Weighted Subgroup Analysis in Regression Discontinuity Designs*

Alvaro Carril† Andre Cazor‡ Maria Paula Gerardino§
Stephan Litschig¶ Dina Pomeranz‖

June 29, 2023

Abstract

Subgroup analyses to explore mechanisms and conduct falsification tests are increasingly common in regression discontinuity designs (RDDs). However, existing methods fail to account for other differences across subgroups beyond the characteristic of interest. As a result, evidence of subgroup effect heterogeneity - or lack thereof - is often difficult to interpret. This paper introduces a new approach to conduct regression discontinuity subgroup analysis while holding other observable characteristics constant, based on inverse probability weighting. Observations from subgroup zero with high (low) estimated probability scores to belong to subgroup one are up-weighted (down-weighted). Successful balancing of observables across subgroups helps to isolate effect heterogeneity due to the subgroup characteristic of interest from effect heterogeneity driven by observed confounders. The approach is illustrated with data from two studies that use RDD subgroup analysis to explore a particular mechanism (Fujiwara, 2015) or run a falsification exercise (Solis, 2017).

*We thank Alberto Abadie, Matias D. Cattaneo, Thomas Fujiwara, David Lee, Ben Olken, Alex Solis, Jeffrey M. Wooldridge and Kaspar Wuthrich for helpful commments and feedback.

†Princeton University, acarril@princeton.edu
‡Pennsylvania State University, acazor@uc.cl
§Inter-American Development Bank, mariage@iadb.org
¶National Graduate Institute for Policy Studies, s-litschig@grips.ac.jp
‖University of Zurich, dina.pomeranz@uzh.ch

1 Introduction

Subgroup or effect heterogeneity analysis has become standard practice in regression discontinuity designs (RDDs).[1] Many of those papers conduct a descriptive subgroup analysis, showing RDD results by gender or race for example. A more ambitious objective for subgroup analysis is to explore a specific mechanism or conduct a falsification exercise, arguing that the effect should be smaller, larger or absent altogether for observations with a given characteristic of interest. However, existing methods do not allow researchers to account for other differences across subgroups beyond the characteristic of interest. As a result, evidence of subgroup effect heterogeneity - or lack thereof - is often difficult to interpret.

This paper outlines a new approach to help isolate effect heterogeneity by a given subgroup characteristic of interest from effect heterogeneity by other observed covariates that are correlated with that characteristic. We call it "weighted subgroup analysis". The approach requires two subgroups of interest but is otherwise fairly general. It is likely most useful in (sharp or fuzzy) RD designs since it minimizes functional form assumptions as further discussed below. The basic idea of weighted subgroup analysis is as follows. Suppose a researcher wants to know whether a larger effect of a given intervention for females (subgroup one) compared to males (subgroup zero) is in fact driven by some other pre-intervention attribute, such as age. The goal of weighted subgroup analysis is to construct a weighted subgroup sample of males that matches the female sample as much as possible in terms of age and then run separate RDDs in the weighted male and unweighted female samples. Under the standard continuity assumption, these separate RDDs identify the causal effect of crossing the cutoff for males, and for females, respectively. And to the extent that observable balancing is successful, the difference between male and female subgroup effects reflects the heterogeneity by gender isolated from the heterogeneity by age. This difference provides the contribution of gender to the average effect plus the effect of potential imbalances in unobserved factors.

The approach proceeds in three steps. First, construct a weighted subgroup sample that matches the other subgroup sample as much as possible in terms of mean observables using inverse probability weighting (IPW) (Horvitz and Thompson, 1952). Observations from subgroup zero with high estimated probability scores to belong to subgroup one

[1] Similarly, Hsu and Shen (2019) find that in 2015-2016, 89% percent of papers in the QJE, AER, JPE, REStud, AEJ:Applied and AEJ:Economic Policy that use RDDs report subgroup analyses in their results.

are up-weighted, while subgroup zero observations with low estimated scores are down-weighted. Intuitively, since subgroup zero units with high scores will have covariate values that are similar to subgroup one units, up-weighting them will improve balance. Second, analyze the extent of balance improvement in terms of average absolute standardized differences in covariate means across subgroups. While balance is guaranteed to improve as long as the probability score exhibits some variation, there is no guarantee that the weighted subgroup zero means match the subgroup one covariate means. If observables are still not balanced, go back to the first step and re-estimate the score, adding interaction terms between covariates or higher-order terms. Third, conduct conventional RDD analysis in the weighted subgroup zero and unweighted subgroup one samples, and test for effect heterogeneity across subgroups using a simple t-test.

In randomized and other non-RD designs, the standard approach to investigate effect heterogeneity is based on specifications that include the subgroup of interest indicator and an interaction term of the treatment indicator with that subgroup indicator. This approach is convenient because it easily accommodates additional covariates and their interactions with treatment status, thus holding other observables constant when testing for differential effects for the subgroup of interest. However, the simple treatment-subgroup-interaction approach is not generally valid in the RDD setting, unless the relationship between the outcome and the running variable is exactly the same across subgroups. A valid and common approach to RDD subgroup analysis instead allows for separate slopes and curvature by subgroup (i.e. full interaction of the running variable polynomial with the subgroup indicator). The problem remains, however, that other characteristics may vary systematically across subgroups, thus making it difficult to interpret differential subgroup impacts. Simply including these additional covariates and their interaction with the treatment indicator would easily lead to specification errors and is thus typically avoided in practice. Instead, extant heterogeneity analysis in RDDs implicitly or explicitly (e.g. Becker et al. (2013)) assumes that relevant covariates are uncorrelated with the characteristic of interest at the cutoff. Our approach relaxes this assumption by only requiring that *unobserved* confounders are balanced across subgroups at the cutoff.[2]

A potential approach would be to run separate subgroup-specific RDDs in cells defined by specific covariate combinations and then aggregate the resulting estimates. While theoretically attractive because the subgroup effects are nonparametrically identified for

[2]See (Hsu and Shen, 2019) for alternative tests that study whether an intervention is 1) beneficial for at least some subpopulations, 2) has any impact on at least some subpopulations, and 3) has a heterogeneous impact across subpopulations.

each cell, this approach quickly runs into sample size issues, as well as potential weak instrument problems in fuzzy RDDs (Feir et al., 2016). Weighted subgroup analysis provides a practical solution by balancing covariates across subgroups only on average, rather than within each cell. Another potential approach is to include interactions of the covariates with running variable polynomials, but this has not been formally investigated in the RDD setting as far as we know.[3] Overall, weighted subgroup analysis provides the first practical method for RD designs to account for differences across subgroups beyond the characteristic of interest.

We illustrate weighted subgroup analysis with two examples. In the first, (Fujiwara, 2015) shows that the introduction of electronic voting technology in Brazil had a stronger effect on valid voting in municipalities with high illiteracy rates (a roughly 5 percentage point higher increase than in more literate municipalities). While these results might indicate that mostly less educated voters were effectively enfranchised by the new technology, the literacy rate also correlates with geographic location. The more illiterate populations are concentrated in the Brazilian Northeast. The observed effect heterogeneity could therefore be driven by political or cultural confounders that vary with geography. Put differently, it is unclear whether illiteracy alone drives this effect heterogeneity since the two samples differ so much in terms of location.

We show that inverse probability weighting of the subgroup of municipalities with low illiteracy rates leads to a substantial balance improvement in terms of a more similar share of municipalities from the Brazilian Northeast as in the subgroup of municipality with high illiteracy rates. Specifically, the average absolute standardized difference of proportions belonging to different regions of Brazil is reduced by an order of magnitude (from 0.65 to 0.06). The corresponding weighted RD estimate for the low illiteracy subgroup is reduced by half and produces a differential effect that is about twice as large as without weights (a roughly 10 percentage point differential effect compared to about 5 percentage points without weights). These results strengthen the argument that the introduction of electronic voting technology mostly enfranchised less educated voters, even holding confounders associated with geography constant.

A second study we use to illustrate our approach is (Solis, 2017), who analyzes credit access and college enrollment in Chile exploiting a cutoff in a nationwide entrance exam.

[3] Calonico et al. (2019) aim to clarify the conditions under which adding covariates in the RDD identifies the average effect of treatment at the cutoff. An earlier version of their paper suggests that future work could study how the interaction approach might be used to investigate treatment effect heterogeneity at the cutoff.

The paper finds that access to credit leads to higher immediate college enrollment for preselected candidates who are eligible for student loans. They are 18 percentage points more likely to enroll immediately in college when they are above the threshold score for access to credit. At the same time, there is virtually no effect for nonselected students, i.e. those who are ineligible for the student loan program because their parents are deemed too rich. The author's interpretation of the null result for the nonselected subgroup is that scoring above the college entrance exam cutoff only relaxed a credit constraint for the preselected subgroup, rather than also providing a positive signal about student ability to students themselves or to admissions officers. But an alternative interpretation is that the nonselected subgroup was simply academically weaker and would not have wanted to go to college in any case.

We first document a substantial imbalance between preselected and nonselected students in terms of high school academic performance and other covariates. Weighting completely eliminates the imbalances but the weighted RDD estimates remain unchanged compared to no weights. This suggests that even with the same average high school performance and other characteristics as preselected students, nonselected students' college enrollment did not respond to them scoring above the college admission cutoff. We show that contrary to the Fujiwara (2015) study, in the Solis (2017) study the R-squared improvement with covariates is only marginal, suggesting that only relevant confounders matter for unbiased subgroup effect estimation.

The paper proceeds as follows. Section 2 discusses the limitations of existing approaches to subgroup analysis in RD designs. Section 3 presents the weighted subgroup analysis approach. Section 4 illustrates the approach using the two examples mentioned above (Fujiwara, 2015; Solis, 2017). Finally, Section 5 concludes.

2 Existing approaches to RDD subgroup analysis

The standard approach to investigate effect heterogeneity in randomized studies (and also often used in other quasi-experimental estimation strategies) is to include a dummy for the subgroup of interest and an interaction term of the treatment indicator with that subgroup dummy. This section illustrates that the simple treatment-subgroup-interaction approach is not generally valid in the RDD setting without strong assumptions. For it to be valid, the relationship between the outcome and the running variable would have to be the same across subgroups. And even when the model allows for separate slopes or curvature by subgroup (i.e. full interaction of the running variable polynomial with the subgroup indicator), the problem remains that other characteristics may vary systematically across subgroups.

2.1 RDD with subgroup-treatment interaction term

Consider aiming to estimate the differential impact of a treatment on two subgroups, $G = 0$ and $G = 1$. There are two estimands of interest, corresponding to the RD-gaps in the outcome Y at the cutoff in each of the two subgroups:

$$\lim_{X_i \downarrow 0} \mathrm{E}[Y_i|X_i, G_i = 0] - \lim_{X_i \uparrow 0} \mathrm{E}[Y_i|X_i, G_i = 0] = \alpha_{R0} - \alpha_{L0} = \alpha_0 \tag{1}$$

and

$$\lim_{X_i \downarrow 0} \mathrm{E}[Y_i|X_i, G_i = 1] - \lim_{X_i \uparrow 0} \mathrm{E}[Y_i|X_i, G_i = 1] = \alpha_{R1} - \alpha_{L1} = \alpha_1 \tag{2}$$

where the cutoff is normalized to zero and the notation is adapted from Lee and Lemieux (2010).

Now consider a linear spline specification, augmented only with the subgroup dummy G and an interaction term of the subgroup dummy with treatment assignment ZG, where $Z_i = \mathrm{I}[X_i \geq 0]$. In terms of the parameters above, the model can be written as

$$Y_i = \alpha_{L0} + \alpha_0 Z_i + (\alpha_1 - \alpha_0) Z_i G_i + (\alpha_{L1} - \alpha_{L0}) G_i + \delta_1 X_i + \delta_2 X_i Z_i + \epsilon_i. \tag{3}$$

To see the correspondence between the regression specification in (3) and the parameters of interest (the αs), simply evaluate the regression function at a given point. For example, when X, Z, and G are all zero, the predicted value is the mean of Y in subgroup zero just before crossing the cutoff, i.e. α_{L0}. Just above the cutoff, the mean of Y in subgroup zero is $\alpha_{L0} + \alpha_0 = \alpha_{R0}$, so the coefficient on Z is the discontinuity in Y in subgroup zero. Similarly, when X and Z are zero and G is one, the predicted value is the mean of Y in subgroup one just before crossing the cutoff, i.e. α_{L1}. Just above the cutoff, the mean of Y in subgroup one is $\alpha_{L1} + \alpha_0 + (\alpha_1 - \alpha_0) = \alpha_{R1}$, so the coefficient on ZG is the difference in discontinuities in Y between subgroups one and zero.

Figure 1 illustrates the specification bias that arises when both true RD gaps are zero ($\alpha_{R0} = \alpha_{L0} = \alpha_{R1} = \alpha_{L1}$, i.e. the treatment has no effect in either subgroup) yet the relationship between Y and X is not the same in the two subgroups (i.e. different slope in each subgroup). The two solid lines show the linear approximations to the conditional expectation functions in the two subgroups, $\mathrm{E}[Y_i|X_i, G_i = 0]$ and $\mathrm{E}[Y_i|X_i, G_i = 1]$ within a neighborhood h around the cutoff. The dashed lines represent the slope estimates from equation (3), which are allowed to differ to the left and to the right of the cutoff but are assumed constant across subgroups. Crucially, the slope estimates are necessarily between the true slope parameters because OLS tries to minimize deviations from the regression line across subgroups. Now as long as the slope estimates are biased, the intercept and discontinuity estimates at the cutoff are necessarily biased as well. In Figure 1, the discontinuity estimate for the $G = 1$ subgroup is upward biased and for the $G = 0$ subgroup the discontinuity estimate is downward biased.

2.2 RDD with full subgroup interaction

The specification bias above is easily fixed if the model allows for separate slopes by subgroup (i.e. a linear spline fully interacted with G_i), which yields

$$Y_i = (1 - G_i)\big(\alpha_{L0} + \alpha_0)Z_i + \delta_1 X_i + \delta_2 X_i Z_i\big) + G_i\big(\alpha_{L1} + \alpha_1 Z_i + \delta_3 X_i + \delta_4 X_i Z_i\big) + \epsilon_i. \tag{4}$$

The problem remains, however, that other characteristics may vary systematically across subgroups. Including additional interactions with the treatment indicator easily leads to specification errors. Instead, extant heterogeneity analysis in RDDs implicitly or explicitly (e.g. Becker et al. (2013)) assumes that relevant covariates are uncorrelated with the characteristic of interest at the cutoff. Our approach relaxes this assumption by only requiring that *unobserved* confounders are balanced across subgroups at the cutoff.

3 Weighted RDD subgroup analysis

3.1 Identification

In the sharp RDD, the parameter of interest is the average treatment effect (ATE) at the cutoff, i.e., $E[\beta_i|X_i = 0]$, where $\beta_i = Y_i(1) - Y_i(0)$ is the difference in potential outcomes with and without treatment for unit i. If potential outcomes are continuous in X, the outcome RD-gap identifies this average effect. Similarly, as long as potential outcomes in each subgroup are continuous, each subgroup outcome RD-gap identifies a corresponding subgroup average effect. Using the notation above we have

$$\lim_{X_i \downarrow 0} E[Y_i|X_i, G_i = 0] - \lim_{X_i \uparrow 0} E[Y_i|X_i, G_i = 0] = E[\beta_i|X_i = 0, G_i = 0] \tag{5}$$

and

$$\lim_{X_i \downarrow 0} E[Y_i|X_i, G_i = 1] - \lim_{X_i \uparrow 0} E[Y_i|X_i, G_i = 1] = E[\beta_i|X_i = 0, G_i = 1]. \tag{6}$$

However, the difference in subgroup average effects may be driven at least in part by observed and unobserved factors that are correlated with the subgroup characteristic, rather than by the subgroup characteristic itself. To make this idea precise, consider the standard approach for subgroup analysis in non-RD settings, which is to assume a linear model for β_i, such as

$$\beta_i = E[\beta_i|X_i, G_i, W_i] + U_i = \beta_0 + \beta_X X_i + \beta_G G_i + \beta_W W_i + U_i. \tag{7}$$

In this setup, β_0 is the average effect at the cutoff (when $X_i = 0$) in subgroup zero for observations with $W_i = 0$ and $U_i = 0$, while β_X, β_G and β_W respectively represent the contributions of running variable X, characteristic G and observed confounder W to the average effect. U represents unobserved effect heterogeneity, potentially correlated with G but uncorrelated with W by construction. The setup easily accommodates multiple observables. Simply substituting the equation for β_i into an RDD equation leads to the specification problems discussed in Section 2.

The goal of subgroup analysis is to identify β_G. Figure 2 shows conditional expectation functions $E[W_i|X_i, G_i = 0]$ and $E[W_i|X_i, G_i = 1]$ to illustrates covariate imbalance across subgroups as well as the idea of weighted subgroup analysis. The observable W is continuous in X in each subgroup as required for identification of the corresponding subgroup average effects. However, due to a marked level difference in means of W at the cutoff, it is difficult to attribute any difference in subgroup average effects to the subgroup characteristic alone. The figure also shows the conditional mean function in the weighted subgroup zero where the observable has the same mean at the cutoff as in subgroup one $E\big[W_i|X_i, G_i = 0, E[W_i|X_i = 0, G_i = 1]\big]$. Intuitively, any remaining difference in subgroup average effects cannot be driven by W if the means of W at the cutoff are balanced across subgroups.

In what follows, we formalize these ideas. Consider the two subgroup average effects and their difference:

$$
\begin{aligned}
E[\beta_i|X_i = 0, G_i = 1] &= E[\beta_0 + \beta_X X_i + \beta_G G_i + \beta_W W_i + U_i|X_i = 0, G_i = 1] \\
&= \beta_0 + \beta_G + \beta_W E[W_i|X_i = 0, G_i = 1] + E[U_i|X_i = 0, G_i = 1],
\end{aligned}
$$

$$
\begin{aligned}
E[\beta_i|X_i = 0, G_i = 0] &= E[\beta_0 + \beta_X X_i + \beta_G G_i + \beta_W W_i + U_i|X_i = 0, G_i = 0] \\
&= \beta_0 + \beta_W E[W_i|X_i = 0, G_0 = 0] + E[U_i|X_i = 0, G_i = 0],
\end{aligned}
$$

and

$$
\begin{aligned}
E[\beta_i|X_i = 0, G_i = 1] - E[\beta_i|X_i = 0, G_i = 0] = \beta_G& \\
&+ \beta_W \big[E[W_i|X_i = 0, G_i = 1] - E[W_i|X_i = 0, G_i = 0]\big] \\
&+ E[U_i|X_i = 0, G_i = 1] - E[U_i|X_i = 0, G_i = 0].
\end{aligned}
$$

The last equation makes it clear that the difference in subgroup effects is due not only to the subgroup characteristic, but also due to potential differences in mean observed and unobserved factors at the cutoff. Now consider the weighted subgroup zero average effect:

$$
\begin{aligned}
&E\big[\beta_i|X_i = 0, G_i = 0, E[W_i|X_i = 0, G_i = 1]\big] \\
&\quad = E\big[\beta_0 + \beta_X X_i + \beta_G G_i + \beta_W W_i + U_i|X_i = 0, G_i = 0, E[W_i|X_i = 0, G_i = 1]\big] \\
&\quad = \beta_0 + \beta_W E[W_i|X_i = 0, G_i = 1] + E\big[U_i|X_i = 0, G_i = 0, E[W_i|G_i = 1]\big].
\end{aligned}
$$

Finally, consider the difference between the subgroup one and weighted subgroup zero average effects:

$$
\begin{aligned}
E[\beta_i|X_i = 0, G_i = 1] - E\big[\beta_i|X_i = 0, G_i = 0, E[W_i|X_i = 0, G_i = 1]\big] = \beta_G& \\
&+ \beta_W \big[E[W_i|X_i = 0, G_i = 1] - E[W_i|X_i = 0, G_i = 1]\big] \\
&+ E[U_i|X_i = 0, G_i = 1] - E\big[U_i|X_i = 0, G_i = 0, E[W_i|G_i = 1]\big].
\end{aligned}
$$

To the extent that observables are balanced at the cutoff across subgroups, β_G is identified as the difference between the subgroup one and weighted subgroup zero effects as long as unobservables are also balanced.

3.2 Estimation

Weighted RDD subgroup analysis requires three steps. First, balance subgroups based on propensity score weighting. Second, assess the balance improvement with weights at or around the cutoff. And finally conduct conventional RDD analysis in the weighted $G = 0$ and unweighted $G = 1$ subgroups.

Step 1: Inverse Probability Weighting

Our weighting approach is based on inverse probability weighting (IPW) in the spirit of Horvitz and Thompson (1952). It involves weighting observations from subgroup zero by the inverse of their conditional probabilities to belong to subgroup one given a set of covariates. Observations from subgroup zero with high estimated probability scores to belong to subgroup one are up-weighted, while subgroup zero observations with low estimated scores are down-weighted. Intuitively, since subgroup zero units with high scores will have covariate values that are similar to subgroup one units, up-weighting them will improve balance.

To estimate the probability score, we first restrict the sample to observations close to the cutoff using determined bandwidths. We then follow the standard approach for propensity score weighting. Estimate a logit model in order to calculate a predicted probability to belong to subgroup G_i=1:

$$P(G_i = 1|X_i, W_i) = \frac{e^{h(X_i,W_i)}}{1 + e^{h(X_i,W_i)}} = P(X_i, W_i), \tag{8}$$

where $h(W_i)$ is a starting specification that includes the covariates W_i as linear or interaction terms. Restrict the sample to the common propensity score support and weight observations by the inverse propensity score. Specifically, the weight attached to the i-th observation is

$$G_i + (1 - G_i)\frac{P(X_i, W_i)}{1 - P(X_i, W_i)}, \tag{9}$$

where p is the unconditional probability of belonging to subgroup $G_i = 1$. The weights in (10) implies that observations from subgroup zero with high (low) estimated probability scores are up-weighted (down-weighted).

Step 2: Assessing Covariate Imbalance Reduction

After weighting, we can check whether this process removed imbalances in covariates between the two subgroups of interest by comparing mean differences in the unweighted and weighted samples. To assess the statistical significance, we use a t-test for individual coefficients and an F-test for overall balance, as typically done in balance tables. Thus, t-test for the difference in means at the cutoff for a specific covariate W_{ki} is given by:

$$W_{ki} = (1 - G_i)\big(\theta_{00k} + \theta_{01k}X_i\big) + G_i\big(\theta_{10k} + \theta_{11k}X_i\big) + \epsilon_i \tag{10}$$

Then, we use standardized mean differences (SMD) or economic or substantive balance. This is:

$$SMD_k = \frac{\theta_{10k} - \theta_{00k}}{sd(W_{ki})} \tag{11}$$

To assess average balance across covariates we take a simple average of the SMDs in absolute terms.

For joint significance of all K covariates, we estimate the following equation for the unweighted and weighted samples, respectively:

$$G_i = \gamma_1 W_{1i} + ... + \gamma_K W_{Ki} + \delta X_i + Q_i \tag{12}$$

Finally, we use an F-statistic to test the joint null hypotheses that all γ coefficients are zero.

Step 3: Estimating the RDD by Subgroup

Once we have settled on a propensity score specification that eliminates or strongly reduces the imbalances of observables between $G_i = 0$ and $G_i = 1$, we can proceed to estimate the differential treatment effect. The weighted and unweighted subsample specifications are[4]:

$$\begin{aligned} Y_i = (1 - G_i)\frac{P(W_i)}{(1 - P(W_i))}&\big(\alpha_{L0} + (\alpha_{R0} - \alpha_{L0})Z_i + \delta_1 X_i + \delta_2 X_i Z_i\big) \\ &+ G_i\big(\alpha_{L1} + (\alpha_{R1} - \alpha_{L1})Z_i + \delta_3 X_i + \delta_4 X_i Z_i\big) + \epsilon_i \end{aligned} \tag{13}$$

[4]Weights are rescaled so as to keep the number of observations constant

By estimating model in (14), we test for effect heterogeneity across subgroups using a simple t-test.

Notice that conventional (robust or clustered) standard errors do not capture sampling variability coming from the fact that the propensity score is estimated. To deal with this issue, our Stata command (block) bootstraps standard errors and confidence intervals.

4 Real examples using our approach

In this section, we apply our methodology using two well published papers that use the standard approach to study subgroup analysis. We compare these results with those under our approach.

Electronic voting in Brazil (Fujiwara, 2015)

Fujiwara (2015) studies the introduction of electronic voting technology in Brazilian elections. Prior to 1998 there were only paper ballots in Brazil which was challenging for illiterate adults who represent about 23%. Additionally, another challenge is that State legislator, governor, federal deputy, senator, and president elections are all held on the same day. For the elections prior to 1998, citizens needed to write down the chosen name or number based on written instructions. Consequently, in practice large numbers of votes were invalid (blank or error-ridden).

In 1998 electronic voting (EV) was introduced, which has mainly four characteristics: (1) Guided step by step voting process; (2) Candidates photographs as visual aids; (3) Number-based interface (like a phone keypad); and (4) Error messages for voters about to cast invalid votes. EV was available in municipalities with more than 40,500 registered voters[5].

This paper uses a sharp RDD to estimate the effect of the introduction of electronic voting technology. Figure A2 plots the main outcome of interest (valid votes in state legislature elections, as a share of turnout) against the forcing variable (registered voters in 1996) for three different elections. A clear jump is visible in the 1998 election (in circles). A little over 75% of the votes are valid on the municipalities below the cutoff, and this figure rises to close to 90% as the cutoff is crossed and EV is introduced. The

[5]Figure A1 depicts the paper ballot above. And below is depicted the EV, in particular, the initial screen of the device, and the screen just before a vote can be confirmed.

fact that no discontinuity is visible for the elections held in 1994 (when all municipalities used paper ballots) and 2002 (when EV was completely phased in) provides a falsification test indicating that municipalities "just above" and "just below" the cutoff are indeed valid treatment and control groups.

In the paper, Fujiwara (2015) presents heterogeneous effects to study whether there are differences between low and high illiteracy populations using subgroup analysis under the standard approach. We compare them with those that are obtained using our approach.

Table 1 document a striking geographical imbalance across high and low illiteracy groups. Inverse probability weighting of the low illiteracy subgroup leads to a substantial balance improvement. Table 2 show that unweighted estimates overstate the effect on valid votes out of turnout in low illiteracy populations. The weighted RD estimate for the low illiteracy subgroup is less than half as large as without weights. Thus, weighted results show that the impact of the technology adoption on valid votes found is mainly due to their positive effect in less educated population.

However, we have some drawbacks given the available data. First, limited number of covariates that are not directly related to literacy (income per capita, poverty rate, share of urban population). And second, moderate sample size precludes zooming in on imbalance close to the cutoff.

College access and credit constraints in Chile (Solis, 2017)

Solis (2017) studies whether access to credit explain the gap in schooling attainment between children from richer and poorer families. In Chile, most universities are private and financed by tuition fees. They are high and equivalent to 47 percent of the median family income. In terms of the duration of a average program, they are designed for 5 years but students take an average of 6.5 years to graduate. Thus the cost of studying a career represent large financial burden for the families.

To reduce the financial burden, private and public loans are available for students. Access to private student loans is restricted by minimum income requirements for the parents in formal jobs. Thus, half the families in the country make too little formal income to qualify. Then, students of poor families can not get private loans and so many of them are looking for a job while they are studying. However, non-college jobs pay too little to work and save for college.

In terms of public loans, Chile has introduced two college loan programs that are available to students scoring above a threshold on the national college admission test. Students from the richest quintile of the income distribution are not eligible. The threshold is above 475 points to get a standardized government loan that is roughly equivalent to 950 SAT points. Most of the poorer students get lower than 475 points, so they are not eligible to get a public loans.

Because of the threshold on the national college admission test in order to get a loan, Solis (2017) estimates a RDD to study whether credit access affect to immediate college enrollment. The top panel of Figure A3 shows the effect of loan eligibility on immediate enrollment for pre-selected students. At the eligibility cutoff, where access to loans changes sharply for preselected students (those that are preselected for a loan), we observe that the enrollment rate for barely eligible students is twice the rate for barely ineligible students. On ahother hand, for the non-selected students. the bottom of Figure A3 suggests that there is no jump. These results illustrates the impact of getting a loan on college enrollment.

However, results in Solis (2017) have the drawback that characteristics across pre-selected and selected student could be different (e.g. gender, household size, etc). Then, the differential effect in college enrolment for pre-selected and non-selected students is not because of credit access, but for example pre-selected students in the sample could come from more educated families that allows them to immediately decide to enroll in a college. In this sense, our approach allow us to address this issue.

Table 3 shows substantial imbalance between preselected and nonselected students in terms of high school academic performance and a long list of other covariates. IPW completely eliminates the imbalances. Table 4 shows the RDD results using the weights that make both groups balanced. We notice that R-squared improvement with covariates is only marginal. As a result, weighting does not make any difference for the estimated null effect. Thus, results with weights still support the Solis' interpretation.

5 Conclusions

In this document, we propose a new approach to balanced RD subgroup analysis based on inverse probability weighting (IPW) (Horvitz and Thompson, 1952). This approach requires three steps. First, construct weighted subgroup samples using IP weights. Second, assess the balance improvement. And third, conduct conventional RDD analysis in the weighted subsample(s) and test for effect heterogeneity across subgroups using a simple t-test.

To show the utility of our approach, we replicate two well-published papers that use subgroup analysis. First, for the Fujiwara (2015) example, we show that unweighted estimates overstate the effect in low illiteracy populations. Results with the new approach strengthen the argument that mostly less educated voters were effectively enfranchised by the new technology. Second, for the Solis (2017) example, weighting does not make a difference for the null effect. Results support the zero signaling value of scoring above the cutoff interpretation.

References

Becker, Sascha, Peter Egger, and Maximilian von Ehrlich, "Absorptive Capacity and the Growth and Investment Effects of Regional Transfers: A Regression Discontinuity Design with Heterogeneous Treatment Effects," *American Economic Journal: Economic Policy*, 2013, *5* (4), 29–77.

Calonico, Sebastian, Matias Cattaneo, Max Farrell, and Rocıo Titiunik, "Regression Discontinuity Designs using Covariates," *Review of Economics and Statistics*, 2019, *101* (3), 442–451.

Feir, Donna, Thomas Lemieux, and Vadim Marmer, "Weak Identification in Fuzzy Regression Discontinuity Designs," *Journal of Business & Economic Statistics*, 2016, *34* (2), 185–196.

Fujiwara, Thomas, "Voting Technology, Political Responsiveness, and Infant Health: Evidence from Brazil," *Econometrica*, 2015, *83* (2), 423–464.

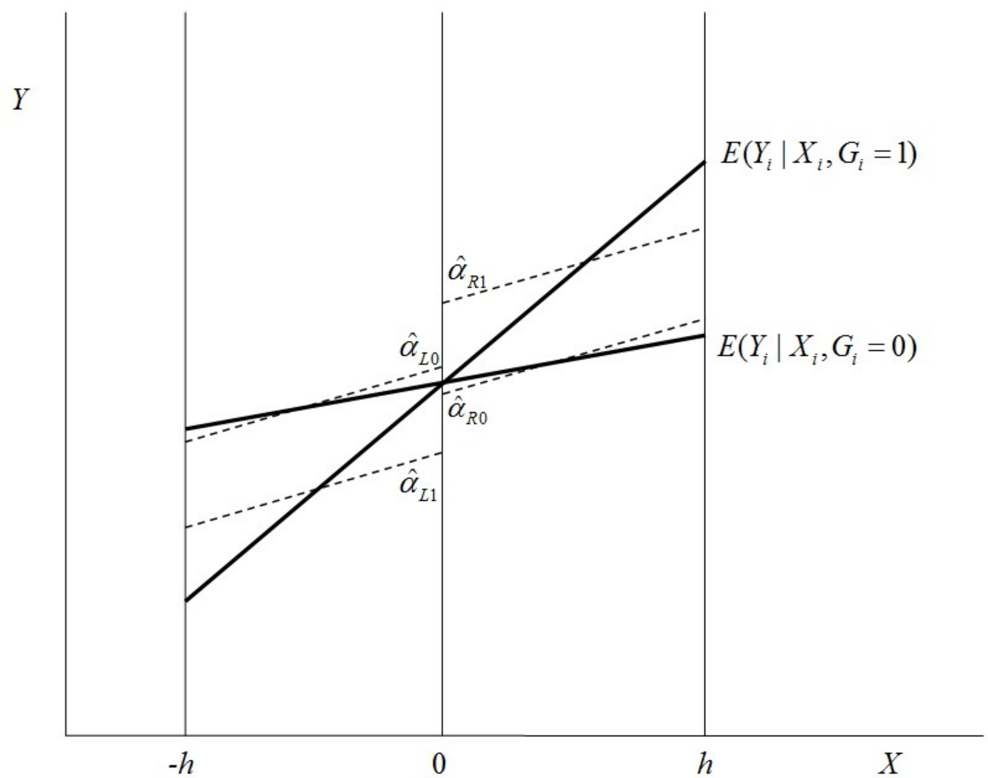
Horvitz, Daniel G. and Donovan J. Thompson, "A Generalization of Sampling without Replacement from a Finite Population," *Journal of the American Statistical Association*, 1952, *47*, 663–685.

Hsu, Yu-Chin and Shu Shen, "Testing for Treatment Effect Heterogeneity in Regression Discontinuity Design," *Journal of Econometrics*, 2019, *208* (2), 468–486.

Lee, David S. and Thomas Lemieux, "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 2010, *48* (2), 281–355.

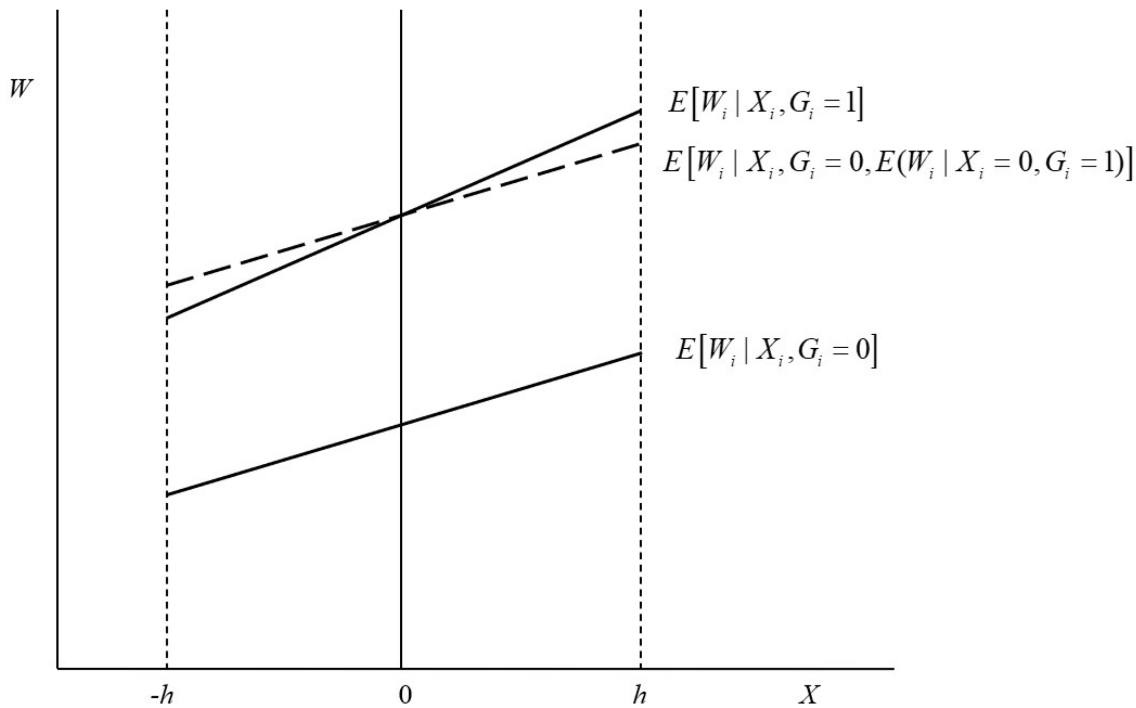
Solis, Alex, "Credit Access and College Enrollment," *Journal of Political Economy*, 2017, *125* (2), 562–622.

Figure 1: Specification bias



Notes: This figure illustrates the specification bias that arises when the true RD gaps in both subgroups are zero and the slope parameters of the running variable and the outcome are different across subgroups, yet the model imposes that the slope parameters are the same.

Figure 2: Covariate imbalance and balancing across subgroups



Notes: This figure illustrates covariate imbalance that arises when the subgroups have differences in the observables and then weighting one subgroup allows to get balancing across subgroups.

Table 1: Balance Improvement for Fujiwara Example

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Original balance				Balance after propensity score-weighted			
	Low illiteracy 16.1% (n=279)	High illiteracy 45.8% (n=279)			Low illiteracy 22.6% (n=279)	High illiteracy 45.8% (n=279)		
	Mean	Mean	St.mean diff.	P-value	Mean	Mean	St.mean diff.	P-value
Central Region	0.065	0.050	0.062	0.467	0.051	0.050	0.003	0.968
North Region	0.043	0.158	-0.381	0.000	0.194	0.158	0.119	0.272
Northeast Region	0.014	0.706	-1.440	0.000	0.665	0.706	-0.086	0.301
South Region	0.337	0.018	0.834	0.000	0.019	0.018	0.003	0.932
Southeast Region	0.541	0.068	1.027	0.000	0.072	0.068	0.008	0.869
population - 1991	49.891	47.472	0.143	0.091	49.92	47.47	0.145	0.112
Abs(St. mean diff.)			0.648				0.061	
F-statistic				252.826				0.879
P-value				0.000				0.494

Notes: Columns (1) and (5) show the mean of each variable for entities with below-median of illiteracy. Columns (2) and (6) show the means for entities with above-median share of illiteracy. Columns (3) and (7) show the standardized mean differences. Columns (4) and (8) show the p-values of t-tests for statistical significance of the difference in means between the two groups.

Table 2: Subgroup analysis by high and low Illiteracy Brazilian Municipalities: The Outcome is valid votes out of turnout

	(1)	(2)	(3)	(4)
	Panel A: Nonweighted			
$1\{S \geq \text{cutoff}\} \times 1\{\text{ Illiteracy} > p(50)\}$	0.148***	0.150***	0.152***	0.176***
	(0.020)	(0.015)	(0.021)	(0.033)
$1\{S \geq \text{cutoff}\} \times 1\{\text{ Illiteracy} < p(50)\}$	0.092***	0.113***	0.095***	0.089***
	(0.021)	(0.016)	(0.022)	(0.033)
Difference Estimate	0.056*	0.037*	0.057*	0.087*
	(0.029)	(0.022)	(0.030)	(0.047)
Observations	265	558	229	116
R-squared	0.484	0.408	0.463	0.432
R-squared Covariates included as Controls	0.608	0.536	0.602	0.621
	Panel B: Propensity score-weighted			
$1\{S \geq \text{cutoff}\} \times 1\{\text{ Illiteracy} > p(50)\}$	0.148***	0.150***	0.152***	0.176***
	(0.020)	(0.015)	(0.021)	(0.033)
$1\{S \geq \text{cutoff}\} \times 1\{\text{ Illiteracy} < p(50)\}$	0.024	0.047	0.020	0.060
	(0.038)	(0.035)	(0.036)	(0.057)
Difference Estimate	0.124***	0.103***	0.132***	0.116*
	(0.043)	(0.038)	(0.042)	(0.066)
Observations	265	558	229	111
R-squared	0.520	0.500	0.525	0.531
R-squared Covariates included as Controls	0.614	0.718	0.619	0.625
Bandwidth	$\pm 11,873$	$\pm 20,000$	$\pm 10,000$	$\pm 5,000$

Notes: Robust standard errors in brackets. Reduced Form estimations. Panel B was obtained using propensity score weighting. Column (1) provides the Imbens and Kalyanaraman (2012) optimal bandwidth. ***p<0.01, **p<0.05, *p<0.1

Table 3: Balance Improvement for Solis Example

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Original balance				Balance after propensity score-weighted			
	Not Preselected (n=6793)	Preselected (n=7645)			Not Preselected (n=5859)	Preselected (n=7645)		
	Mean	Mean	St.mean diff.	P-value	Mean	Mean	St.mean diff.	P-value
Both parents work	0.191	0.142	0.131	0.000	0.163	0.142	0.056	0.001
Mother housewife	0.469	0.514	-0.090	0.000	0.569	0.514	0.110	0.000
Father has formal work	0.594	0.527	0.135	0.000	0.597	0.527	0.141	0.000
Mother has formal work	0.314	0.281	0.073	0.000	0.283	0.281	0.005	0.762
Household sizee	4.476	4.489	-0.007	0.669	4.711	4.489	0.119	0.000
Self-reported income	1.515	1.286	0.392	0.000	1.296	1.286	0.017	0.225
Father college graduate	0.112	0.050	0.227	0.000	0.054	0.050	0.014	0.297
Mother college graduate	0.088	0.045	0.171	0.000	0.047	0.045	0.005	0.708
Father drop high school	0.411	0.473	-0.125	0.000	0.431	0.473	-0.085	0.000
Mother drop high school	0.415	0.461	-0.094	0.000	0.428	0.461	-0.067	0.000
Father years of education	11.52	10.68	0.224	0.000	10.64	10.68	-0.011	0.508
Mother years of education	11.33	10.67	0.187	0.000	10.64	10.67	-0.009	0.597
Female	0.491	0.591	-0.201	0.000	0.587	0.591	-0.008	0.643
Private school	0.077	0.012	0.322	0.000	0.011	0.012	-0.005	0.589
Voucher school	0.525	0.502	0.045	0.007	0.512	0.502	0.019	0.254
Public school	0.394	0.482	-0.177	0.000	0.473	0.482	-0.017	0.323
High school GPA	52.43	55.17	-0.285	0.000	54.81	55.17	-0.038	0.003
Abs(St. mean diff.)			0.170				0.043	
F-statistic			58.92					0.759
P-value				0.000				0.743

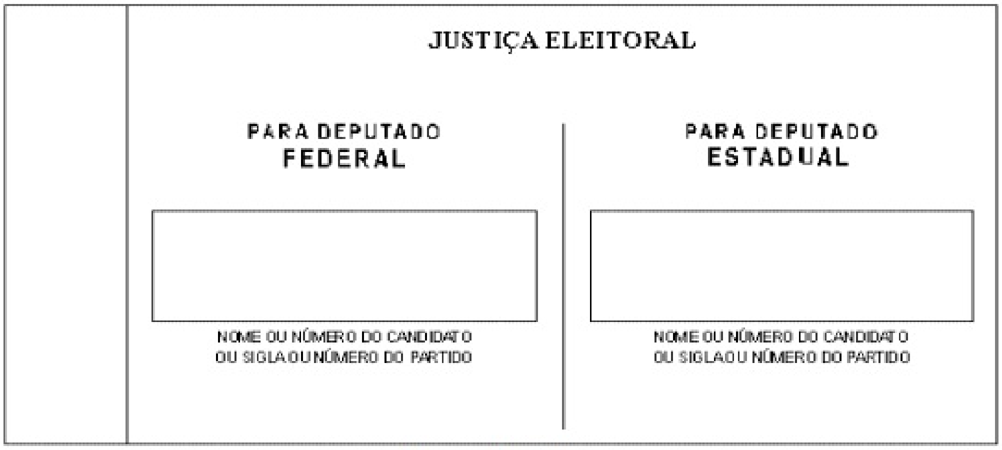
Notes: Columns (1) and (5) show the mean of each variable for entities with below-median of illiteracy. Columns (2) and (6) show the means for entities with above-median share of illiteracy. Columns (3) and (7) show the standardized mean differences. Columns (4) and (8) show the p-values of t-tests for statistical significance of the difference in means between the two groups.

Table 4: Subgroup analysis by preselected and not preselected: The Outcome is immediate college enrollment

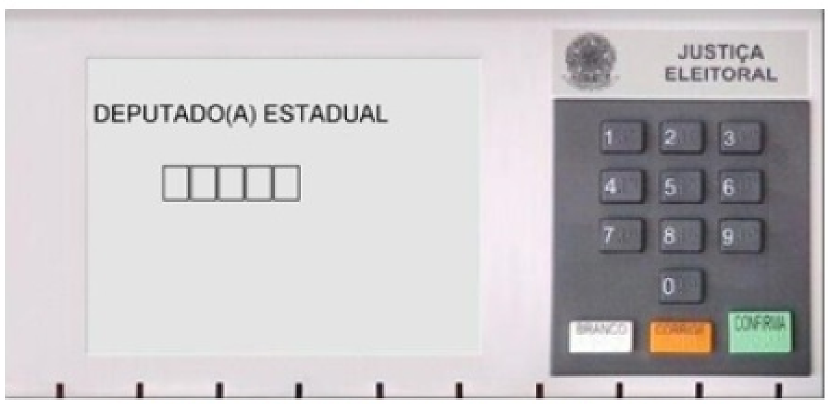
	(1)	(2)	(3)
		Panel A: Nonweighted	
$1\{T \geq \tau\} \times$ Preselected	0.175***	0.170***	0.188***
	(0.006)	(0.006)	(0.020)
$1\{T \geq \tau\} \times$ Not Preselected	0.003	0.007	0.025
	(0.006)	(0.006)	(0.018)
Difference Estimate	0.172***	0.163***	0.164***
	(0.008)	(0.008)	(0.027)
Observations	147638	475165	14438
R-squared	0.102	0.351	0.050
R-squared Covariates included as Controls	0.133	0.358	0.088
		Panel B: Propensity score-weighted	
$1\{T \geq \tau\} \times$ Preselected	0.175***	0.170***	0.188***
	(0.006)	(0.006)	(0.020)
$1\{T \geq \tau\} \times$ Not Preselected	0.004	0.013**	0.005
	(0.006)	(0.006)	(0.018)
Difference Estimate	0.171***	0.157***	0.184***
	(0.008)	(0.009)	(0.027)
Observations	137938	436736	13504
R-squared	0.122	0.354	0.070
R-squared Covariates included as Controls	0.142	0.360	0.097
Bandwidth	±44	All	±4
Spline	Linear	4th-Order Polynomial	Linear

Notes: Robust standard errors in brackets. Reduced Form estimations. Panel B was obtained using propensity score weighting. ***p<0.01, **p<0.05, *p<0.1

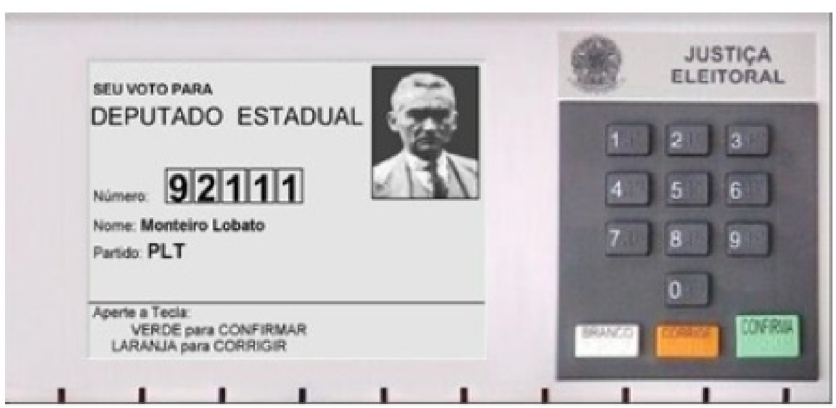
Figure A1: Examples of the voting technologies



Paper ballot

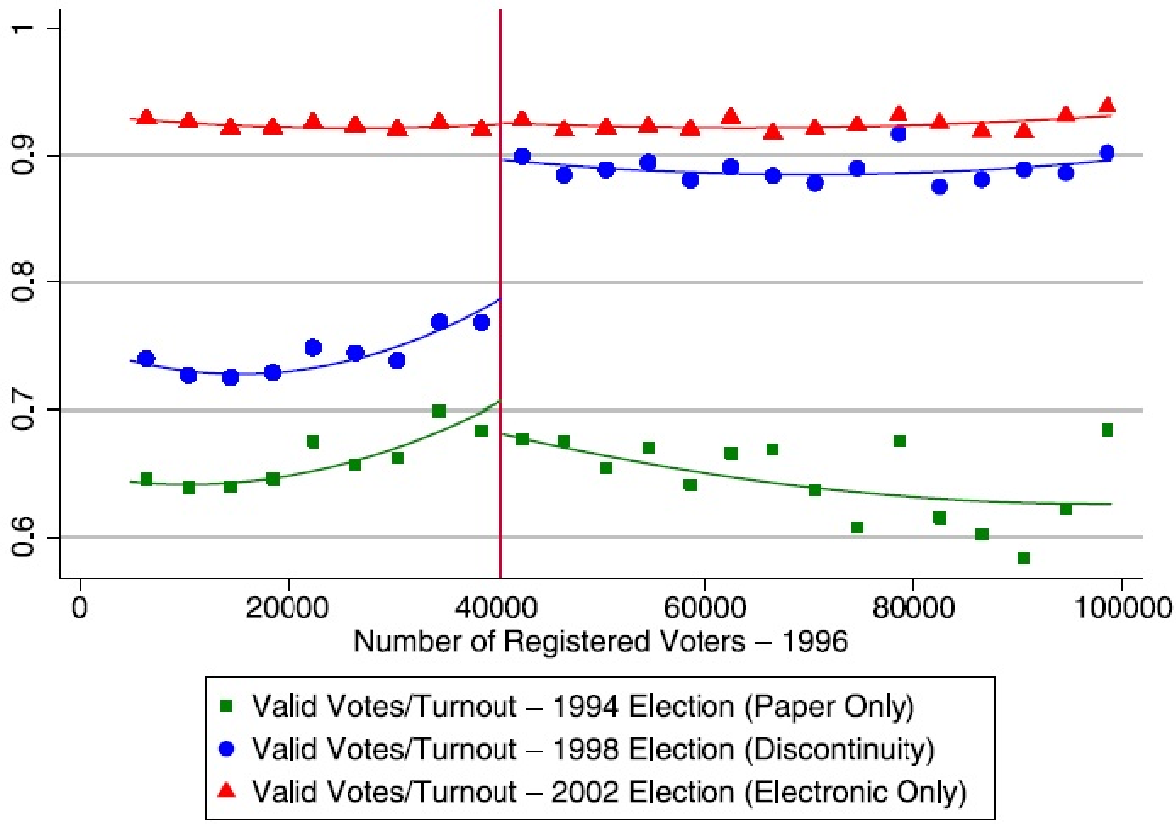


Initial screen of the voting technology



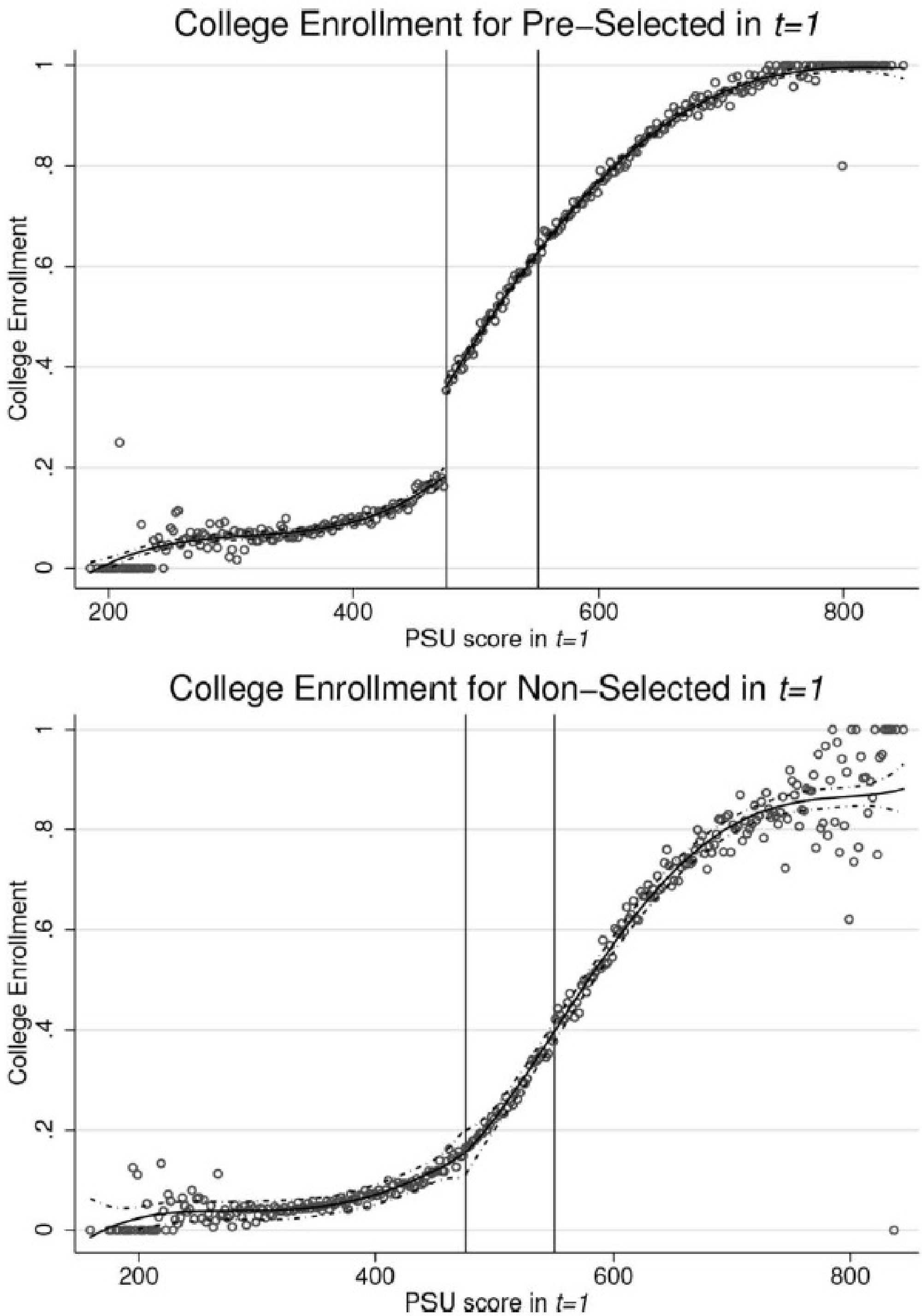
Voting for (fictional) candidate number 92111 (name: Monteiro Lobato, party: PLT)

Figure A2: Valid votes/turnout



Notes: Local averages and parametric fit. Each marker represents the average value of the variable in a 4000-voter bin. The continuous lines are from a quadratic fit over the original ("unbinned") data. The vertical line marks the 40,500-voter threshold.

Figure A3: RD for immediate college enrollment



Notes: Each dot represents average college enrollment within bins of 2 PSU points. The top figure considers PSU first-time takers who applied for benefits and were classified as eligible for loans by the tax authority (preselected students). The bottom figure considers students who did not complete the FUAS or were classified in the richest income quintile (nonselected). Cohorts 2007–9 are pooled together. The vertical lines at 475 and 550 correspond to the loan cutoff and the Bicentenario scholarship, respectively. The dashed lines represent fitted values from the estimation of equation (1) where f() is a fourth-order polynomial at each side of the cutoff and 95 percent confidence intervals.