How Far is Too Far?
Estimation of an Interval for Generalization of a Regression Discontinuity Design Away from the Cutoff*

Magdalena Bennett†

This draft: January 17, 2020

Click **here** for most updated version

Abstract

Regression discontinuity designs are a commonly used approach for causal inference in observational studies. Under mild continuity assumptions, the method provides a robust estimate of the average treatment effect for observations directly at the threshold of assignment. However, it has limited external validity for populations away from the cutoff. This article proposes a strategy to overcome this limitation by identifying a wider interval around the cutoff for estimation using a Generalization of a Regression Discontinuity Design (GRD). In this interval, predictive covariates are used to explain away the relationship between the assignment score and the outcome of interest for the pre-intervention period. Under the partially-testable assumption of conditional time-invariance in absence of the treatment, the generalization bandwidth can be applied to the post-intervention period, allowing for the estimation of average treatment effects for populations away from the cutoff. To illustrate this method, GRD is applied in the context of free higher education in Chile to estimate effects for vulnerable students.

Keywords: Causal Inference; External Validity; Observational Study; Regression Discontinuity Design; Representative Matching

*For comments and suggestions, I thank Peter Bergman, Sebastian Calonico, Sarah Cohodes, Alex Eble, Elizabeth Tipton, Jose Zubizarreta, and participants at the AEFP 2019 Conference and SREE 2019 Spring Conference. All errors are my own.

†PhD Student in Economics and Education, Education Policy and Social Analysis Department, Teachers College at Columbia University, 525 W 120th St., New York, NY 10027; email: mb3863@columbia.edu

1 Introduction

Regression discontinuity designs (RD) are commonly used in observational studies to estimate the effect of policies that are assigned based on a continuous score or running variable. In these settings, the probability of assignment to treatment changes discontinuously at a specific value of the score. Assuming that potential outcomes change smoothly across that threshold, any discontinuity in observed outcomes will be due to the treatment (Lee 2008; Lee & Lemieux 2010). However, one of the main drawbacks of the regression discontinuity approach is its lack of external validity: It only provides estimates for effects that are local to observations right at the cutoff (LATE).

One of the main challenges for estimating effects away from the cutoff is the correlation between the running variable and the outcome. Without taking this correlation into account or making stronger assumptions, it is not possible to identify causal effects beyond the LATE. Naive attempts to model such relationship with flexible functions have shown to be highly sensitive to the choice of model (Manski 2013; Angrist & Rokkanen 2015; Gelman & Imbens 2019), while other generalization methods that attempt to break the linkage between the assignment score and the outcome impose conditions on the entire sample that may fail conditional ignorability assumption tests (Angrist & Rokkanen 2015).

This paper proposes a new method, the Generalization of Regression Discontinuity Design (GRD), to identify a bandwidth for generalizability and estimate average treatment effects for populations away from the cutoff. GRD uses representative template matching (RTM) (Silber et al. 2014; Bennett et al. 2019) leveraging pre-intervention data and predictive covariates to identify an interval in the pre-intervention period where covariates can explain away the relationship between the running variable and the outcome. Under the assumption that conditional potential outcomes in the absence of treatment would be the same across time, the generalization bandwidth can be applied to the post-intervention period to

recover target average treatment effects for the population within that interval. GRD has several advantages: it presents a gradual approach to external validity, relying on a data-driven bandwidth for generalizability, avoids extrapolation by using matching to adjust for covariates, and it generates a matched sample that resembles the observed characteristics of a population of interest.

The GRD method sits in between a regression discontinuity approach and a difference-in-differences (DD) estimation. Unlike a RD design, GRD is able to provide causal estimates away from the cutoff, even for a particular population of interest by adjusting the distribution of covariates that are matched through RTM. In contrast to a DD approach, GRD provides estimates for a subpopulation where identification assumptions are not imposed to the entire sample and do not rely on extrapolation. The combination of RD designs with a DD approach has been widely used in prior literature as a way to enhance a RD (Chicoine 2017; Grembi et al. 2016), but not in the context of a gradual generalization of this method.

This paper relates to the literature on regression discontinuity designs as a form of local randomization (Lee 2008; Lee & Lemieux 2010; Cattaneo et al. 2015; Keele et al. 2015). As shown by Lee (2008), if the running variable is comprised of a random element that cannot be precisely controlled, then the treatment assignment is "as good as random" for a neighborhood around the threshold. Under conditions set by Cattaneo et al. (2015) for identifying such a neighborhood, matching within a narrow bandwidth close to the cutoff would recover the effect of the intervention at the cutoff.

Applying the previous idea of local randomization to a two-period setting, a difference-in-differences approach should be able to reproduce RD estimates, given that groups immediately above and below the cutoff should have very similar outcomes in the pre-intervention period. Then, using a GRD approach, the bandwidth can be gradually increased around the cutoff up to the point where predictive covariates can no longer explain away the relationship between the running variable and the outcome, or balancing restrictions do not hold.

There are two distinct advantages to taking a gradual approach for generalizability using matching. The first one is that even if predictive covariates are not able to break the linkage between the running variable and the outcome for the complete sample, GRD is still able to recover an effect for a subpopulation within a generalization bandwidth. The second one is that there is no need to rely on extrapolation or parametric assumptions if there is not enough overlap between covariates, making estimates local to a specific population but more robust given that assumptions need to hold only for a specific interval.

Following Keele, Small, Hsu, & Fogarty (2019), GRD emulates one of the most "convincing" cases of difference-in-differences to inform the selection of the largest external validity bandwidth, where there is no significant difference between outcomes in the treatment and control group for the pre-intervention period after closely controlling for predictive covariates. The fact that there is no difference in pre-intervention outcomes emulates the idea of local randomization, as treatment and control comparison in the post-intervention period provides similar results to the difference-in-differences estimate. Keele et al. (2019) also propose a useful framework to conduct sensitivity analysis to hidden biases in the context of a difference-in-differences approach, which can also be used for the RTM setting.

The use of predictive covariates to generalize the regression discontinuity design has been used before in the literature. Angrist & Rokkanen (2015) show that under a conditional independence assumption (CIA), the link between the running variable and the outcome can be broken by controlling for predictive confounders. The authors use this approach to estimate the effect of attending selective public schools on inframarginal students. After testing the conditional independence assumption using a regression approach, Angrist & Rokkanen (2015) were able to extrapolate away from the cutoff for only one of their samples.[1] Rokkanen (2015) developed a latent variable model to generalize a regression discontinuity design, building on their previous work. Unlike GRD, however, these methods propose an

[1]Their second sample, using 7th grade test scores, failed the residual test and the authors could not generalize beyond the traditional RD estimate.

"all or nothing" approach, relying on the idea that predictive covariates or a latent factor can break the linkage between the running variable and the outcome across the entire sample of analysis.[2] GRD, on the other hand, avoids this stringent condition on the whole sample by using a gradual approach to generalization and makes explicit the population that it can be generalized for.

This paper also relates to the literature of *comparative RD* design, where an external sample that was not subject to the treatment at any level of the running variable is used as a counterfactual (Wing & Cook 2013; Wing & Bello-Gomez 2018). For instance, Wing & Cook (2013) combine both a difference-in-differences approach and a regression discontinuity design to extrapolate away from the cutoff under the assumption of time-invariant effect of the assignment variable. GRD also leverages pre-intervention data as an external sample to inform the bandwidth for generalizibility; however, by using a matching approach to balance covariates directly, it does not rely on extrapolation and maintains the units of analysis intact. Additionally, GRD only assumes conditional time invariance for potential outcomes in the absence of treatment *within* the generalization interval.

Additional methods have been proposed to generalize the RD to other populations such as Cattaneo et al. (2018) and Bertanha & Imbens (2019). However, given that they apply to different contexts, such as fuzzy regression discontinuity designs or RD with multiple cutoffs, these methods are outside the scope of this paper.

To illustrate the use of GRD, I apply the method to the introduction of free higher education in Chile at the end of 2015, where students in the bottom 50% of the income distribution were eligible to receive this benefit. This example lends itself nicely to a regression discontinuity approach and, given the availability of additional pre-intervention data, also allows for the demonstration of how a GRD approach would work.

[2]Angrist & Rokkanen (2015) use a bandwidth around the cutoff for estimating effects on infra-marginal students to avoid bias of changing counterfactuals, but the bandwidth is not chosen in relation to the interval where the CIA holds.

The structure of this paper is as following. Section 2 describes the framework for the generalization of a regression discontinuity design away from the cutoff, as well as the implementation of the method in practice. Section 3 shows the performance of GRD using simulated data for different scenarios. Section 4 shows the application of the methodology in the context of free higher education in Chile. Finally, section 5 concludes.

2 Generalization of regression discontinuity design

2.1 Framework

Consider a two-period setup, where $t = 0$ refers to a pre-intervention and $t = 1$ a post-intervention period.[3] The intervention itself is noted by $D_{it} = 1$ for those individuals i in period t in the treatment group and $D_{it} = 0$ for individuals in the control group. In a sharp regression discontinuity (RD) design, the intervention is assigned based on a running variable R_{it}, where all individuals i in period 1 are treated if their running variable lies above or below a threshold of eligibility (for illustration purposes $D_{i1} = \mathbb{1}(R_{i1} < c)$); otherwise, they are assigned to the control group. Each individual i in period t is also associated to a set of predictive observable covariates, $\mathbf{X}_{it}$, a set of unobserved confounders $\mathbf{u}_{it}$, and an observed outcome, Y_{it}. The estimand of interest in this case will be the Average Treatment Effect on the Treated (ATT), for the post-intervention period $t = 1$.

Let $Y_{it}^{(0)}$ and $Y_{it}^{(1)}$ be the potential outcomes for unit i in period t, where $Y_{it}^{(0)}$ represents the potential outcome under control and $Y_{it}^{(1)}$ the potential outcome under treatment. Under an additive treatment effect, potential outcomes for unit i in period t under treatment z can be expressed as a function of observed covariates $\mathbf{X}$, unobserved confounders $\mathbf{u}$, and the running variable R, as following:

$$Y_{it}^{(z)} = g(\mathbf{X}_{it}, \mathbf{u}_{it}, R_{it}) + z_{it} \cdot \tau(\mathbf{X}_{it}, \mathbf{u}_{it}, r_{it}) + \alpha_t + \varepsilon_{it}$$

[3] Data in this case can be either panel or cross-sectional data.

where g is an unknown function and ε_{it} is a disturbance term with mean 0.

The observed outcome and the potential outcomes are related as following:

$$Y_{it} = D_{it}Y_{it}^{(1)} + (1 - D_{it})Y_{it}^{(0)}$$

Given that for $t = 0$, $D_{i0} = 0\ \forall i$, we know that $Y_{i0} = Y_{i0}^{(0)}$. For a traditional RD setting, assuming continuity of all observed and unobserved variables across the cutoff, as well as the potential outcomes, I know that for the pre-intervention period:

$$\lim_{R \to c^-} \mathbb{E}[Y_{i0}] = \lim_{R \to c^+} \mathbb{E}[Y_{i0}]$$

so there is no effect at the cutoff.

As previously stated, let the potential outcome under control for the pre-intervention period be a function of the running variable R, observable covariates $\mathbf{X}$, and a set of unobservable confounders $\mathbf{u}$:

$$Y_{i0}^{(0)} = g(\mathbf{X}_{i0}, \mathbf{u}_{i0}, R_{i0}) + \alpha_0 + \varepsilon_{i0}$$

Then, the expectation of the potential outcome conditional on R can be expressed as following:

$$Y_0^{(0)}(R) = \mathbb{E}[Y_{i0}^{(0)}|R] = \mu(R) + \alpha_0$$

where the conditional expectation of the potential outcome under control (and, by extension, the observed outcome) for the pre-intervention period depends on the running variable R.

On the other hand, the conditional expectation of the potential outcomes under treatment can be expressed as following:

$$Y_0^{(1)}(R) = \mathbb{E}[Y_{i0}^{(1)}|R] = \mu(R) + \tau(R) + \alpha_0$$

where $\tau(R)$ is the causal effect of the intervention as a function of the running variable R.

To find the generalization bandwidth, assume that for an interval H^* around the cutoff c, the running variable R can be expressed as:

$$R_{i0} = h(\mathbf{X}_{i0}) + \eta_{i0} \quad \forall \; R_{i0} \in H^* \tag{1}$$

Equation (1) shows that for an interval H^*, the running variable can be fully determined by observed covariates $\mathbf{X}$ in addition to a disturbance term, η, where $\mathbf{X} \perp \eta$ and $\mathbf{u} \perp \eta$.

Then, if the interval H^* exists, conditioning on a specific set of covariates $\mathbf{X} = \mathbf{X}_T$ for the pre-intervention period:

$$Y_0^{(0)}(R)|\mathbf{X}_T, R \in H^* = \mathbb{E}[g(\mathbf{X}_T, \mathbf{u}_{it}, R_{it})|R, R \in H^*] + \alpha_0$$

$$= \mathbb{E}[g(\mathbf{X}_T, \mathbf{u}_{it}, h(\mathbf{X}_T) + \eta_{it}))|R] + \alpha_0 = \mathbb{E}[g(\mathbf{X}_T, \mathbf{u}_{it}, h(\mathbf{X}_T) + \eta_i))] + \alpha_0 = \mu_{\mathbf{X}_T, \mathbf{u}} + \alpha_0$$

Thus, for any $R_1, R_2 \in H^*$,

$$\mathbb{E}[Y_0^{(0)}(R_1)|\mathbf{X}_T] = \mathbb{E}[Y_0^{(0)}(R_2)|\mathbf{X}_T] \tag{2}$$

Equation (2) shows that as long as there is no systematic confounding within the

generalization bandwidth *after controlling for covariates* $\mathbf{X}$, H^* can be identified for the pre-intervention period. In fact, the width of H^* will be fully determined by the widest interval for which equation (2) holds.

Figure 1 illustrate the original setup for the two periods, showing the curves for conditional expectation of potential outcomes $Y_t^{(0)}(R) = \mathbb{E}[Y_{it}^{(0)}|R]$ and $Y_t^{(1)}(R) = \mathbb{E}[Y_{it}^{(1)}|R]$. For the pre-intervention period only the potential outcome under control $Y_{i0} = Y_{i0}^{(0)}$ is observed (Figure 1a), while for the post-intervention period the potential outcome under treatment $Y_{i1}|R_i < c = Y_{i1}^{(1)}$ is observed for individuals under the cutoff c, and the potential outcome under control $Y_{i1}|R_i \geq c = Y_{i1}^{(0)}$ for units above the cutoff.

The generalization bandwidth will be determined by the widest interval H around the cutoff for which condition (2) holds. Then, for each interval $H = [H_-, H_+] \in \mathcal{H}$, where $\mathcal{H}$ is the set of all possible intervals that contain the cutoff c, a potential solution $S \in \mathcal{S}$ will be given by:

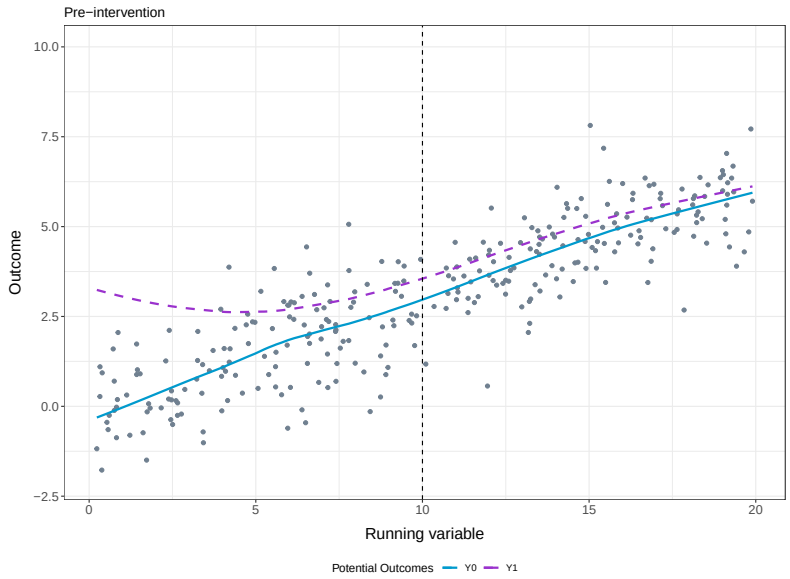
$$S: \ H \in \mathcal{S} \text{ iff } \ \mathbb{E}[Y_{i0}|\mathbf{X}_H, R_i \in H_{c-}] = \mathbb{E}[Y_{i0}|\mathbf{X}_H, R_i \in H_{c+}] \tag{3}$$

Where $H_{c-} = [H_-, c)$ and $H_{c+} = [c, H_+]$ represent the sub-interval of H below and above the cutoff c, respectively, and $\mathbf{X}_{H+}$ is the distribution of the observed covariates within interval H_{c+} (treated units). Then, the generalization bandwidth will be the interval H within the solution set $\mathcal{S}$ with the largest length:

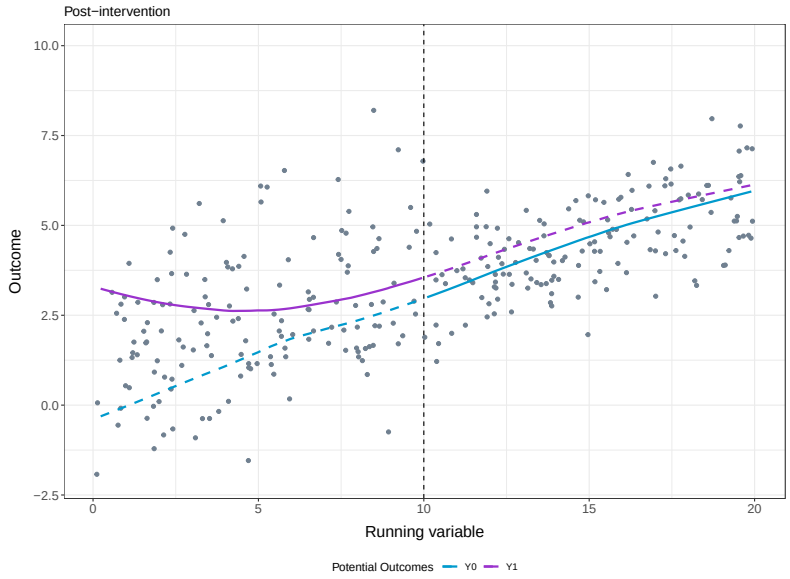
$$H^* = \max\{|H| \mid H \in \mathcal{S}\}$$

Once the generalization bandwidth H^* is identified for $t = 0$, and under the following assumptions:

$$Y_0^{(0)}(R|\mathbf{X}) = Y_1^{(0)}(R|\mathbf{X}) + (\alpha_1 - \alpha_0) \ \ \forall \ R \in H^* \tag{4}$$



(a) Pre-intervention period setup with potential outcomes



(b) Post-intervention period setup with potential outcomes

Figure 1: Setup for the GRD method in the pre- and post-intervention period with potential outcomes

$$Y_1^{(1)}(R)|\mathbf{X}\perp\mathbf{u} \quad \forall\ R \in H^* \tag{5}$$

the generalization bandwidth H^* can be applied to the post-intervention period $t = 1$.

Assumption (4) refers to the idea that absent the treatment, the conditional expectation of potential outcome under control for the pre-intervention period and post-intervention period would have the same trend *within the generalization bandwidth*. This implies that the distribution $f_{\mathbf{u}|R}$ of unobserved covariates does not depend of the period t within interval H^*, which can be partially tested for the portion of H^* in the post-intervention period that covers the control group (H^*_- or H^*_+, depending on treatment assignment). Assumption (5), on the other hand, shows that to be able to obtain a causal estimate within the interval H^*, the potential outcomes under treatment *conditional on observed covariates* have to be independent of unobserved confounders, so $Y_1^{(1)}(R)|\mathbf{X}$ will depend on R only through the potential heterogeneity of the treatment effect.

Figure **??** illustrates how, under assumption (4) and after controlling for predictive covariates (e.g. through matching), the conditional expectation of potential outcomes under control for observations with the same covariates, $Y^{(0)}|\mathbf{X}$, is constant within H^*.

Finally, the estimate of interest, the average treatment effect on the treated (ATT) within the generalization bandwidth H^* , can be determined as following:

$$ATT = \mathbb{E}[Y_{i1}^{(1)} - Y_{i1}^{(0)}|R \in H^*_{c-}, \mathbf{X}]$$

Given (3) and (4), the average treatment effect on the treated can be re-written as following:

$$ATT = \mathbb{E}[Y_{i1}|R \in H^*_{c-}, \mathbf{X}] - \mathbb{E}[Y_{i1}|R \in H^*_{c+}, \mathbf{X}]$$

The next sub-section shows a practical implementation of GRD using a representative matching approach.

2.2 Generalizing a regression discontinuity in practice

In order to implement the previous method, a representative template matching approach (RTM) is used based on the original template matching procedure proposed by Silber et al. (2014) and extended by Bennett et al. (2019). Similar to other matching methods, representative template matching has several properties that are appealing. Unlike commonly used parametric adjustment procedures, matching allows us to compare observations that are alike while maintaining the units of analysis intact (Rosenbaum & Silber 2001). In that same line, matching does not need to rely on the imposition of functional forms for adjustments, reducing the potential risk of bias due to parametric structures (Imbens 2015). Finally, matching has the advantage of separating the adjustment procedure from the outcomes, preventing potential manipulation (Rubin 2008).

However, unlike other matching procedures, RTM allows matching of multiple groups under the same balancing restrictions. Additionally, RTM matches all groups to a sample of a target population of interest, facilitating the estimatation of a Target Average Treatment Effect (TATE). In this case, RTM is implemented using a mixed integer programming approach which allows for the direct balance of covariates, while at the same time substantially reducing computing time of the optimization process even in large samples (see Bennett et al. (2019) for a thorough implementation of representative template matching).

Figure 2 shows a diagram of the template matching methodology and its different stages. Figure 2a shows traditional bipartite matching, were there is (usually) a smaller treatment group and a larger control group. In this setting, the goal is to match treatment units to control units that are similar in observable characteristics, represented in this case by the shapes of the different elements. Figure 2b represents the case of template matching when there are only two groups, a treatment and a control group. In this case, a representative sample from the target population is selected (i.e. template) and then matched independently

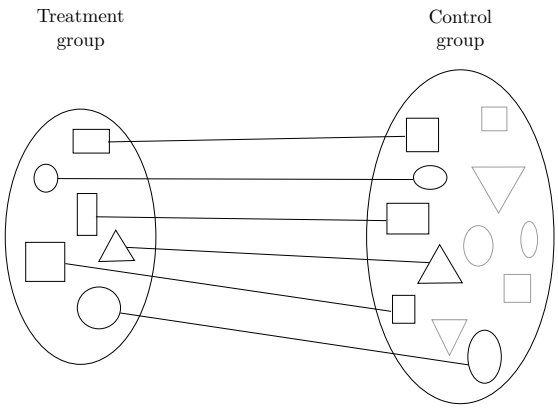
with the treatment and the control group. In this way, by construction, if each matched group is balanced with the template, both matched groups will also be balanced with each other. Finally, Figure 2c shows the extension of template matching with multiple doses or groups.

Using the idea of representative template matching, the steps for selecting the largest generalizability bandwidth are outlined below. Figure 3 illustrate the set up for this method with simulated data, and 4 shows the different stages of the procedure. The main idea of the GRD algorithm is to start with a template sample representative of a population very close to the threshold, and update it in each iteration with a representative sample of a population within a broader interval. This process is repeated until there are significant differences between the outcome within such interval.

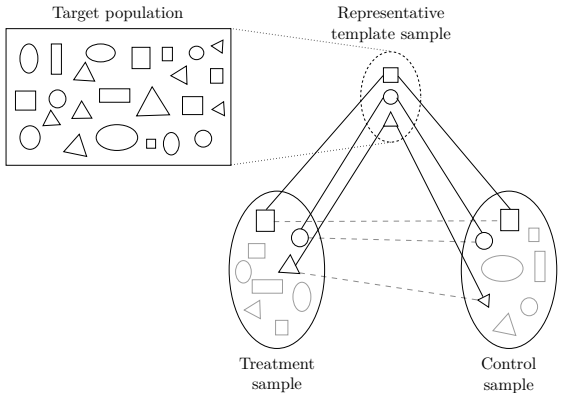
Assuming a treatment assignment based on $D_{it} = \mathbb{1}(R_{it} < c)$, and a set of intervals around the cutoff defined by H_j (Figure 4 (a)), with $j = 1, 2, 3...$, the step are as follows:

1. **Template selection**: For a bandwidth H_j, a template T_j of size N is selected out of S samples of the same size. The template is selected using Mahalanobis distance to choose the random sample S that most closely resembles the target population within H_j. The size N of the template depends on the number of observations within the narrowest bandwidth $H_1 < c$ (e.g. $N = n_1^c$, where n_1^c is the number of observations in H_1 under the cutoff c), and it is maintained through out the entire process to control for minimum detectable effects. The bandwidth H_1, on the other hand, is chosen based on Cattaneo et al. (2015), which identifies a narrow bandwidth around the cutoff where local randomization can be assumed under a set of assumptions (Figure 4 (b)).

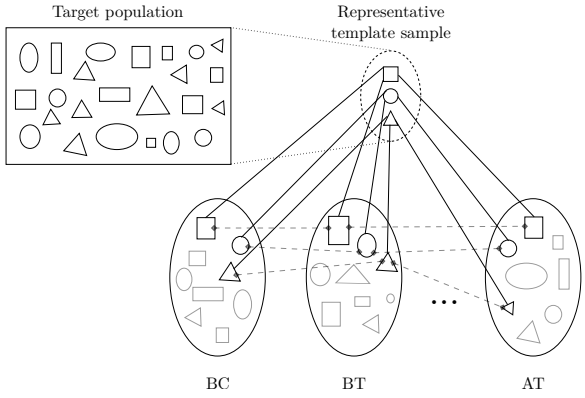
2. **Balance constraints**: To explain away the relationship between the running variable and the outcome of interest, a set of predictive covariates is chosen. For these covariates, balance constraints $\mathcal{B}$ can be set to restricted mean differences, fine balance, or



(a) Traditional matching



(b) Template matching for treatment and control groups



(c) Template matching for multiple difference-in-differences groups

Figure 2: Template matching diagram

exact matching, among many others. More predictive covariates should be matched more closely (e.g. fine balance) than less predictive covariates (e.g. restricted mean difference) (Pimentel et al. 2015). The balancing restrictions are maintained throughout

the different intervals H_j.

3. **Grid setup**: In order to obtain the estimated conditional potential outcome $Y^{(0)}(R)$, a grid G needs to be setup for matching, dividing the running variable R in sub-intervals. For consistency, the grid can be formed by quantiles depending on the total size of the sample and the size of the template.[4] Dividing the running variable into grid intervals (bins) of equal sample size allows for consistency in the matching procedure and proper identification of overlapping characteristics. Let $G_j^{(1)}$ be the bandwidth for treated units between G_j and the cutoff, where G_1 is the closest to the cutoff.

4. **Matching and Re-Matching**: Once the balancing restrictions are set, and the template T_j is selected for the treatment population within bandwidth H_j, each interval of the grid is matched to the template T_j (Figure 4 (c)).[5] Due to potential lack of overlap, matching may not be feasible for the entire grid, especially on its extremes. A feasibility bandwidth $\mathcal{F}_j$ is identified as the largest continuous interval within the grid for which all observations can be matched to the template T_j (Figure 4 (c))).

5. **Outcome assessment for pre-intervention period**: Using the previously matched samples within $\mathcal{F}_j$, a local polynomial regression estimator is fitted for the outcome of interest over the running variable, $f(R, T_j)$[6] (Figure 4 (d)), with a $(1-\alpha)\%$ CI of $[f(R, T_j)_l, f(R, T_j)_h]$. By comparing the estimated function to the level of the outcome at the cutoff score $f(c, T_j) = f_{c-j}$, a local generalization bandwidth H_g^j is identified for template T_j if $|H_g^j| > |H_j|$, such that:[7]

$$H_g^j = [\max\{f(R, T_j)_h < f_{c-j}\}, \min\{f(R, T_j)_l > f_{c-j}\}]$$

[4]For a sample size of $10,000$ for $t = 0$ and a template size $N = 500$, the grid could be set to be deciles of the running variable, allowing for matching of 2:1 between each grid interval and the template.

[5]Due to computational advantages, the use of `cardmatch()` in the `R designmatch` package (Zubizarreta et al. 2018) is highly recommended for this stage.

[6]Other methods can be used to estimate the relationship between the running variable and the outcome of interest, but in this case, the function `lprobust` from the `nprobust` (Calonico et al. 2019) was used.

[7]Assuming a positive relationship between Y and R.

The process is repeated for $H_{j+1} = G^{(1)}_{j+1}$, until $|H^j_g| \subset |H_j|$.

If $|H^j_g| \subset |H_j|$, then the generalization bandwidth is

$$H^*(\alpha, \mathcal{B}) = H^{j-1}_g$$

.

6. **Estimation of the Average Treatment Effect on the Treated (ATT)**: Finally, using the bandwidth H^* found in the previous step, the template T^* is matched and re-matched to both the treatment and control group within the generalization bandwidth for the post intervention period (Figure 5), including a new balancing constraint that minimizes the distance to the cutoff for the control group. This restriction does not change the balance on matching covariates, but prioritizes units from the control group that are closer to the treatment group in terms of the running variable. Given that there is additional uncertainty introduced by the estimation of the generalization bandwidth in the pre-intervention period, in practice a matched difference-in-differences estimator is used, where units from the pre and post-intervention period are matched to the same units of template. Then, the average treatment effect on the treated τ_{ATT} is estimated as following:

$$\hat{\tau}_{ATT} = \sum_{k=1}^{N} \frac{Y_{k(1)1} - Y_{k(0)1} - (Y_{k(1)0} - Y_{k(0)0})}{N}$$

Where $k(z)t$ represent a unit from matched group $k = 1, ..., N$ belonging to treatment assignment $z = 0, 1$, from period $t = 0, 1$. Variance of the estimator using a paired t-test can be obtained as following:

$$Var(\hat{\tau}_{ATT}) = \sum_{k=1}^{N} \frac{(d_k - \hat{\tau}_{ATT})^2}{N-1}$$

Where $d_k = Y_{k(1)1} - Y_{k(0)1} - (Y_{k(1)0} - Y_{k(0)0})$ is the difference between treatment and control for the post minus the pre-intervention period for each matched pair k.

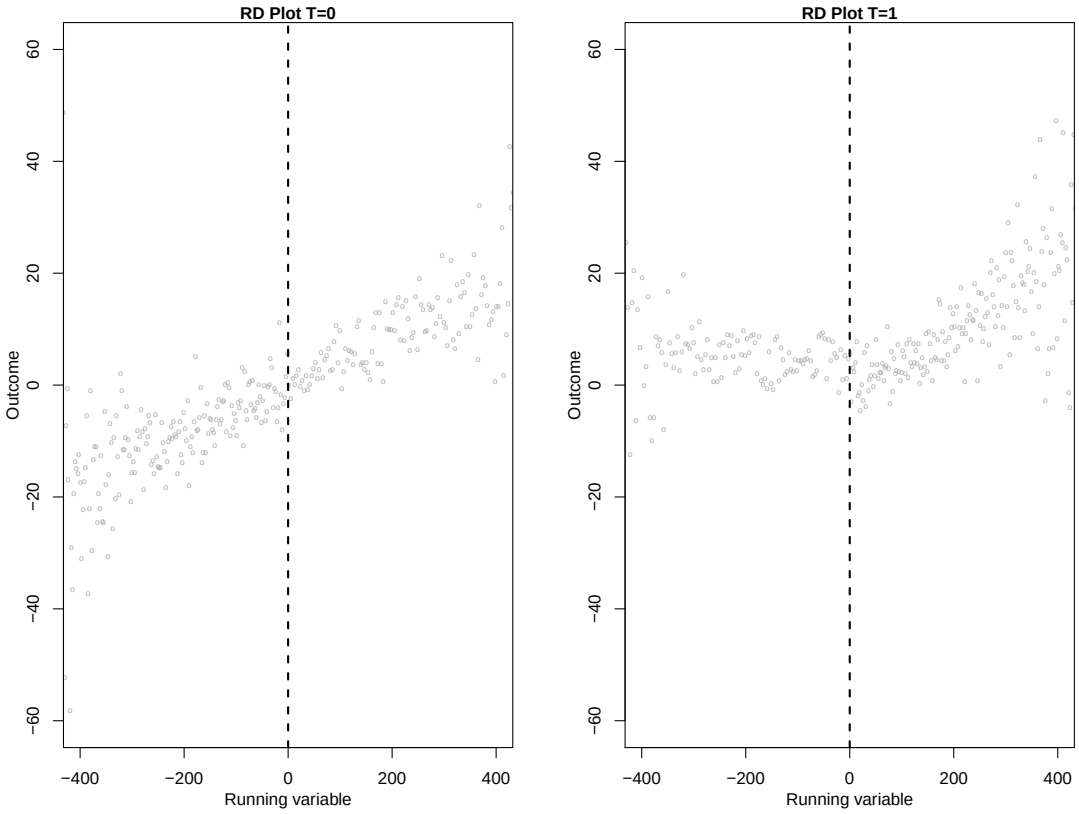


Figure 3: Mean outcome by running variable bins in original data setup for GRD procedure for pre-intervention ($T = 0$) and post-intervention ($T = 1$) periods

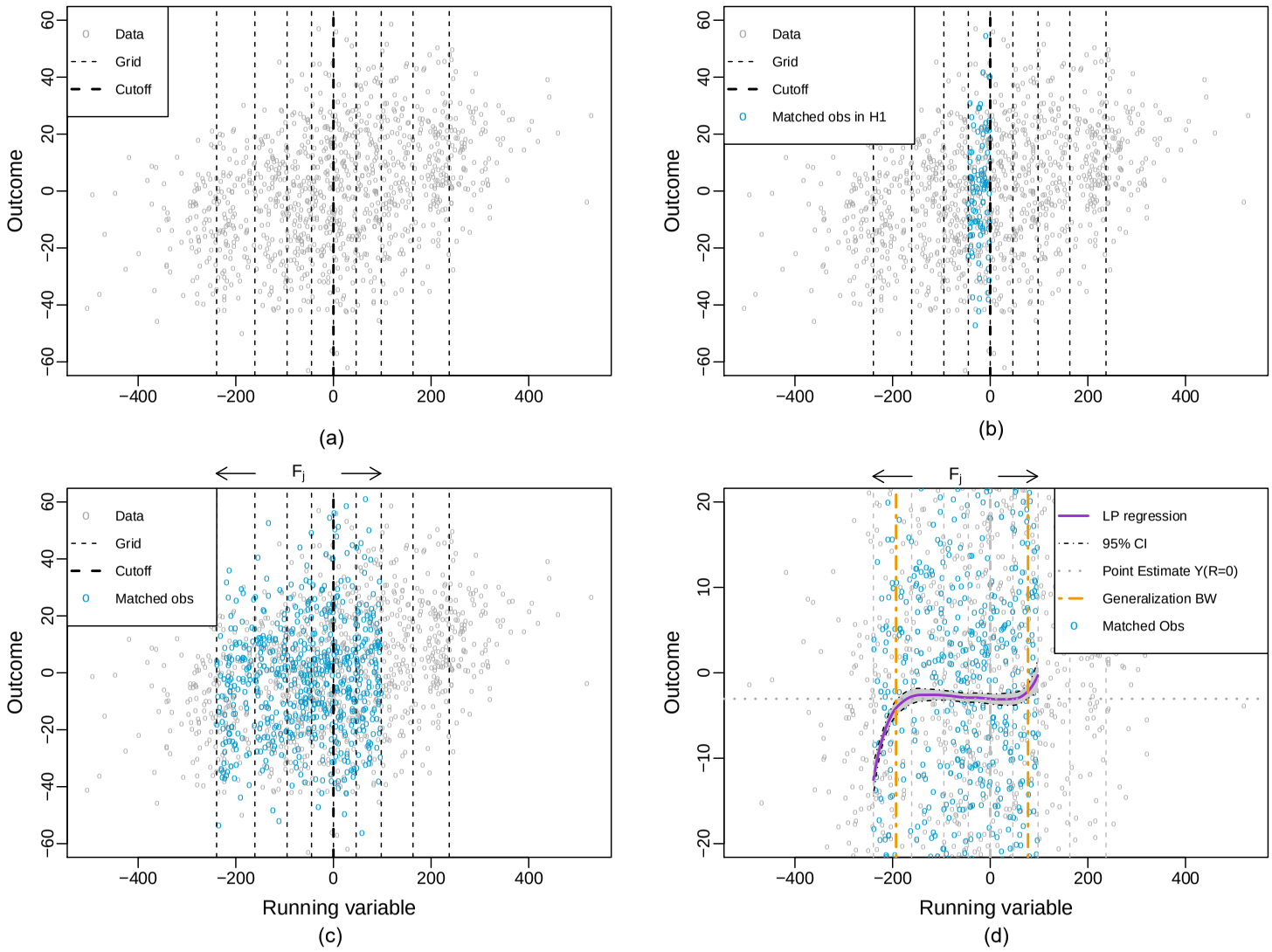


Figure 4: Illustration of generalization of RD using GRD procedure for pre-intervention period

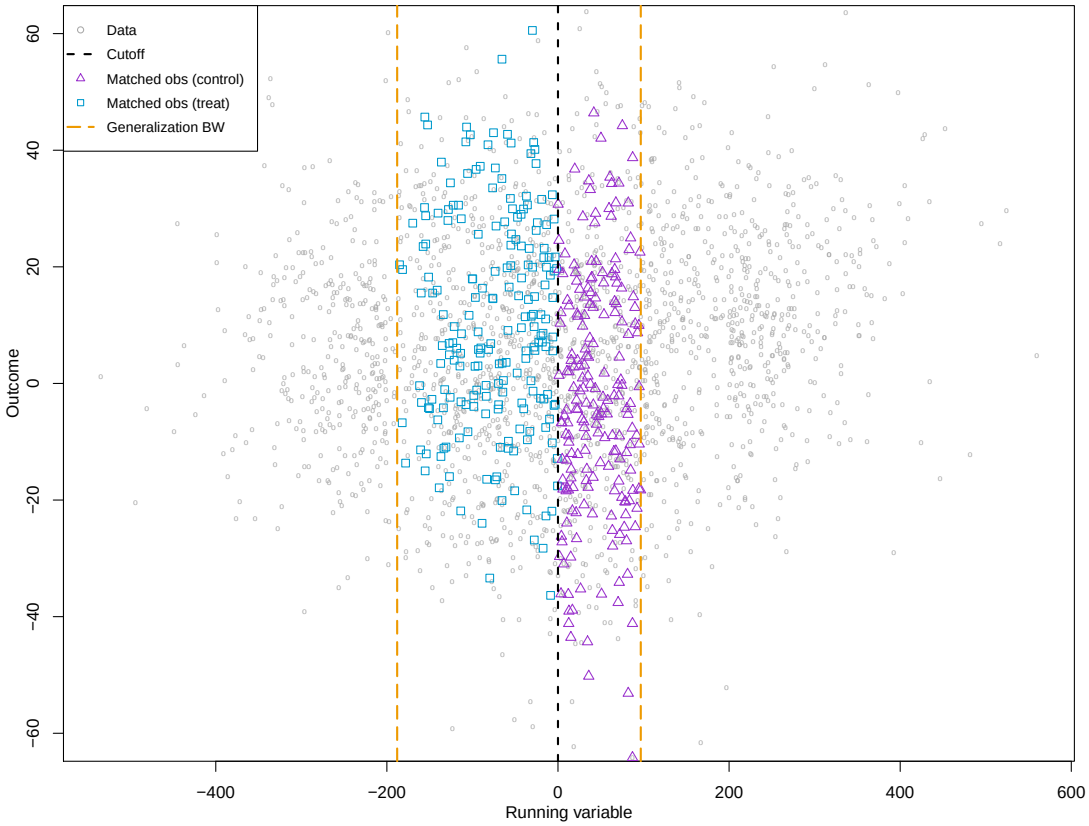


Figure 5: Matched data for post-intervention period within generalization bandwidth

It is important to note that the generalization bandwidth is dependent on the balancing restrictions that are imposed. The more stringent the balancing constraints are, the closer the match (if feasible), but could potentially result in a more narrow generalization bandwidth.[8] This trade-off will be context-dependent for now, and is left up to the researcher to propose the appropriate balancing restrictions according to the specific context.[9]

The maximum bandwidth H^* found for the pre-intervention $t = 0$ period is applied to the post-intervention $t = 1$ period, under the assumption that, absent the intervention, the relationship between the running variable and the outcome would not have changed across

[8]The generalization bandwidth could be narrower under more stringent balancing restrictions because it would be harder to find matching units for the template. However, on the other hand, more restrictive balancing constraints might be better at explaining away the relationship between the running variable and the outcome of interest.

[9]In future work, optimization strategies for the trade-off between balancing constraints and overlap will be developed.

time. However, this assumption poses the challenge of having to match post-intervention units under the same balancing restrictions as the pre-intervention units. One of the advantages that RTM provides in this setting is that, unlike other approaches, it allows for matching of all four groups (treatment and control group before and after the intervention) under the same balancing restrictions. Unlike propensity score approaches used for difference-in-differences, units are not only matched on the "average" characteristics of a matched pair (Keele et al. 2019) to obtain a set of four matched observations, but the exact same balancing constraints can be enforced for all four units in the matched group. This does not only provide more precision in the estimated effects, but will also provide estimates that are less sensitive to hidden biases.

3 Simulations

In this section, simulated data is generated to assess the performance of GRD under different scenarios against a traditional RD estimator. For simplicity, simulations include one observed covariate (x_{it}), one unobserved confounder (u_{it}), a running variable r_{it} with mean 0, and an observed outcome y_{it}, for two periods $t = 0, 1$. Data is generated using normal distributions for the observed and unobserved covariates ($X \sim \mathcal{N}(0, 10)$ and $U \sim \mathcal{N}(0, 10)$), and the running variable r is generated as a linear combination of x and u plus a disturbance error, $r_{it} = \beta_{r,x} x_{it} + \beta_{r,u} u_{it} + \varepsilon_{it}$, with a range of $[-600, 600]$. Outcome y is also generated as a linear function of x and u and a error term, $y_{it} = \beta_{y,x} x_{it} + \beta_{y,u} u_{it} + \nu_{it}$, in addition to a treatment effect τ for treated units. Parameters $\boldsymbol{\beta}$ and τ depend on the scenario, as well as the sample size for each period.

The three dimensions that define the different simulation scenarios are:

- Sample size: Large (20,000 obs) and small samples (2,000 obs).
- Correlation between x and r: Low (0.33) and high (0.66).

- Treatment effect: constant ($\tau = 0.2\sigma$) and heterogeneous ($\tau = 0.2\sigma + 0.0025\sigma \cdot r$), where $Var(y_{i0}) = \sigma$.

For each scenario, the true generalization bandwidth is set to $[-200, 200]$, and balancing restriction on the observed covariate x_{it} is set to fine balance on the covariate deciles. The template size for large samples is 1,000 observations, while for the small sample size is 100. GRD estimates for the generalization bandwidth are compared to traditional RD estimates obtained using the `rdrobust` package in R (Calonico et al. 2018).

Three measures are used to assess the performance of the estimator based on its error term, $\epsilon = \hat{\tau} - \tau$: bias, variance, and root mean square error (RMSE). Bias is measured as the ratio of the difference between the estimated and the true effect over the true effect, and variance is estimated as a percentage of the true effect. Using the error term of the estimator to assess performance allows for the comparison of the traditional RD against the GRD estimator for heterogeneous effects. As the bandwidth estimated by the GRD also varies with each simulation, introducing variance for the true effect in the heterogenous scenario across simulations, the performance of the estimator is assessed without confounding the variance of the bandwidth estimation with the estimator variance.

Table 1 shows the results for 500 simulations under different data generating processes according to the parameters mentioned above. For small sample sizes, the GRD estimator performs as well as the robust RD estimator: Even though the generalization estimator presents slightly higher bias, variance is lower. Simulation for large samples show a similar story, though in most cases there is a significant reduction in bias for both estimators. Particularly for large samples under heterogeneous effects, the GRD estimator presents small biases, similar in magnitude to the RD robust estimator, and even higher precision than the latter. In this case, however, GRD is obtaining estimates away from the cutoff of eligibility, and not exclusively at the threshold.

Table 1: Error term performance between `rdrobust` and GRD estimator for different scenarios

(a) True effect $\tau = 0.2\sigma$

		Small Sample			Large Sample		
		Bias (%)	**Var (%)**	**RMSE**	**Bias (%)**	**Var (%)**	**RMSE**
Low correlation	**Robust RD**	-0.009	0.619	1.573	-0.002	0.060	0.490
	GRD	0.025	0.552	1.486	-0.042	0.060	0.521
High correlation	**Robust RD**	-0.003	0.608	1.559	0.012	0.064	0.511
	GRD	-0.008	0.599	1.547	-0.030	0.057	0.495

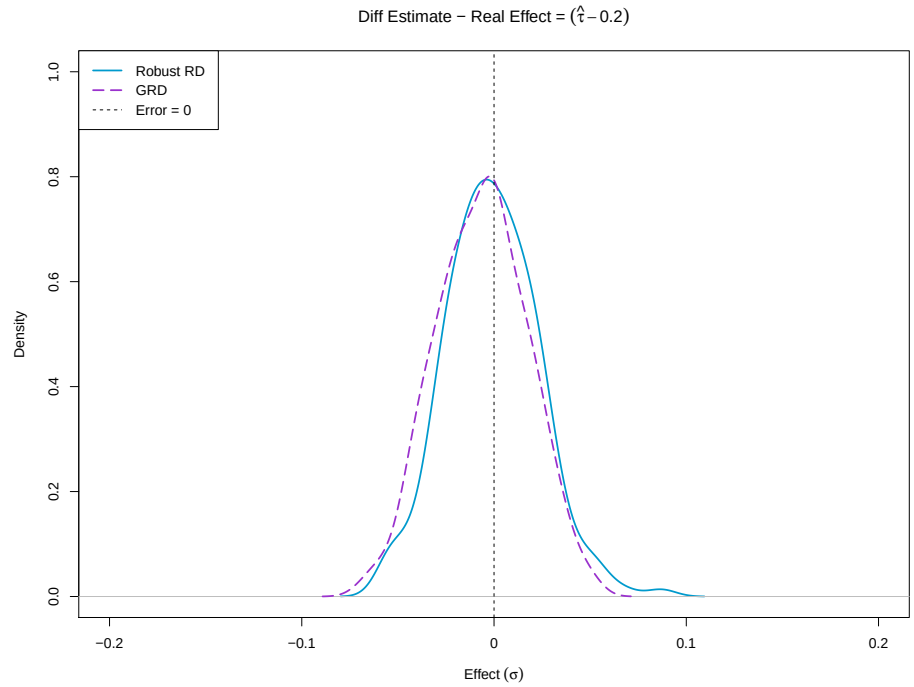
(b) True effect $\tau = 0.2\sigma + 0.0025\sigma \cdot r$

		Small Sample			Large Sample		
		Bias (%)	**Var (%)**	**RMSE**	**Bias (%)**	**Var (%)**	**RMSE**
Low correlation	**Robust RD**	0.006	0.594	1.540	-0.001	0.061	0.493
	GRD	0.017	0.318	1.612	0.008	0.027	0.464
High correlation	**Robust RD**	-0.024	0.621	1.577	-0.001	0.083	0.576
	GRD	-0.011	0.279	1.457	0.003	0.034	0.497

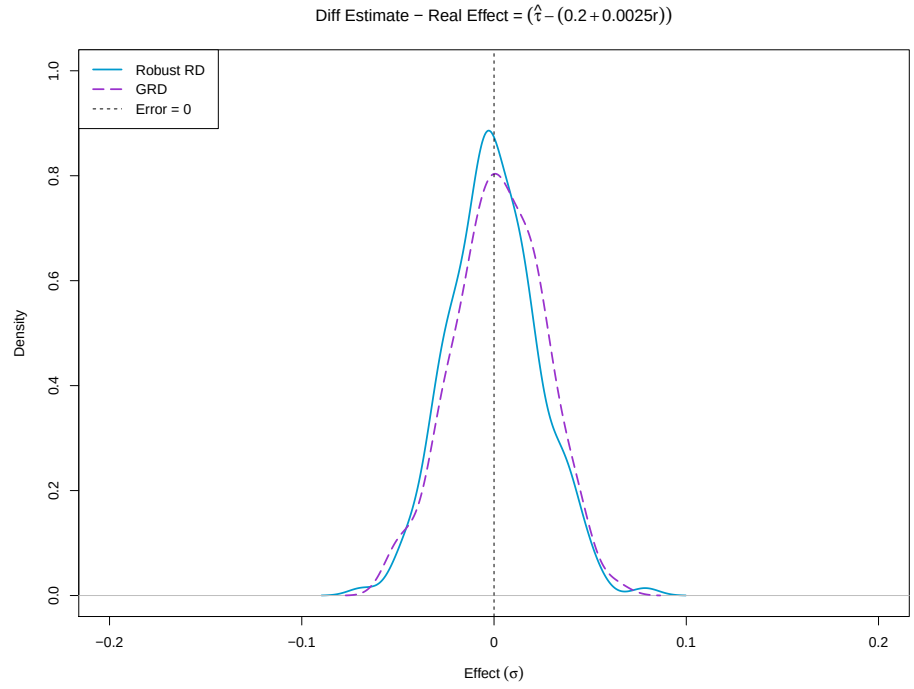
Notes: Performance for error term of RD robust estimator (Calonico et al. 2018) and GRD, calculated as $\epsilon = \hat{\tau} - \tau$, where $\hat{\tau}$ is the estimate for each method and τ is the true effect.

Figure 6 shows the error distribution for both estimators for the scenario of high correlation and large sample size.

In terms of the generalization bandwidth (Table 2), estimation of the generalization interval is more precise in large samples, particularly with low correlation. For example, in large samples, the estimated lower bound of the generalization bandwidth is -183.5 ($SD = 23.7$)for low correlation between x and r, while for high correlation scenarios, the estimated lower bound is -147.2 ($SD = 38.5$). The estimation is conservative compared to the true generalization bandwidth, especially in high correlation scenarios, given the use of a smooth local polynomial function.



(a) Error term distribution for RD and GRD under constant treatment effect



(b) Error term distribution for RD and GRD under heterogenous treatment effect

Figure 6: Error term distribution for RD and GRD estimators under large samples with low correlation between x and r

Table 2: Mean generalization bandwidth for different settings of simulations

	Small sample	**Large sample**
Low correlation	-201.92	-183.54
	(52.18)	(23.70)
High correlation	-166.63	-147.17
	(42.23)	(38.53)

Notes: Standard deviation shown in ()

4 Application: Free Higher Education in Chile

In this section, an application for the previously described method is presented, using the introduction of Free Higher Education in Chile as a case study. First, the intervention is described, as well as the specific context of higher education in Chile and the data used for the analysis. The results for traditional RD are shown for application and enrollment outcomes. The implementation for the generalized RD in this setting is also outlined, and report results for university enrollment and applications.

4.1 Introduction of Free Higher Education and the Chilean Context

Even though recent work has shown that attending college is not a beneficial decision for everyone (Heckman et al. 2016; Hastings et al. 2014), there is concrete evidence that, at least for some students, earning a higher education degree can positively affect life outcomes (Dee 2004; Hastings et al. 2015). However, students face multiple constraints when deciding whether to attend college: previous college preparation, information frictions, and credit constraints, just to name a few (Dynarski & Scott-Clayton 2013). The latter are particularly salient in the policy world, as governments and different private and public entities have invested in several programs and interventions to try to alleviate this restriction.

One of the potential policies that can be adopted to relief students from credit constraints is Free Higher Education (FHE). However, FHE is seen as a controversial inter-

vention. Even though certain advocates say that it promotes a more equal society, there are also detractors that argue that it is a regressive policy (Barr 2003). Additionally, given recent evidence of the heterogeneity in the return of a college degree in Chile (Hastings et al. 2015; Rodriguez et al. 2016), it is not clear whether FHE would provide an inadequate incentive for students to enroll in university.

In the Chilean case, even though until 2015 the government offered a vast array of financial aid instruments for attending college, including scholarships for students up to the third lower income quintile (Bucarey 2018), and income-contingent government-backed loans with low interest rates, there was an important pressure from citizens' movements to implement free higher education. Thus, at the end of 2015 the government implemented a FHE policy. Students that wanted to opt into this policy had to comply with two requirements: (1) be admitted or enrolled in a university that had agreed to participate in the FHE program, and (2) belong to the bottom half of the income distribution (MINEDUC 2016). In December 2015, the Chilean Congress included a decree in the national budget which allowed students belonging to the lowest 50% of the income distribution to study in most universities in Chile for free. The policy benefited around 150,000 students who were admitted to universities in 2016.

In terms of higher education, Chile has a centralized admission system to most universities[10] that operates through a deferred admission mechanism. Students must take the admission test (PSU) in order to apply to university, and their admission score is a weighted average of their PSU scores, high school GPA, and high school ranking.[11] Once their PSU scores are published, students must rank 10 preferences for institution-degrees, and get accepted based on their admission score and the number of slots each institution-degree offers.[12]

[10]All selective and prestigious universities participate of the centralized admission system.

[11]Unlike other countries like the US, admission scores are the only determinant in a student's regular admission to university.

[12]Unlike the US, Chilean students apply to a specific degree within an institution when applying to college

Until 2015, government scholarships or government-backed loans provided to lower income students only covered up to a "referential" tuition cap.[13] The difference between the referential and real tuition needed to be covered by the student either by out-of-pocket payments, loans (government or private), or university scholarships. The introduction of free higher education for vulnerable students, which covered uncapped full tuition, meant that lower income individuals who were admitted into university did no longer have financial constraints regarding the cost of full tuition, allowing them to potentially choose better institutions which tend to be more expensive. However, anecdotal evidence shows that prior to the introduction of FHE most students in lower income deciles that were admitted into university received other financial aid that helped them cover the gap.

It is important to note that even though government financial aid was extensive in terms of scholarships and income-contingent loans, they did require a minimum PSU score to be eligible for them. Specifically, government scholarships required 500 points on average between PSU Math and PSU Language, and government-backed loans required 475 points. Even though these thresholds were not particularly high, especially considering the cutoff scores for admissions, they did affect lower-performance/lower-income students who, in absence of the score requirement, could be eligible for financial aid. FHE policy, on the other hand, did not establish a performance threshold to obtain the benefit and, as a result, might have been more relevant for students in this group.

Additionally, FHE might have had other effects on students' outcomes, other than enrollment. Lower-income students who otherwise would have not applied to college might change their decision due to this new policy. Students' career choices might be affected as well, as the cost of studying a "less profitable" degree is reduced if they intended to pay for their higher education with a loan instead of a scholarship in the absence of the policy.

(e.g. Engineering at Catholic University), and each institution-degree combination has their own cutoff scores and available slots.

[13]Referential tuition was calculated by the government based on groups of institutions and degrees, and were always lower than the real tuition charged by universities.

Finally, FHE could have also reduced the gap in access between lower and higher income students, reducing inequality, especially in the most prestigious universities.

Even though there is still no consensus on the impact that the policy had, recent work by Bucarey (2018) using data from 2008 until 2015 (before the FHE policy was implemented) has shown that the introduction of this new financing policy may actually be detrimental for lower income students if expanded. Leveraging RD estimates from the cutoff PSU score for scholarship eligibility before the introduction of FHE, and combining it with a structural approach, Bucarey (2018) finds that implementing free higher education for the entire population would actually produce an increase in cutoff scores for more selective (and lucrative) institution-degrees, crowding out more vulnerable students.

Given that actual receipt of the treatment was bind to enrollment, the intervention of interest is the eligibility for free higher education, e.g. belonging to the bottom 50% of income distribution. In this analysis, only partial equilibrium effects of the introduction of free higher education in Chile are considered, in the way it was implemented during its first year. It is important, however, to keep in mind potential general equilibrium effects which could provide substantially different results in the long run or on scale-up (Bucarey 2018).

4.2 Data

The data used for the analysis corresponds to administrative data from the Ministry of Education and DEMRE, the department that administrates the Chilean university admission exams called PSU. The data provided by the Ministry of Education comes from the Educational Quality Measurement System (SIMCE), and includes past scores for standardized tests (8th grade for 2014, and 10th grade for 2015 and 2016) at the individual and school level, socioeconomic and demographic information for each student,[14] school characteristics,

[14]Some of the covariates include self-reported household income, parental education, number of books at home, type of health insurance, among others.

academic performance (i.e. grades and ranking), as well as school attendance. The data provided by DEMRE contains complete information from the admission process 2014-2016, which includes two years before FHE was implemented and one year after. The data includes demographic characteristics for each student enrolled in the PSU,[15] individual and family socioeconomic variables, PSU scores, application preferences and admission results.

The Ministry of Education also provided income per capita and socioeconomic deciles at the student level, which were used to assign students to different benefits related to higher education (e.g. scholarships and loans). The Ministry has both the declared income per capita (initial) by the student in the Unique Socioeconomic Assistance Form (FUAS), as well as the verified income per capita (final).

Even though FHE has been implemented for a few years to the date, only the year 2016 is used as the post-intervention period. This decision was made based on the fact that for 2016 the policy was not expected,[16] and did not have an effect on other inputs for admission, such as scores. The fact that the introduction of the policy was not anticipated by students also provides additional robustness to the assumption that other key confounders did not change between the pre- and post-intervention period for our population of interest. The outcomes of interest for the analysis include enrollment rates in university, as well as application decisions.

4.2.1 Sample selection

The final sample focuses on senior high school students who had valid income per capita given their application to the Unique Socioeconomic Assistance Form (FUAS).[17] Given that FHE also required students to be admitted into university, only students with valid math and language PSU scores are analyzed; these are the two mandatory tests students' need to

[15] Over 90% of senior high school graduates from the respective year enroll in the PSU.

[16] FHE was actually introduced in the budget the very last day possible, and almost a month after students had already taken the admission exam.

[17] Most students below the 9th income decile complete the FUAS, which is a requirement to apply not only to scholarships, but also government backed loans.

take in order to apply to college. Additionally, only students who can be matched to previous data provided by the Education Quality Measurement System (SIMCE) for either 8th (2014) or 10th grade (2015-2016) are kept. Finally, due to the implementation of an additional socioeconomic adjustment for some students in 2016, the sample is comprised only by students who were not affected by this change. In 2016, an additional criteria was incorporated to adjust socioeconomic deciles based on other students' assets and circumstances: it does not alter the verified income per capita, but it adjust their assigned decile. In turn, only 23% of students in the sample in 2016 have the same final assignment decile as the one determined by their verified income per capita. Given this new adjustment method, only observations that maintained their decile assignment are used.[18]

To verify that the latter decision did not affect the internal validity of the identification strategy, the income distribution before and after dropping the adjusted observations in 2016 is analyzed. Figure 7 shows the distribution for both samples, showing that there is no bunching or jumps in the distribution around the cutoff of eligibility.

The final sample has 199,334 observations. Table 3 shows some characteristics from the final sample of analysis.

[18]All deciles were approximately affected in the same way by the adjustment, but particularly the bottom 3 deciles which have a higher rate of adjustment (changed to a higher decile). On average, within every verified income decile, students affected by the adjustment increased their assgined decile by one.

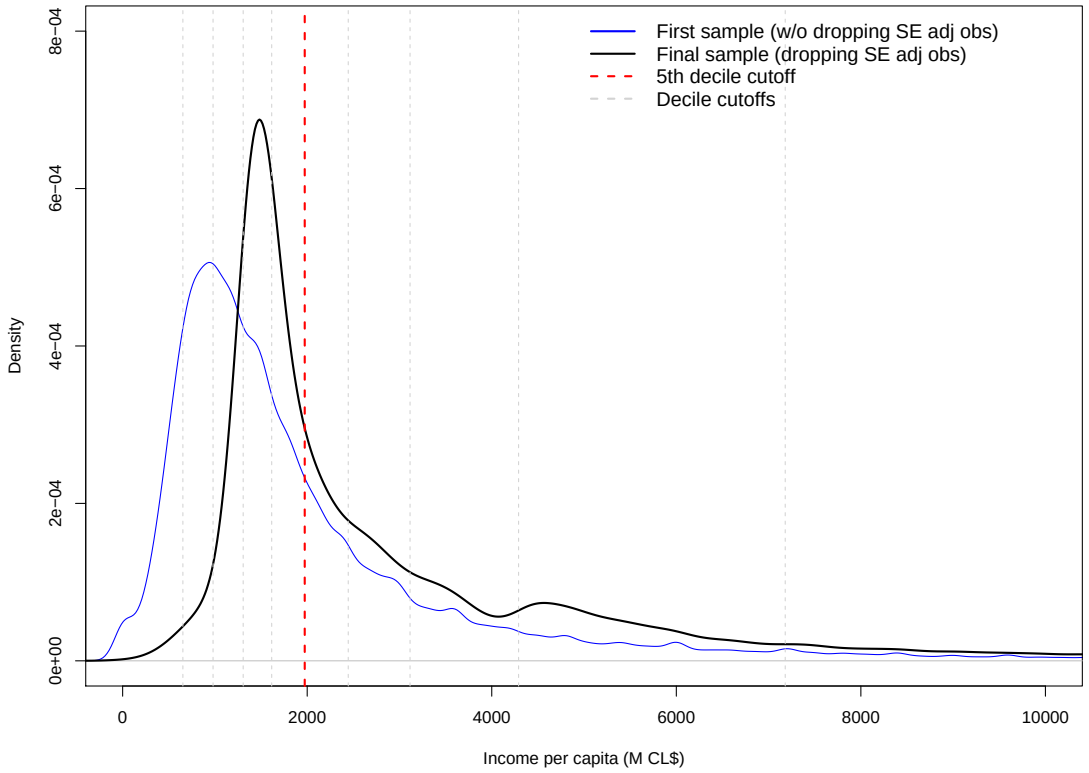


Figure 7: Distribution of student income per capita for sample before and after dropping observations with adjusted deciles

Table 3: Sample characteristics of students by year

	2014	2015	2016
Female	0.55	0.55	0.53
Mother's Ed			
Primary, but less than HS	0.13	0.12	0.09
High school grad	0.38	0.39	0.36
Some technical	0.05	0.05	0.06
Technical grad	0.12	0.12	0.16
Some university	0.03	0.03	0.03
University grad	0.09	0.09	0.17
Missing	0.01	0.02	0.03
Father's Ed			
Primary, but less than HS	0.13	0.12	0.09
High school grad	0.34	0.35	0.32
Some technical	0.04	0.04	0.05
Technical grad	0.09	0.09	0.11
Some university	0.04	0.04	0.05
University grad	0.11	0.11	0.19
Missing	0.06	0.07	0.07
Public health insurance	0.75	0.76	0.63
Lives in Metropolitan Region	0.37	0.36	0.42
SIMCE score (student)	276.71	279.56	286.61
HS ranking score	592.40	588.67	595.67
GPA score	565.40	559.58	571.14
Language PSU score	509.45	509.05	530.82
Math PSU score	515.31	512.16	535.56
Public school	0.36	0.32	0.27
Average SIMCE score (school)	262.84	270.33	277.75
School SES group	3.03	2.56	3.10
Observations	76654	94952	27728

4.3 Results

4.3.1 RD findings

Figure 8 shows the allocation of FHE by income decile. There is a clear sharp discontinuity after the 5th decile, which is consistent with the allocation of the policy given the government requirements. In this case, the eligibility criteria (i.e. belonging to the 5th income decile or lower) does not necessarily imply that the student will get FHE. One of the most important requirements to get FHE besides income eligibility is being admitted into a university that subscribed to the policy. However, it is clear that the treatment variable used for the analysis, belonging to the bottom 50% of the income distribution, did affect the probability of assignment to FHE.

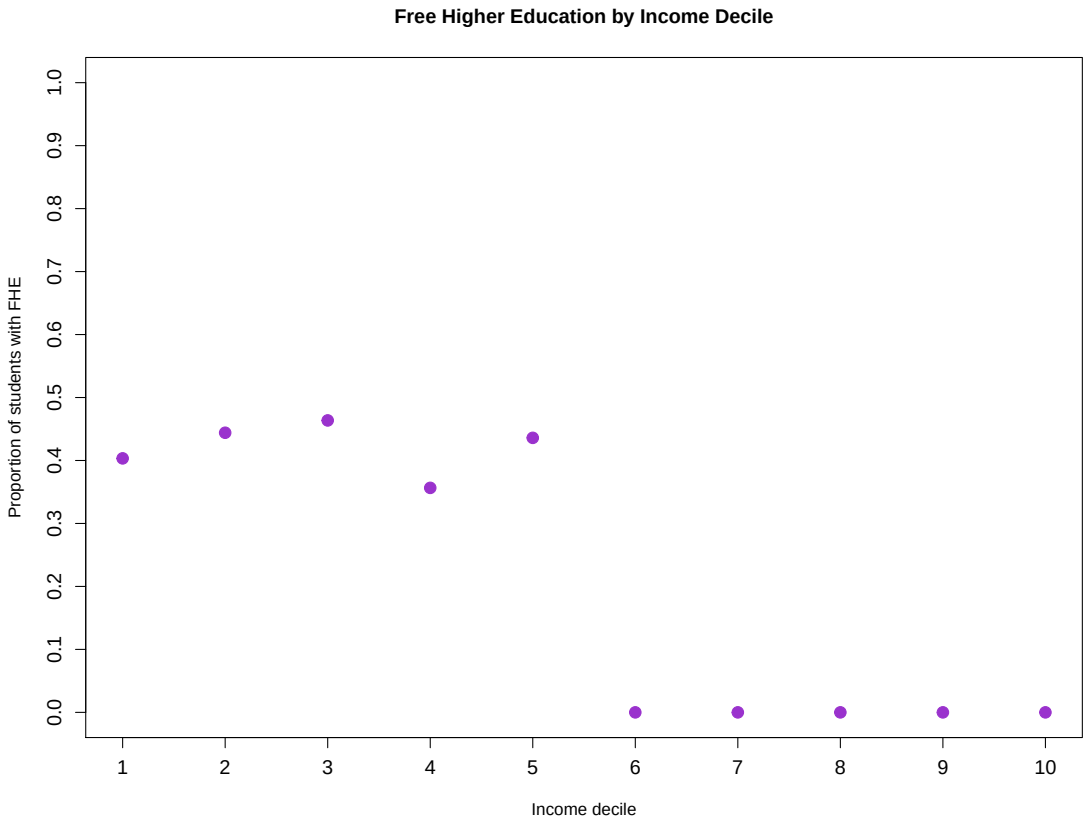


Figure 8: Proportion of students by income decile assigned to Free Higher Education in 2016

In this section, focus is on two different outcomes: application to university and

enrollment. The first one measures the magnitude of the effect that being socioeconomically eligible for FHE had on applying to a university; [19] The second one captures the effect of the same treatment on actually enrolling in college.[20]

In the Chilean context, enrollment in university depends on the admission score (i.e. being selected subject to application) and students' decisions, such as whether they apply to university and where they apply. Given that both in the traditional RD strategy, as well as in GRD, determinants of the admission score[21] are closely adjusted for, it is possible that the main differences in enrollment comes from adjustment in preferences/decisions triggered by the FHE policy. For instance, eligible students that would have not applied to college in previous years due to cost of tuition, might have changed their minds once the policy was introduced.

For illustration purposes, Figure 9 shows the average enrollment in university both pre- and post- intervention by income per capita. The figure shows that while there is a smooth curve along the 5th decile cutoff in 2015, there is a clear jump in enrollment in 2016 at this same threshold, which is an indication of an effect around the cutoff.

To check whether there are discontinuities at the cutoff related to other characteristics, a set of 10 covariates used for adjustment is plotted against income per capita in 2016. Figure 10 shows the plots for 10 different covariates, where there are no significant discontinuities around the cutoff.

Table 6a shows the results for both outcomes using a traditional RD approach from the `rdrobust` package (Calonico et al. 2018).[22] The effect on application is positive, but

[19] An application is considered successful if a student marked at least one preference for an institution-program during the application period and previously took the mandatory tests required for the program.

[20] Notice that enrollment implies that the student had at least one successful application.

[21] Admission scores differ by university-program, given that it is a weighted average of PSU, GPA, and ranking scores, in addition to history and science PSU scores, for which the matching strategy does not control for. The latter tests are optional, and many non-selective programs do not require them and/or weight them significantly less than the other three scores.

[22] In the traditional RD strategy, the same set of covariates is used as controls as in the matching strategy

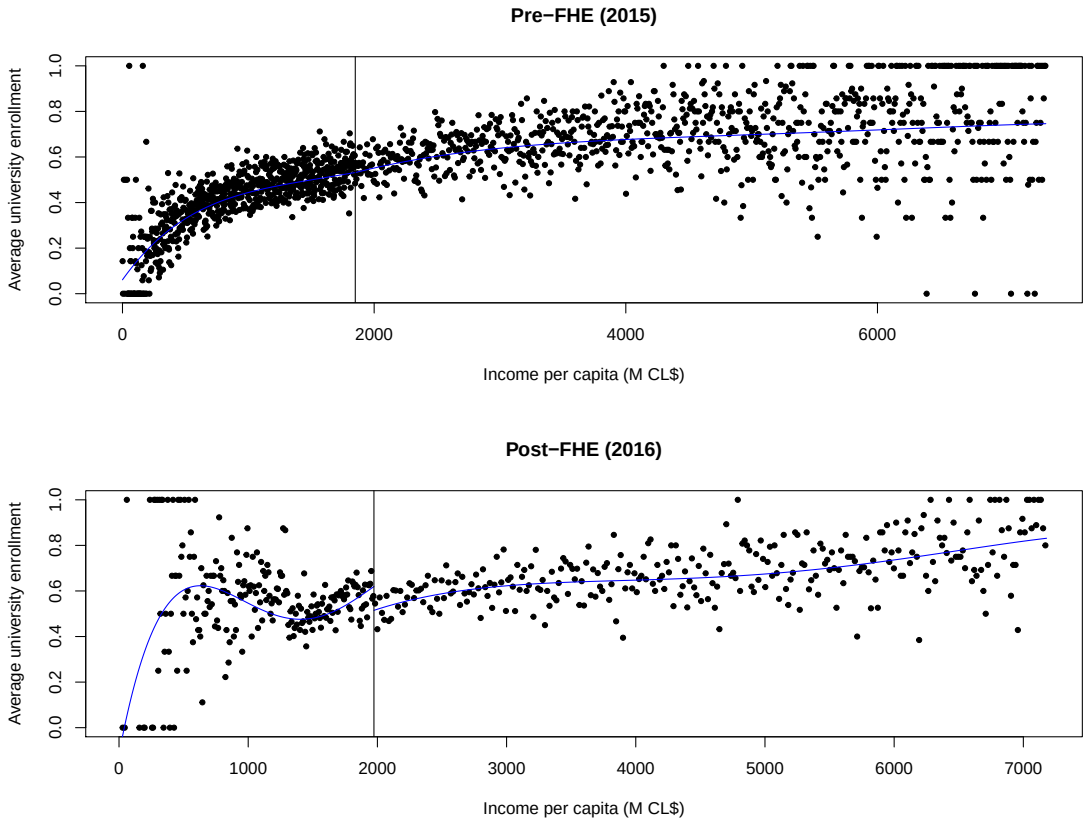


Figure 9: Mean university enrollment rates before (2015) and after (2016) FHE was implemented by income per capita

not statistically significant at conventional levels (point estimate = 0.035; p-value = 0.105). The effect on enrollment, however, is larger and statistically significant (point estimate = 0.07; p-value = 0.02). These results show that the policy had a positive and significant effect on outcomes, ranging from 6% to 13% increase with respect to the control group at the cutoff.

for the GRD to maintain consistency.

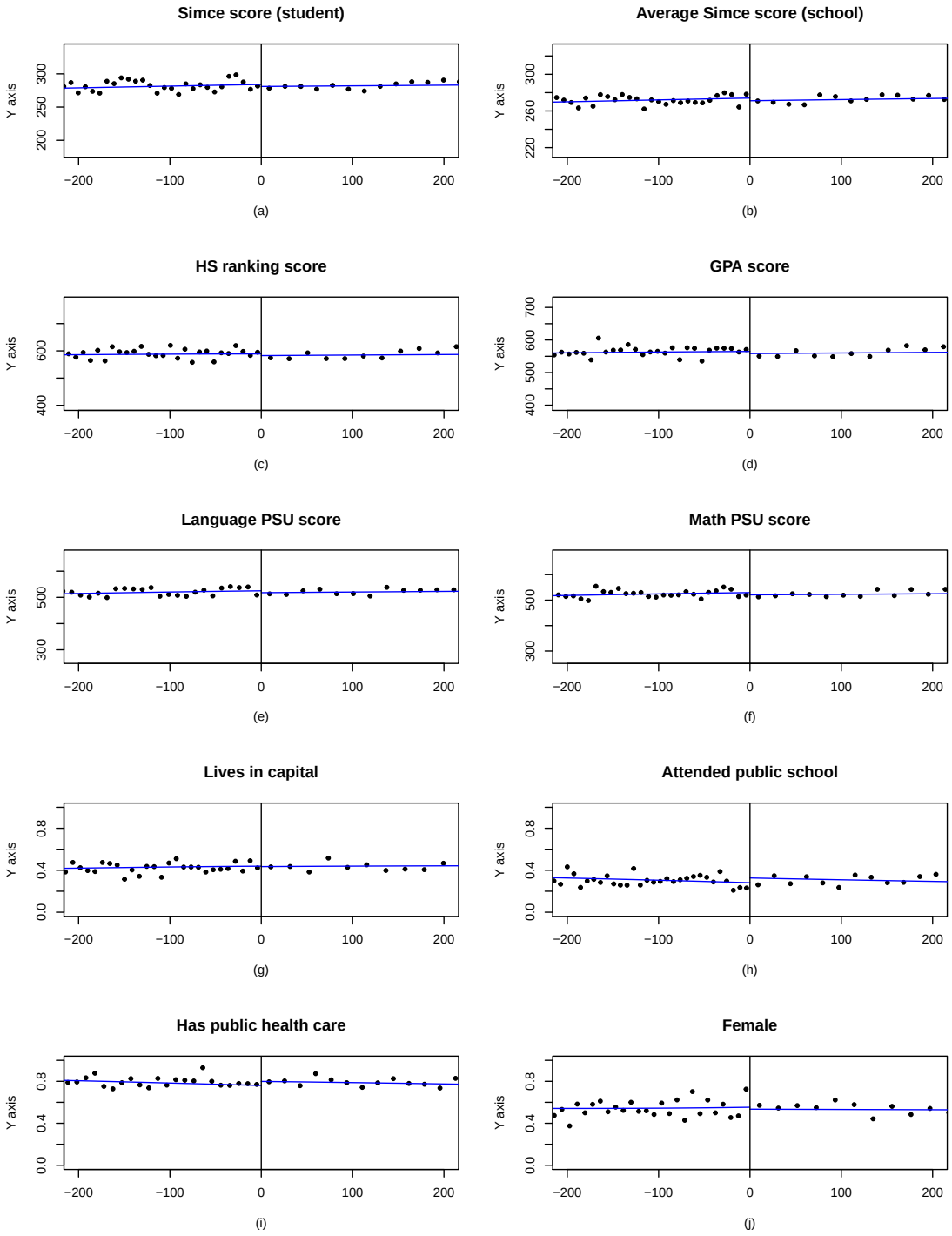


Figure 10: Average covariates for students by adjusted income per capita in 2016

4.3.2 GRD findings

To implement the generalized regression discontinuity design as proposed in subsection 2.2, a template T of size 1000 is chosen from 100 S random samples.[23] Based on the total sample size of the post-intervention period and the size of the template, the grid for matching is built by dividing the running variable into 20 quantiles. Starting with a representative template of the treatment units in the narrowest bandwidth, each grid bin is matched under the following balancing constraints to the template:

- Restricted Standardized Mean Difference: For the following covariates, restricted their mean difference is imposed to be no greater than 0.05 SD. *Student's 10th grade SIMCE score (standarized test), school average 10th grade SIMCE score, school SES group, student's high school ranking score, GPA score, PSU Language score, PSU Math score, reside in the Metropolitan region, attended a public school, has public health care.*

- Fine Balance: The fine balance restriction allows for balancing of the marginal distribution of categorical covariates. The following covariates were matched under this constraint: *Gender, Mother's education (8 categories), Father's education (8 categories), PSU Language score (deciles), PSU Math score (deciles), GPA score (quintiles)*

Given that parental education, as well as school characteristics and student's performance are highly predictive of household income, that set of covariates is matched more closely than other characteristics. Additionally, matching closely on admission scores ensures that the probability of admission would only be affected by the decision of whether and/or where to apply made by each student. The intervention might have also affected these preferences.

After conducting representative template matching gradually getting away from the

[23]The starting size of the template corresponds roughly to the number of observations for the treatment group within the narrow bandwidth.

cutoff under the balancing restrictions previously described, the generalization bandwidth for both application and university enrollment was estimated. The generalization interval for the application is [-M CL$500.26, M CL$1,254.86], and for enrollment is [-M CL$500.26, M CL$300.88] around the cutoff. For comparability reasons, the narrowest bandwidth is used for both outcomes.

The generalization interval estimated for the pre-intervention period is [-M CL$500.26, M CL$300.88], and it is applied to the post-intervention period. The assumption of conditional-time invariance can be tested for the post-intervention period using the control side. Figure 11 shows a local polynomial fitted to the control matched units in the post intervention period. The lack of trend in the local polynomial function provides additional evidence that the conditional time-invariance assumption at least partially holds for the control-side of the distribution.

The generalization interval in the post-intervention period includes students from the 4th and 5th income decile on the treatment side, and on the treatment side, comprise of 50% of the overall treated population. To benchmark the results of the matching procedure, Table 4 shows the mean characteristics for the complete 2016 sample before matching, for both the treatment and control group. Table 5 shows the balance between both the matched treatment and the control group within the generalization bandwidth for the pre- and post-intervention period. Panel (a) of the same table shows the variables matched on restricted standardized means, while panel (b) shows the results for fine balance.

Figure 12 shows the balance for covariates in 2016 for the entire sample, the sample within the generalization bandwidth, and the matched sample for the GRD.

Figure 13 shows the RD plot for the post-intervention period, highlighting the narrowest bandwidth used for the RD estimate, as well as the maximum generalization bandwidth found using the previously described method for the enrollment outcome.

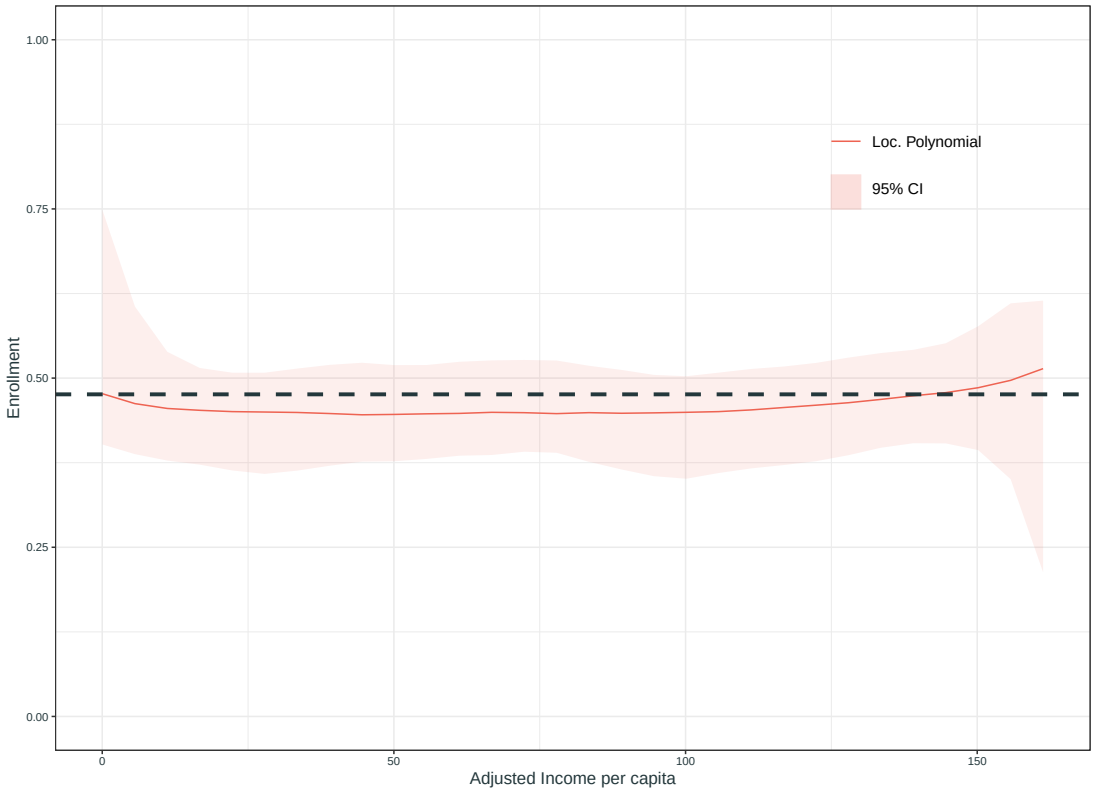


Figure 11: Local polynomial fit for post-intervention period matched sample for control units: Partial test for conditional time-invariance assumption.

Table 4: Sample characteristics for the year 2016 for students below (T) and above (C) the 5th income decile

	Treatment group (SE decile ≤ 5)	Control group (SE decile > 5)	Std Dif T-C
Female	0.55	0.52	0.05
Mother's education (years)	11.37	13.55	-0.67
Father's education (years)	11.52	13.83	-0.76
Language PSU score	504.08	556.90	-0.51
Math PSU score	507.69	562.75	-0.53
GPA score	554.88	586.99	-0.33
Ranking score	579.84	611.11	-0.26
SIMCE 10th grade (student)	274.90	298.03	-0.48
SIMCE 10th grade (school)	266.91	288.32	-0.62
SES group school	2.68	3.50	-0.82
Lives in Metropolitan region	0.40	0.44	-0.08
Public school	0.35	0.19	0.38
Public health insurance	0.82	0.45	0.83

Within the generalization bandwidth, both the treatment group and the control group are matched under the same balancing constraints used before, minimizing the distance

Table 5: Balance across matched samples for pre- and post-intervention period for observations within the generalization bandwidth for external validity of university enrollment

(a) Restricted mean balance (0.05 SD)

	Pre-intervention			Post-intervention		
	Mean T	Mean C	Std Dif	Mean T	Mean C	Std Dif
SIMCE (student)	279.06	278.90	0.00	277.53	278.06	-0.01
SIMCE (school)	269.41	270.29	-0.03	268.98	268.99	-0.00
School SES group	2.60	2.60	-0.00	2.64	2.64	0.00
Ranking (score)	583.22	583.73	-0.00	583.79	583.01	0.01
GPA (score)	556.22	555.86	0.00	557.22	557.92	-0.01
PSU Language	511.83	510.55	0.01	510.31	509.39	0.01
PSU Math	513.47	513.92	-0.00	512.44	512.29	0.00
Lives in capital city	0.40	0.41	-0.02	0.40	0.41	-0.02
Public school	0.31	0.31	0.00	0.33	0.33	-0.00
Has public health insurance	0.81	0.80	0.01	0.81	0.80	0.03

(b) Distributional balance (near fine balance). The balance is maintained for all 6 covariates and 35 categories.

	Pre-intervention		Post-intervention	
	C	T	C	T
Gender				
Male	453	453	453	453
Female	547	547	547	547
Father's Education				
Primary or less	168	168	168	168
Less than HS	130	130	130	130
High school grad	355	355	355	355
Some technical	42	42	42	42
Technical grad	101	101	101	101
Some university	43	43	43	43
University grad	83	83	83	83
Missing	78	78	78	78
Mother's Education				
Primary or less	146	146	146	146
Less than HS	121	121	121	121
High school grad	416	416	416	416
Some technical	60	60	60	60
Technical grad	138	138	138	138
Some university	31	31	31	31
University grad	59	59	59	59
Missing	29	29	29	29
Language PSU (deciles)				
1	70	70	70	70
2	83	83	83	83
3	97	97	97	97
4	96	96	96	96
5	109	109	109	109
6	104	104	104	104
⋮	⋮	⋮	⋮	⋮

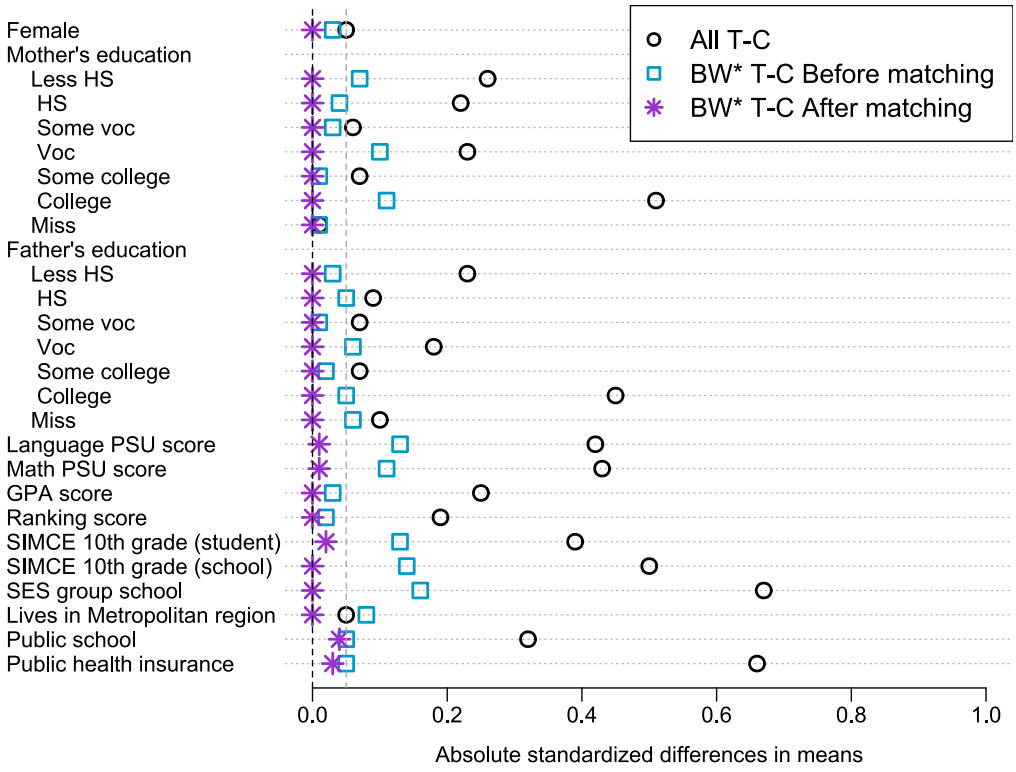


Figure 12: Balance between treatment and control groups for (i) entire sample 2016, (ii) entire sample within generalization bandwidth BW^*, and (iii) matched sample within BW^*

to the cutoff for the control group. Table 6b shows the estimated effect for the average treatment on the treated within the generalization bandwidth using a paired t-test approach. Using the GRD approach, the effect for the treated population within the bandwidth of generalization 5.2 percentage points (p-value = 0.02) for applications and 7.7 percentage points (p-value $<$ 0.01) for enrollment.

Figure 14 shows the difference in application and enrollment between the treated and control group on the sample within the generalization bandwidth of adjusted income per capita, for both the pre-intervention and post-intervention periods (difference-in-differences estimates). Figure 14a shows the mean difference in outcomes by year and treatment group for all students within the generalization bandwidth, without adjusting for covariates. As it can be seen from the figure, there are significant differences in application and enrollment

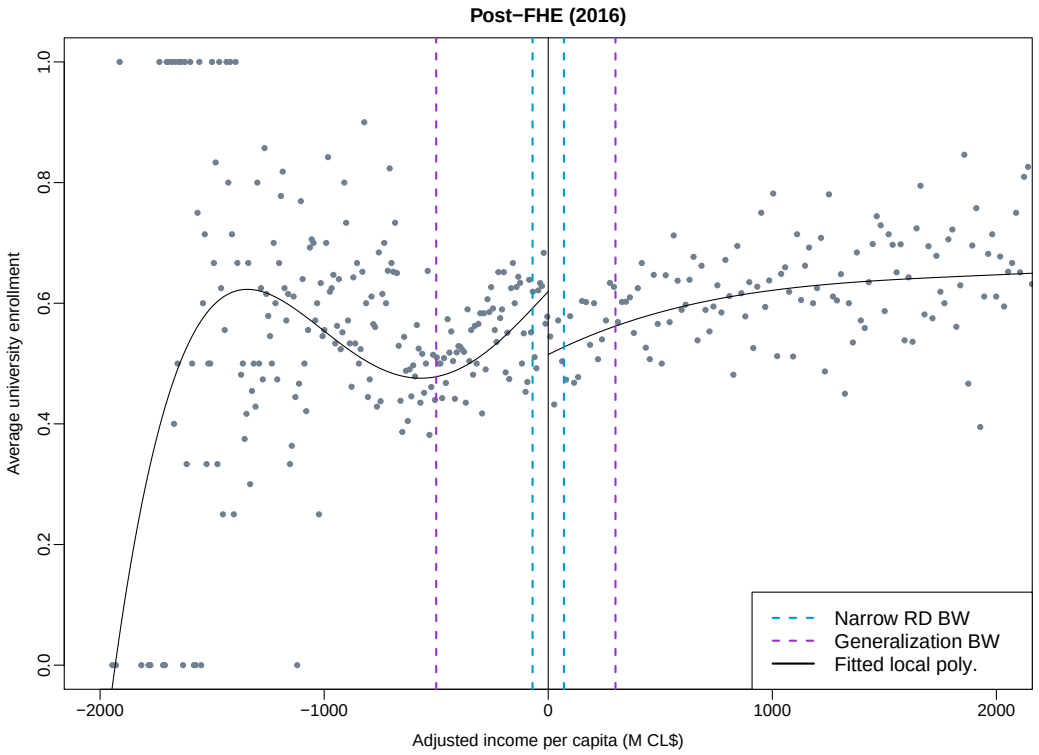


Figure 13: Regression discontinuity plot for year 2016 showing narrowest and generalization bandwidth

in the pre-intervention period, which are due to the relationship between income and the outcomes. Figure 14b shows the difference in application and enrollment for matched students using representative template matching. This figure shows that when the predictive covariates are taken into account, the relationship between the running variable and the outcome can be explained away for the generalization bandwidth, getting a null difference in application and enrollment rates for students in 2015, and a positive treatment effect for the post-intervention year using matching difference-in-differences estimators.

Comparing the results for both outcomes, most of the effect of eligibility to FHE on enrollment appears to be due to an increase in applications. This provides some evidence that the main gap that FHE closed for vulnerable students was on the actual decision to apply to university (Figure 14).

Table 6: Results for outcomes of interest at the cutoff (RD) and within a generalization bandwidth (GRD)

(a) Results for traditional RD

	Application to University	Enrollment to University
Effect	0.035	0.069**
	[-0.007, 0.077]	[0.026, 0.112]
Effective N Obs	6,588	6,458
Mean control	0.606	0.515

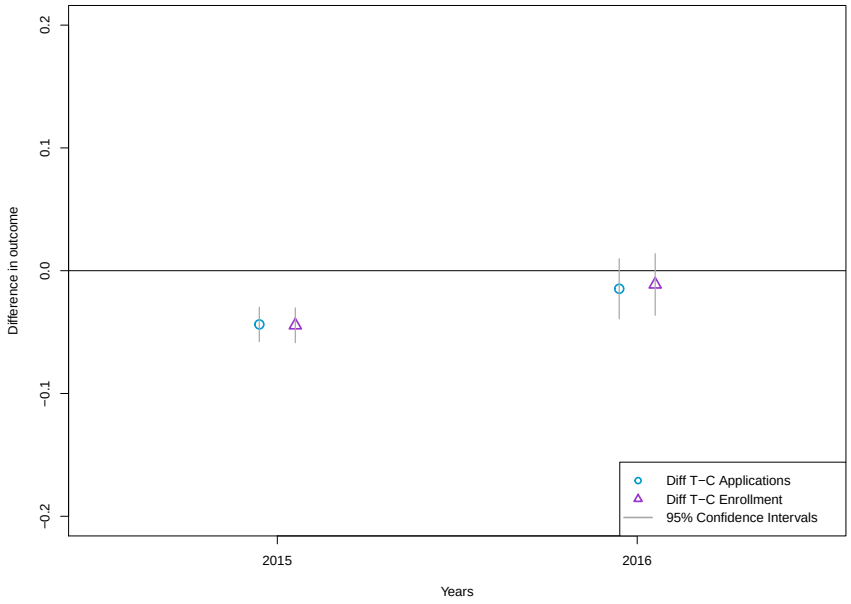
(b) Results for GRD method

	Application to University	Enrollment to University
Effect	0.052**	0.077***
	[0.008, 0.096]	[0.029, 0.125]
N Obs	2,000	2,000
Mean control	0.568	0.472

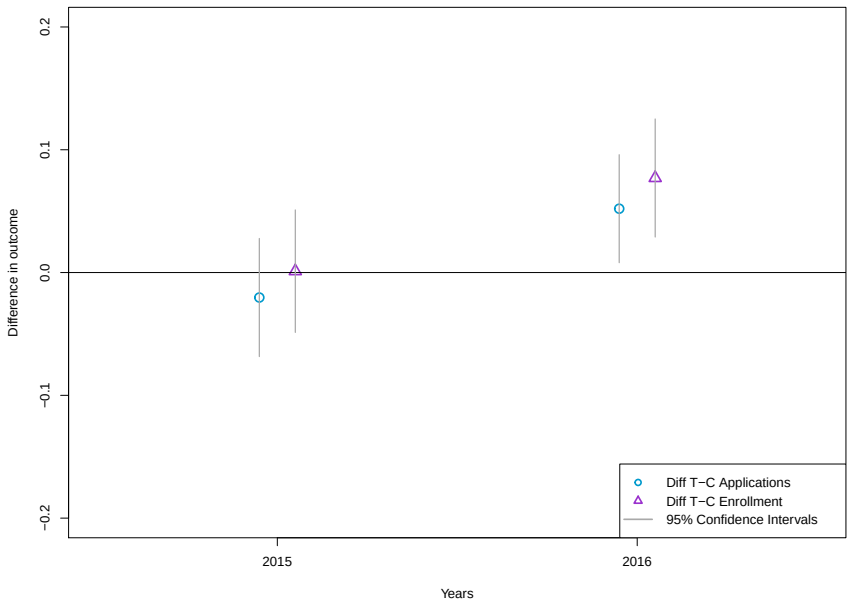
Generalization Bandwidth [-M$500,M$301]

95% CI in squared parenthesis.

Statistical significance at *: 10%, **: 5%, and ***: 1%



(a) Difference in outcomes for all students in the generalization bandwidth



(b) Difference in outcomes for matched students in the generalization bandwidth using representative template matching

Figure 14: Difference in applications and enrollment for students in the treated and control group for the generalization bandwidth before and after FHE

Sensitivity Analysis for Hidden Bias

To assess the sensitivity of the previous results to hidden bias, Rosenbaum bounds are estimated for the effects on application and enrollment previously found (Rosenbaum 1987, 2002, 2015). In particular, the adapted method proposed by Keele et al. (2019) for a difference-in-differences approach is used. These bounds quantify the change in probability of assignment between the treatment and control group that would need to occur due to an unmeasured confounder to qualitatively change the interpretation of the effects found. More specifically, the parameter Γ_c represents the ratio between the probability of assignment to treatment between the treatment and control group that would explain away the results that were previously found in application and enrollment rates.

The approach proposed by Keele et al. (2019) in a difference-in-difference setting for binary outcome switches the use of the sensitivity parameter Γ to Γ^2 to account for the fact that results derive from matched quadruplets, and not pairs. Table 7 shows the results for both outcomes.

Table 7: Results for sensitivity analysis for hidden bias for GRD estimates on application and enrollment to university

	Application to University	Enrollment to University
Γ_c	1.640	1.616
$\Pr(Z_{i1} = 1)$	0.622	0.618
$\Pr(Z_{i1} = 0)$	0.378	0.382

Γ_c: Critical parameter for unobserved bias.

$\Pr(Z_{i1})$ represents how probabilities should change for assignment to $\{0, 1\}$ to change qualitative results

The sensitivity parameter Γ relates to the probabilities of assignment to treatment ($P(Z_i = 1)$) and control ($P(Z_i = 1)$) as following:

$$P(Z_i = 0) = \frac{\Gamma_c - 1}{\Gamma_c}$$

$$P(Z_i = 1) = \frac{1}{\Gamma_c}$$

This means that if there is an unobserved confounder that is not accounted for, then such confounder would need to shift the probabilities of assignment between treatment and control to 0.62 vs. 0.38 in order to explain away any significant effect at a 5% significance level for both application and enrollment. The main advantage of a sensitivity analysis to hidden bias is that it provides a quantifiable measure of how probabilities would have to shift so that the qualitative conclusions of the study change. In this case, even thought the results for application and enrollment are moderately sensitive to potential hidden bias, if such bias is associated with the running variable, it should work in favor of the effect, so it is unlikely that is coming in this context from variables associated to income.

5 Conclusions

Regression discontinuity designs provide a robust setting for estimating local average treatment effects in observational studies. However, their results might be of skewed policy usefulness given that estimates are limited to the population just around the cutoff score.

In this paper, it is shown that is possible to generalize regression discontinuity designs for a broader population of interest. By using a representative template matching approach, in contrast to other parametric methods, the relationship between the running variable and the outcome of interest can be explained away for a generalization bandwidth away from the cutoff by leveraging pre-intervention data to inform the selection of such bandwidth.

Even though the method proposed requires additional data to inform the selection of the external validity bandwidth, it provides a sound counter-factual under many policy settings. Additionally, because matching is used instead of other adjusting methods to explain away the relationship between the running variable and the outcome, there is no need to rely on extrapolation to sustain the breakage of this link; even though it limits the population which can be generalized for, both treatment and control populations are balanced

in observed characteristics under the balancing constrains imposed by the researcher in a self-weighted sample.

The proposed method also lends itself to multiple extensions that can be useful for policy, such as two-dimensional regression discontinuity designs, or potential selection at either side of the cutoff. The use of matching also allows for the straightforward implementation of sensitivity analysis to hidden biases, as proposed by Keele et al. (2019). Finally, this method allows for the estimation of a target average treatment effect for a population of interest, leaving it up to the researcher to select the sample they wish to make inference on, and increasing the potential usefulness of the method for policy-based evidence.

References

Angrist, J., & Rokkanen, M. (2015). Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff. *Journal of the American Statistical Association*, *110*(512), 1331-1344.

Barr, N. (2003). Financing higher education: Comparing options. *London School of Economics*.

Bennett, M., Vielma, J., & Zubizarreta, J. (2019). Building representative matched samples with multi-valued treatments in large observational studies. *Working Paper - Columbia University*.

Bertanha, M., & Imbens, G. (2019). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*.

Bucarey, A. (2018). Who pays for free college? crowding out on campus. *Job Market Paper, MIT*.

Calonico, S., Cattaneo, M., Farrel, M., & Titiunik, R. (2018). rdrobust: obust data-driven statistical inference in regression-discontinuity designs. *R package version 0.99.4*.

Calonico, S., Cattaneo, M., & Farrell, M. (2019). `nprobust`: Nonparametric kernel-based estimation and robust bias-corrected inference. *arXiv working paper*.

Cattaneo, M., Frandsen, B., & Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference*, *3*(1), 1–24.

Cattaneo, M., Keele, L., & Titiunik, R. (2018). Extrapolating treatment effects in multi-cutoff regression discontinuity designs. *arXiv working paper*.

Chicoine, L. (2017). Homicides in mexico and the expiration of the u.s. federal assault weapons ban: A difference-in-discontinuities approach. *Journal of Economic Geography*, *17*, 825–856.

Dee, T. (2004). Are there civic returns to education? *Journal of Public Economics*, *88*, 1697-1720.

Dynarski, S., & Scott-Clayton, J. (2013). Financial aid policy: Lessons from research. *NBER Working Paper Series*.

Gelman, A., & Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, *37:3*, 447–456.

Grembi, V., Nannicini, T., & Troiano, U. (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics*, *8*(3), 1–30.

Hastings, J., Neilson, C., Ramirez, A., & Zimmerman, S. (2015). (un)informed college and major choice: Evidence from linked survey and administrative data. *Journal of Human Resources*, 126-151.

Hastings, J., Neilson, C., & Zimmerman, S. (2014). Are some degrees worth more than others? evidence from college admission cutoffs in chile. *NBER Working Paper Series*.

Heckman, J., Humphries, J., & Veramendi, G. (2016). Returns to education: The causal effects of education on earnings, health and smoking. *NBER Working Paper Series*.

Imbens, G. (2015). Matching methods in practice: Three examples. *Journal of Human Resources*, *50*(2), 373-419.

Keele, L., Small, D., Hsu, J., & Fogarty, C. (2019). Patterns of effects and sensitivity analysis for differences-in-differences. *arXiv Working Paper*.

Keele, L., Titiunik, R., & Zubizarreta, J. (2015). Enhancing a geographic regression discontinuity design through matching to estimate the effect of ballot initiatives on voter turnout. *Journal of the Royal Statistical Society: Statistics in Society Series A*, *178*(Part 1), 223–239.

Lee, D. (2008). Randomized experiments for non-random selection in u.s. house elections. *Journal of Econometrics*, *142*, 675–697.

Lee, D., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*, 281-355.

Manski, C. (2013). Public policy in an uncertain world: Analysis and decisions. *Harvard University Press*.

MINEDUC. (2016). Gratuidad. *Retrieved from http://www.gratuidad.cl*.

Pimentel, S., Kelz, R., Silber, J., & Rosenbaum, P. (2015). Large, sparse optimal matching with refined covariate balance in an observational study of the health outcomes produced by new surgeons. *Journal of the American Statistical Association*(110), 515-5127.

Rodriguez, J., Urzua, S., & Reyes, L. (2016). Heterogeneous economic returns to postsecondary degrees: Evidence from chile. *Journal of Human Resources*, *51*(2), 416-460.

Rokkanen, M. (2015). Exam schools, ability, and the effects of affirmative action: Latent factor extrapolation in the regression discontinuity design. *Working Paper, Columbia University*.

Rosenbaum, P. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, *74*, 13–26.

Rosenbaum, P. (2002). *Observational studies*. Springer.

Rosenbaum, P. (2015). Two `R` packages for sensitivity analysis in observational studies. *Observational Studies*, *1*(1), 1–17.

Rosenbaum, P., & Silber, J. (2001). Matching and thick description in an observational study of mortality after surgery. *Biostatistics*, *2*, 217-232.

Rubin, D. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, *2*(2), 808-840.

Silber, J., Rosenbaum, P., Ross, R., Ludwig, J., Wang, W., Niknam, B., ... Fleisher, L. (2014). Template matching for auditing hospital cost and quality. *Health Services Research*, *49*(5), 1446-1474.

Wing, C., & Bello-Gomez, R. (2018). Regression discontinuity and beyond: Options for studying external validity in an internally valid design. *American Journal of Evaluation*, *39*(1), 91-108.

Wing, C., & Cook, T. (2013). Strengthening the regression discontinuity design using additional design elements: A within-study comparison. *Journal of Policy Analysis and Management*, *32*, 853-877.

Zubizarreta, J. R., Kilcioglu, C., & Vielma, J. P. (2018). designmatch: Matched samples that are balanced and representative by design. *R package version 0.3*. (`https://cran.r-project.org/package=designmatch`)