plotted a graph showing how the estimated effect changes as the bandwidth increases, for every bandwidth between 15 days and 365 days. These results are shown in the graph below, with the solid line showing the point estimate and the dashed lines showing the 95% confidence interval.

Several trends are worth noticing. Firstly, as the bandwidth increases, the confidence interval becomes smaller, indicating a rise in certainty as more data points are being used in creating the linear models.

Secondly, the difference in estimated CV effect between more educated and less educated individuals is far more pronounced at lower bandwidth levels. This indicates that the CV effect exerts itself further away from the threshold for less educated groups than for more educated ones - that is, less educated individuals whose birthday falls close to the threshold experience are more likely to turn out to vote than otherwise, even if they are not required to by the policy. The change in turnout for more educated individuals is much more limited to the jump at the threshold, on the other hand.



Estimated CV Effect (Age 69-70 Group)

For the final coefficient, I decided to use a bandwidth of 300 days, as the values appeared to stabilize at around that point, indicating that this is the point at which the threshold ceases to have an effect. The authors of the original study used an algorithm to select a different bandwidth for each group being investigated, so their final coefficients differed from mine.

[Explanation of Final 69-70 Coefficient]

Table 1: CV Effect on Turnout for Age 69-70 Group

Group	Estimate	Confidence.Interval.Min	Confidence.Interval.Max	Std.Error
All	0.058	0.053	0.063	0.0026
More Educated	0.067	0.062	0.072	0.0026
Less Educated	0.040	0.036	0.049	0.0046

Estimated CV Effect (Age 17-18 Group)

For the 17-18 year age group, I also used the bandwidth of 300 days. The coefficients I obtained for this section were significantly different to those of the authors, and with the Estimates for All and Less Educated I was unable to replicate the author's results. This was because the authors had used an additional model to impute the number of those who failed to register, which I was unable to replicate. As a result, the number of less educated individuals was significantly underestimated by my model. This has the result that the Estimate value for the Less Educated group is not statistically significant, and the Estimate value for the Full group is disproportionately influenced by the More Educated group, so its values are very similar to it.

Therefore, no reliable conclusion can be drawn from this group - this is partly due to the fact that the proportion of individuals aged 17-18 who have not completed primary school but are still registered voters is extremely small relative to older age groups, likely reflecting improvement in educational access over time, making primary school more attainable for the 17-18 age group today than it was for those in the 69-70 age group when they were younger.

Table 2: CV Effect on Turnout for Age 17-18 Group

Group	Estimate	Confidence.Interval.Min	Confidence.Interval.Max	Std.Error
All	0.072	0.069	0.075	0.0015
More Educated	0.073	0.070	0.075	0.0010
Less Educated	0.007	-0.016	0.030	0.0117

Extension: Political Inequality by Sex

Table 3: CV Effect on Turnout for Age 69-70 Group

Group	Estimate	Confidence.Interval.Min	Confidence.Interval.Max	Std.Error
All	0.058	0.052	0.064	0.0061
Male	0.066	0.060	0.071	0.0056
Female	0.049	0.044	0.055	0.0056