Regression Discontinuity Designs with Multiple Running Variables

Alden Cheng

April 29, 2020

Abstract

In this paper, I introduce a new estimator for regression discontinuity designs with multiple running variables. My estimator provides efficiency gains relative to the common empirical practice of analyzing each running variable separately. In addition, it can be used to estimate heterogeneous treatment effects over a subset of the running variable space. I derive Bayesian confidence intervals for my estimator, and confirm their validity in simulations. Finally, I demonstrate the performance of my estimator in an empirical application from Londoño-Vélez, Rodríguez, and Sánchez (2020), which studies the effect of a large financial aid program on higher education in Colombia.

The regression discontinuity (RD) design, first introduced by Thistlethwaite and Campbell (1960), has enjoyed a revival in popularity over the past two decades. As Lee and Lemieux (2010) document, the RD design has been used in a wide range of policy evaluations, in areas such as education, labor market programs, health and crime.

In a RD design, treatment status is determined by whether an assignment variable passes a threshold. Under the assumption that the location of observations near this cutoff is as-good-as-random, the treatment effect is identified as the difference in mean outcomes for observations just above and below the cutoff. The multitude of programs with threshold-based eligibility criteria, as well as the transparency of the identification strategy, have both contributed greatly to the popularity of the RD design.

In addition to the common single-dimensional RD design, there are also many instances where treatment is determined by more than one assignment

I thank Alberto Abadie, Josh Angrist, David Autor, David Card, Jonathan Cohen, Amy Finkelstein, Jetson Leder-Luis, Juliana Londoño-Vélez, and all participants of MIT's labor lunch for their comments and suggestions. I am especially grateful to David Card for all his support and advice on this paper ever since its genesis as an undergraduate thesis, and Juliana Londoño-Vélez for generously agreeing to share her data on the Ser Pilo Paga program in Colombia. All mistakes are my own.

variable. I will call such designs multidimensional RD (MRD) designs. For example, eligibility for benefit programs often depend on multiple criteria, and in many school systems students need to achieve a passing score in each of several subject tests in order to move onto the next grade level. MRDs fall under two general categories – cases with dichotomous treatments (the two treatment conditions being either treatment or control), and those with multiple treatments (i.e. more than two mutually exclusive treatment conditions). Throughout this paper, I will focus my discourse on the case with dichotomous treatment, although all of the analysis extends straightforwardly to the case with multiple treatment arms as I describe in the Appendix A2. As with single-dimensional RD designs, treatment effects in MRD designs are identified by the difference between mean outcomes for observations on either side of the boundary separating treated and untreated observations.

When confronted with a MRD design, most empirical papers have resorted to analyzing each running variable separately.[1] In this paper, I present an estimation approach based on thin plate regression splines that has two key advantages over this empirical practice. Specifically, my estimator achieves efficiency gains, and recovers heterogeneous treatment effects as functions of the running variables (over a subset of the running-variable-space). I then derive confidence intervals for my estimator using Bayesian inference. In simulations, I show that my estimator does indeed result in more precise estimates, that it is able to capture heterogeneous treatment effects, and that the Bayesian confidence intervals provide valid coverage. Finally, I apply my estimator to an empirical application from Londoño-Vélez, Rodríguez, and Sánchez (2020) on the effect of a large financial aid program in Colombia. Once again, my estimator provides efficiency gains in this setting, and in addition, my estimates of heterogeneous treatment effects yield economically interesting insights.

The rest of this paper is organized as follows. Section 1 introduces single-dimensional RD as well as MRD, and outlines the assumptions that underpin these RD designs. Section 2 discusses current estimation approaches, and proposes a novel estimation method which uses thin plate regression splines. Section 3 presents simulation results comparing this estimator with methods typically used in empirical papers. In Section 4, I demonstrate the utility of my estimator in an empirical application. Section 5 concludes.

1 Background and Notation

The defining feature of the RD design is that there is some running variable X_i (also known as an assignment variable) that determines a binary treatment W_i depending on whether it crosses a threshold. For MRD, there is more than one running variable, and throughout this paper I will assume just two for simplicity,

[1] Matsudaira (2008), and Londoño-Vélez, Rodríguez, and Sánchez (2020) are just two of many examples of this.

so that $X_i = (X_{1i}, X_{2i})'$.[2] Without loss of generality, I will assume the threshold to be zero, so that in the case of sharp MRD, we have:[3]

$$W_i = \mathbb{I}\left[X_{1i} \geq 0\right] \cdot \mathbb{I}\left[X_{2i} \geq 0\right]. \tag{1}$$

Adopting the usual potential outcomes notation, let $Y_i(1)$ and $Y_i(0)$ denote the potential outcomes for individual i when she is treated and not treated respectively. Also, denote the treated and untreated regions in running variable space by $R_1 \equiv \{x_i | x_{1i} \geq 0, x_{2i} \geq 0\}$ and $R_0 \equiv \{x_i | x_{1i} < 0 \text{ or } x_{2i} < 0\}$ respectively, and call $\mathbb{F} \equiv \left\{x \in \mathbb{R}^2 | x_{1i} \cdot x_{2i} = 0\right\}$ the treatment frontier. The identifying assumption for RD designs is that mean potential outcomes are continuous at the treatment frontier.[4] Denoting an open ball of radius ϵ centered at x by $B_\epsilon(x)$, this assumption can be written as:

$$\lim_{\epsilon \to 0} \mathbb{E}\left[Y_i(w) | X_i = x', x' \in B_\epsilon^1(x)\right] = \lim_{\epsilon' \to 0} \mathbb{E}\left[Y_i(w) | X_i = x', x' \in B_{\epsilon'}^0(x)\right], \tag{2}$$

for $w \in \{0, 1\}$, for all $x \in \mathbb{F}$, and $B_\epsilon^w(x) \equiv B_\epsilon(x) \cap R_w$. Under this assumption, the treatment effect at any point on the treatment frontier, $x \in \mathbb{F}$, is identified:

$$\begin{aligned}
\tau(x) &\equiv \mathbb{E}\left[Y_i(1) - Y_i(0) | X_i = x\right] \\
&= \lim_{\epsilon \to 0} \mathbb{E}\left[Y_i(1) | X_i = x', x' \in B_\epsilon^1(x)\right] - \lim_{\epsilon' \to 0} \mathbb{E}\left[Y_i(0) | X_i = x', x' \in B_{\epsilon'}^0(x)\right] \\
&= \lim_{\epsilon \to 0} \mathbb{E}\left[Y_i | X_i = x', x' \in B_\epsilon^1(x)\right] - \lim_{\epsilon' \to 0} \mathbb{E}\left[Y_i | X_i = x', x' \in B_{\epsilon'}^0(x)\right]
\end{aligned} \tag{3}$$

As equation (3) shows, the MRD treatment effect is identified as the difference between two limits of the conditional expectation function (CEF): the limit along a sequence in the treated region minus the limit along a sequence in the untreated region. This suggests a natural way to estimate the treatment effect. Specifically, I will estimate the two CEFs – $g_1(x)$ and $g_0(x)$ – using only observations in the treated and untreated regions respectively, and take the difference to obtain an estimate of the treatment effect at each point x. One can then average these estimates over the treatment frontier to obtain the average treatment effect over the treatment frontier.[5] Details concerning the estimation method are covered in the next section.

[2] The framework and estimation extend to cases with more than two running variables, although the estimation is likely to work less well in practice due to the curse of dimensionality.

[3] While I have described treatment as determined by an "AND" condition here, this is without loss of generality in the sense that cases where the treatment is assigned by an "OR" condition (and/or one or both of the running variables have to fall below the threshold) can be transformed into the formulation above by appropriately redefining the treatment (and/or switching the sign(s) of the running variable(s)). Even more generally, the estimator in this paper can be applied as long as connected regions of the running variable space determine the treatment that each individual receives.

[4] It is common to strengthen this assumption to continuity of mean potential outcomes over the entire running variable space, since it is difficult to think of a plausible scenario where this assumption is suspect but continuity at the treatment frontier is not.

[5] Alternatively, if one is interested in the average treatment effect over a subset of the treatment frontier, one can also average the treatment effect estimates over the subset of interest.

The first advantage of this approach is that it yields more precise estimates relative to the common empirical approach of analyzing each running variable separately. In particular, by focusing only on the set of observations $\{i : X_{1i} \geq 0\}$ (or alternatively, $\{i : X_{2i} \geq 0\}$) and estimating a single-dimensional RD based on the running variable X_{2i} (or X_{1i}), one is discarding a (potentially informative) subset of observations for this estimation. On the other hand, the MRD estimate of the average treatment effect uses *all* of the observations, which results in efficiency gains.[6]

The second advantage of this framework is that it allows us to estimate heterogeneous treatment effects as a function of the running variables for values of the running variables in the treatment frontier. This is especially useful when economic theory suggests that the treatment effect may vary as a function of the running variables.

2 MRD Estimation

2.1 Literature on MRD

Most empirical papers have dealt with MRD designs by analyzing each running variable separately. The few empirical papers that take an alternative approach have tended to estimate the CEFs parametrically.

Papay, Willett and Murnane (2011) suggest parameterizing the CEF as a linear function of the running variables and their interaction, and to choose the bandwidth via a form of cross-validation that places greater weight on observations close to the treatment frontier.[7] However, this choice of functional form is very restrictive and may not result in a good fit in many instances.

Dell (2010) parameterizes the CEF in a MRD design as a cubic function of the two running variables. However, the use of high-order polynomials for RD estimation is typically regarded as a bad idea. This is because high-order polynomials often have poor performance at the boundaries, and estimates at the boundaries tend to be affected by observations far away (Gelman and Imbens, 2019). These issues with high-order polynomials are exacerbated in the multidimensional case.

A rare example of an empirical MRD paper that estimates the CEFs nonparametrically is Snider and Williams (2015), who use local linear regressions for estimation. Yet, the key tuning parameter for local linear regressions – the bandwidth – is chosen in an ad hoc manner. Given the importance of the bandwidth choice for local linear regressions, this is a significant limitation of their approach.

Econometricians have tended to focus on nonparametric estimation methods for MRD. Zajonc (2012) extends the influential framework for bandwidth selection in local linear regressions by Imbens and Kalyanaraman (2012), henceforth

[6] The exact magnitude of these efficiency gains will depend on the distribution of the running variables.

[7] Although their paper considers MRD with multiple treatment arms, the discussion applies to MRD with dichotomous treatments as well.

IK, to multiple dimensions. Essentially, this approach uses local linear regressions to estimate the CEFs, with a bandwidth choice that seeks to minimize the mean-squared error (MSE) of the MRD estimate. However, there are several complications in this extension to multiple dimensions. First, one has to estimate numerous unknown functionals of the underlying data-generating process (DGP) to estimate the optimal bandwidth, which is substantially more challenging in the multidimensional case compared to the single-dimensional setting.[8] Second, the optimal bandwidth is "too large" in the sense that there will be non-negligible bias in the asymptotic distribution of the estimate, which may result in incorrect coverage for the confidence intervals. Finally, one ends up with a different bandwidth choice along each point on the treatment frontier, and Zajonc suggests using the minimum of these without any rigorous justification.

Imbens and Wager (2018) suggest an estimation approach that may be applied to MRD. However, this approach requires the user to specify bounds on the second derivative of the CEFs and the conditional variance functions,[9] quantities for which an empirical researcher may not have good intuition for, especially when the CEFs are multidimensional. In addition, the authors note that if this method is used to estimate heterogeneous treatment effects, it is likely to result in "relatively long confidence intervals".

2.2 Estimation Using Thin Plate Regression Splines

As mentioned in Section 1, MRD estimation boils down to the estimation of the CEFs g_1 and g_0. The choice of estimator should be guided by two principles. First, the estimate should not depend heavily on points far away from the treatment frontier $\mathbb{F}$. Second, the estimate should be relatively flexible, but should not overfit the data.

The approach I take is to estimate the CEFs using thin plate regression splines (TPRS), introduced by Wood (2003). This is a local method, which ensures that estimates near $\mathbb{F}$ will not be very sensitive to points far away. Moreover, the flexibility of the estimate is regularized by a scalar penalty term which penalizes more flexible functional forms. Since this tuning parameter is a scalar, it is far easier to implement than multidimensional local linear regressions, which essentially require the choice of a continuum of tuning parameters. Before elaborating on TPRS, I will start by introducing thin plate splines (Duchon, 1977),

[8] Using this approach for MRD would essentially require estimating a continuum of nuisance parameters along the treatment frontier. Moreover, these nuisance parameters tend to be local quantities (e.g. the second derivatives of the CEFs at every point along the treatment frontier), so that only observations close to a point $x \in \mathbb{F}$ will be useful for estimating the nuisance parameters associated with that point. To see why this is a practical concern, it might be useful to compare this approach for a single-dimensional and multidimensional RD design with the same number of observations. Not only are there many more nuisance parameters to estimate in the multidimensional case, the effective number of observations used to estimate each of these nuisance parameters is also much smaller. It is unclear whether all of these nuisance parameters in the MRD case can be estimated reliably for a modestly sized empirical dataset.

[9] More precisely, in the context of MRD, they require bounds on the operator norm of the Hessian of the CEFs and the conditional variance functions.

since TPRS is essentially just a more computationally tractable version of this.

Consider the following DGP with $x_i \in \Omega$ (an open bounded subset of $\mathbb{R}^d$) and i.i.d. error terms ϵ_i:

$$y_i = l(x_i) + \epsilon_i,\ i = 1, ..., N,$$

where l belongs to the Sobolev space $H^m(\Omega)$. Denote the space of Schwartz distributions by $\mathcal{D}'(\mathbb{R}^d)$, and let $D^{-m}L^2(\mathbb{R}^d) = \{g \in \mathcal{D}'(\mathbb{R}^d) : D^\alpha g \in L^2(\mathbb{R}^d), |\alpha| = m\}$. Then, the thin plate spline is the function in $D^{-m}L^2(\mathbb{R}^d)$ that minimizes:

$$\sum_{i=1}^{N} (y_i - g(x_i))^2 + \lambda J_{md}(g),$$

where we require $2m > d$, and the penalty term J_{md} is defined by:

$$J_{md} \equiv \int ... \int_{\mathbb{R}^d} \sum_{\nu_1 + ... + \nu_d = m} \frac{m!}{\nu_1! ... \nu_d!} \left(\frac{\partial^m g}{\partial x_1^{\nu_1} ... \partial x_d^{\nu_d}} \right)^2 dx_1 ... dx_d.$$

In the most common setting for MRD, where we have $d = 2$ running variables, we may choose the penalty order to be $m = 2$. In this case, the penalty term becomes:

$$J_{22} = \int \int \left(\frac{\partial^2 g}{\partial x_1^2} \right)^2 + 2 \cdot \left(\frac{\partial^2 g}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 g}{\partial x_2^2} \right)^2 dx_1 dx_2.$$

We also need to choose the smoothing parameter λ, which entails a bias-variance tradeoff. This is often implemented via generalized cross-validation (GCV), a modification of leave-one-out-cross-validation (LOOCV) that has several advantages.[10] First, it is less computationally expensive than LOOCV, and second, it is invariant to rotation of the outcome vector and basis matrix. Further details on thin plate splines can be found in an excellent overview by Wahba (1990).

A disadvantage of thin plate splines is its computational cost for $d > 1$ dimensions. In particular, while an efficient $O(N)$ algorithm exists for $d = 1$, computational costs for $d \geq 2$ is generally of order $O(N^3)$. To deal with this issue, I use an approximation to thin plate splines suggested by Wood (2003), called thin plate regression splines (TPRS). The key idea is to use a basis matrix of rank k in this approximation, instead of a basis matrix of rank N as in thin plate splines. This reduces the computational cost of fitting a TPRS to $O(kN^2)$. Further details on the precise definition of TPRS, as well as the numerical methods that yield these computational savings can be found in the original paper by Wood.[11]

[10] Let $\boldsymbol{A}(\lambda)$ denote the influence matrix for the model fit using λ. Then the GCV score is given by:

$$GCV(\lambda) = \frac{N||\boldsymbol{y} - \boldsymbol{A}(\lambda)||^2}{[N - tr(\boldsymbol{A}(\lambda))]^2}.$$

[11] My implementation of TPRS estimation is based on the "mgcv" package in R, maintained by Simon Wood.

2.3 Inference

I conduct inference for my TPRS estimates using a Bayesian approach. Specifically, I adopt a normal prior for the TPRS parameters, which yields a normal posterior for the TPRS estimates. Since the treatment effect $\tau(x)$ is a linear function of the TPRS parameters, this implies that the estimate $\hat{\tau}(x)$ is also normally distributed. I then take the standard approach to form CIs for $\tau(x)$ of level α:

$$\left[\hat{\tau}(x) - q_{1-\alpha/2} \cdot \hat{se}\left(\hat{\tau}(x)\right), \hat{\tau}(x) + q_{1-\alpha/2} \cdot \hat{se}\left(\hat{\tau}(x)\right)\right],$$

where $q_{1-\alpha/2}$ denotes $(1-\alpha/2)$th quantile of a standard normal distribution, and $\hat{se}$ is the estimate of the standard error based on the Bayesian approach above.

I can also estimate the average treatment effect τ over the frontier $\mathbb{F}$ using the formula:

$$\hat{\tau} = \sum_{p=1}^{P} \frac{\hat{f}(x_p)}{\sum_{p'=1}^{P} \hat{f}(x_{p'})} \hat{\tau}(x_p),$$

where $x_1, ..., x_P$ is a grid of points along $\mathbb{F}$, and $\hat{f}$ is an estimate of the density of the running variables, f. The delta method can then be used to obtain an estimate of the standard error $\hat{se}\left(\hat{\tau}\right)$, and we can form a CI for τ using:

$$\left[\hat{\tau} - q_{1-\alpha/2} \cdot \hat{se}\left(\hat{\tau}\right), \hat{\tau} + q_{1-\alpha/2} \cdot \hat{se}\left(\hat{\tau}\right)\right].$$

More generally, this same methodology can be used to estimate the average treatment effect over other subsets of the treatment frontier, and to obtain a standard error for the estimate. More details on the calculations for the standard errors can be found in Appendix A1.

3 Simulations

In this section, I present simulation results on the MRD estimator. These simulations aim to accomplish the following goals. First, I provide evidence that the MRD estimator produces reasonable estimates as well as valid CIs. Second, I show that the MRD estimator tends to outperform estimators commonly used in the empirical literature based on single-dimensional methods applied to the MRD problem. Specifically, for the single-dimensional methods, I consider IK, CCT, as well as a method proposed in Kolesár and Rothe (2018) based on an assumed bound for the second derivative, henceforth KR (BSD).

For my simulations, I consider variants of the following data-generating process (DGP) based on a fifth order polynomial:

$$Y = \begin{cases} \sum_{p+q\leq 5} a_{p,q} X_1^p X_2^q + \tau(X_1, X_2) + \epsilon & X_1 \geq 0, X_2 \geq 0, \\ \sum_{p+q\leq 5} a_{p,q} X_1^p X_2^q + \epsilon & \text{otherwise.} \end{cases}$$

The error terms ϵ are distributed i.i.d. standard normal, and the running variables X_1 and X_2 are each drawn independently from a $2 \cdot Beta(3,3) - 1$ distribution, following the spirit of the simulations in IK. This distribution ensures that the running variables have support on $[-1,1]^2$, and since the marginal distribution is symmetric about zero, about three-quarters of the observations are untreated. I consider the two versions of this DGP, with either constant treatment effects where I set $\tau(X_1, X_2) = 0.5$, or heterogeneous effects where I let $\tau(X_1, X_2) = 0.5 + X_1 - X_2$. Figures 1 and 2 show the CEFs for these two different DGPs. For each different DGP, I run 100 simulations, each with 10,000 observations, and use Bayesian standard errors for MRD inference.

Figure 1: CEF for DGP with Constant Treatment Effects

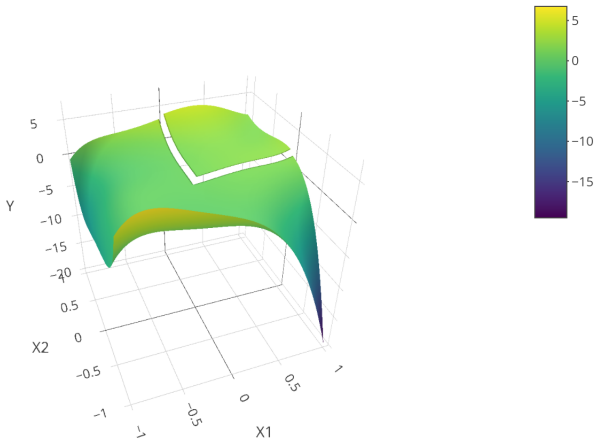
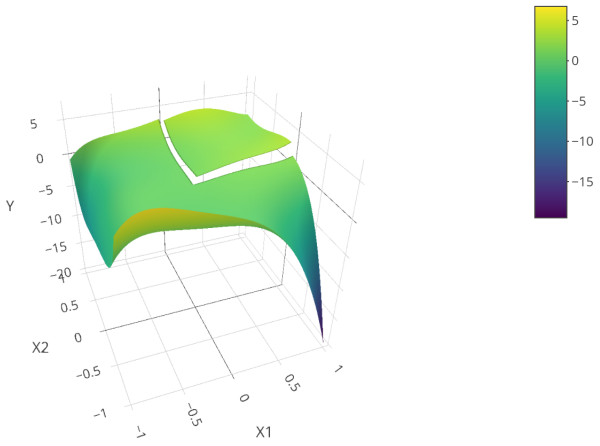


Figure 2: CEF for DGP with Heterogeneous Treatment Effects



The results for the DGP with constant treatment effects are shown in Table 1. We would expect the single-dimensional methods to perform well in this case, since they are restricted to estimating constant treatment effects by design. The results in Panels (a) and (b) correspond to the average treatment effects over the two segments of the frontier $\mathbb{F}$: the positive X_2 axis and the positive X_1 axis respectively, since the single-dimensional estimators yield estimates corresponding to these two subsets of $\mathbb{F}$. For the MRD estimator, I estimate $\tau(x)$ at 10 equally spaced grid points along the positive X_2 axis, and similarly for the positive X_1 axis (ranging from zero to the largest observed value of the relevant running variable in each realization of the DGP).

We observe in the first column that the bias for all of the estimators are very small, and the same can be said for the MSE in the second column. In the third column, we see that the 95 percent CIs for all estimators have roughly the correct coverage. The final column shows that the MRD CIs are roughly 5 to 30 percent shorter than those of the other estimators. These efficiency gains are due to the fact that the MRD estimator uses all of the data for estimation simultaneously, whereas the single-dimensional methods uses only about half of the data when estimating the treatment effect in Panel (a) and likewise for Panel (b).

Table 1: Simulation Results for DGP with Constant Treatment Effects ($\tau = 0.5$)

Panel A. Estimates of the Average Treatment Effect Over {X_1=0, X_2≥0}				
Estimator	Bias	MSE	Coverage	CI Length
IK	0.002	0.009	0.97	0.334
CCT	-0.020	0.016	0.87	0.388
KR	0.001	0.009	0.99	0.419
MRD	-0.026	0.006	0.97	0.319
Panel B. Estimates of the Average Treatment Effect Over {X_1≥0, X_2=0}				
Estimator	Bias	MSE	Coverage	CI Length
IK	-0.005	0.010	0.93	0.350
CCT	0.017	0.012	0.92	0.389
KR	-0.007	0.008	0.98	0.406
MRD	-0.013	0.006	0.95	0.319

Notes: The IK estimator is based on local linear regression with bandwidth selection according to IK (2012). The CCT estimator is based on local linear regression with bandwidth selection and bias correction according to CCT (2014). The KR estimator is based on the method introduced in KR (2018) with an assumption on the bound for the second derivative of the CEF. The MRD estimator is the estimator introduced in this paper. The results shown in this table are based on 100 realizations of the DGP with constant treatment effects. See text for more details on these simulations.

Simulation results for the DGP with heterogeneous treatment effects are shown in Table 2. There is no natural way to estimate heterogeneous treatment effects using the single-dimensional methods. However, a possible way to do so is to apply them separately to different subsets of the treatment frontier, which I do in these simulations in order to give these methods the best chance of success. Specifically, in Panel A, I apply these methods separately to $\{X_1 = 0, X_2 \in [0, c_1]\}, \{X_1 = 0, X_2 \in (c_1, c_2]\}, ..., \{X_1 = 0, X_2 \in [c_9, c_{10}]\}$, where c_{10} is the largest value of X_2 observed in that particular realization of the DGP, and this yields 10 different estimates of τ over of these subsets, allowing us to estimate heterogeneous treatment effects. I estimate the heterogeneous treatment effects using the analogous method for the single-dimensional estimators in Panel B. The row corresponding to the CCT estimator is left empty because its default implementation fails in most of the realizations of this DGP. This happens because when we cut the data so finely, the sample size turns out in many cases to be too small for the default implementation of the CCT method to estimate the optimal bandwidth, and/or the bias correction term and its variability.

Columns 1 and 2 of Table 2 again show that the bias and MSE for all of the estimators seem relatively small.[12]Turning next to the performance of the 95 percent pointwise CIs of the various estimators, we see an even more striking

[12]The bias (or IMSE) in these simulations are computed as the weighted average of the difference (or squared difference) between the treatment effect estimate over a subset of the

difference between the performance of the MRD CIs and those of the other estimators. The third column shows that the coverage of the MRD CIs is close to 95 percent, whereas those for the IK and KR (BSD) CIs are far below 95 percent, at between 50 and 82 percent. This is despite the fact that the MRD CIs are about half the length of the other CIs, as shown in the last column. The poor coverage properties of the IK and KR (BSD) CIs may be due to a small sample issue – when cutting the data so finely to estimate heterogeneous treatment effects, the asymptotic approximations which the CIs are based on perform badly due to the small sample size.

Table 2: Simulation Results for DGP with Heterogeneous Treatment Effects $(\tau(X_1, X_2) = 0.5 + X_1 - X_2)$

Panel A. Estimates of the Treatment Effect Over {X_1=0, X_2≥0}				
Estimator	Bias	IMSE	Coverage (Pointwise)	CI Length
IK	0.008	0.044	0.741	0.982
CCT	-	-	-	-
KR	0.019	0.029	0.821	1.052
MRD	-0.035	0.018	0.933	0.501
Panel B. Estimates of the Treatment Effect Over {X_1≥0, X_2=0}				
Estimator	Bias	IMSE	Coverage (Pointwise)	CI Length
IK	-0.010	0.043	0.507	0.998
CCT	-	-	-	-
KR	-0.020	0.032	0.577	1.071
MRD	-0.023	0.015	0.944	0.502

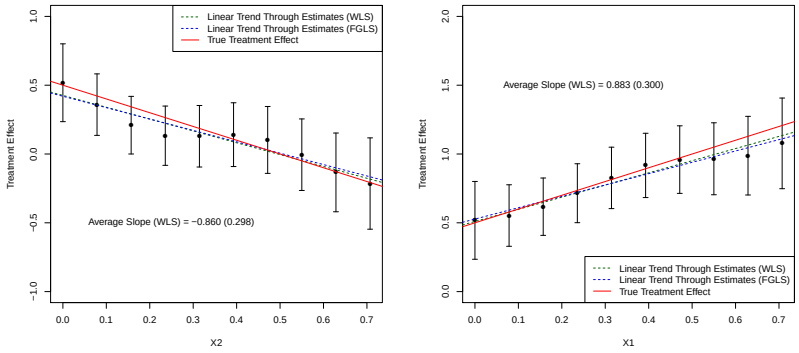
Notes: The IK estimator is based on local linear regression with bandwidth selection according to IK (2012). The CCT estimator is based on local linear regression with bandwidth selection and bias correction according to CCT (2014). The KR estimator is based on the method introduced in KR (2018) with an assumption on the bound for the second derivative of the CEF. The MRD estimator is the estimator introduced in this paper. The results shown in this table are based on 100 realizations of the DGP with heterogeneous treatment effects. The bias and IMSE in these simulations are respectively computed as the weighted average of the difference and weighted average squared difference between the treatment effect estimate over a subset of the treatment frontier and the true average treatment effect over the same subset. The weights are based on the density of the running variables over these subsets. The row corresponding to the CCT estimator is left empty because its default implementation fails in most of the realizations of this DGP. See text for more details on these simulations.

Finally, one may wonder whether the MRD estimates are able to capture the qualitative features of the heterogeneous treatment effects, namely that it is increasing in X_1 and decreasing in X_2. Figure 3 shows the average MRD point estimates as well as the average CIs over the 100 simulations (as solid black dots and bars respectively). The true treatment effect is shown as a solid red line. We see that the MRD estimates are indeed increasing in X_1 and decreasing in X_2, and that they correspond relatively closely to the true treatment effects. Moreover, if we plot a linear fit through the MRD estimates using weighted least squares (WLS) or feasible generalized least squares (FGLS), we see that

treatment frontier and the true average treatment effect over the same subset. The weights are based on the density of the running variables over these subsets.

the resulting fit (shown as a dashed green and blue lines for WLS and FGLS respectively) is very close to the true treatment effect, the slope coefficient is not statistically different from the true slope for the treatment effect.[13]

Figure 3: MRD Estimates of Heterogeneous Treatment Effects in the Simulations



4 Empirical Application

4.1 Institutional Setting

In this section, I present an empirical application based on the *Ser Pilo Paga* (SPP) program in Colombia. In particular, I extend the RD analysis in Londoño-Vélez, Rodríguez, and Sánchez (2020), henceforth LRS, to look at heterogeneous treatment effects as a function of the two running variables – test scores and family wealth.

The SPP is a merit-based financial aid program introduced in Colombia in 2014. Students who score above a threshold on a standardized high school test called the SABER 11, and whose families are poor enough (i.e., their SISBEN wealth index falls below a geography-dependent threshold), are eligible for financial aid if they enroll in a university with High Quality Accreditation. The SPP provides loans that are forgiveable upon graduation, as well as a biannual stipend while recipients attend college. Further details regarding the program can be found in LRS.

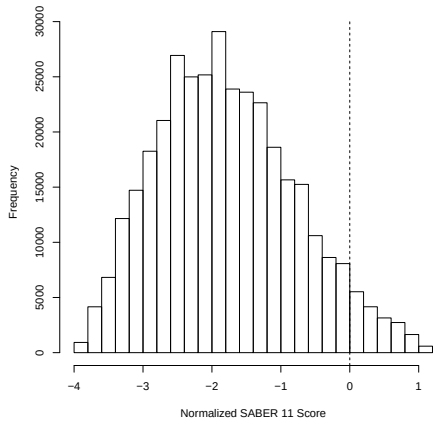
Figure 4 shows histograms of the running variables, which have been normalized so that they have standard deviation one, and zero corresponds to the cutoff. Note also that higher values of the normalized SISBEN wealth index corresponds to *poorer* households. From these histograms, we see that the SPP program is much more selective on the academic dimension than on the wealth

[13]The standard errors for the slope coefficients for the WLS fits are computed via bootstrap. The standard errors for the slope coefficients for FGLS fits are analytic standard errors.

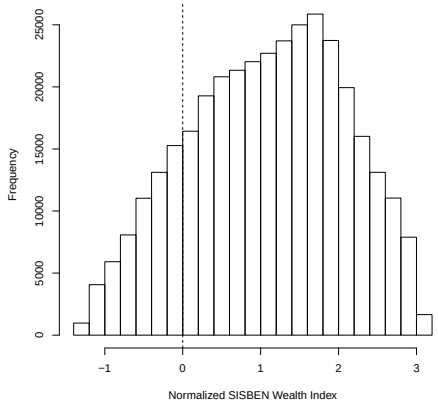
dimension, i.e. most students in the sample are poor enough to qualify, but relatively few score well enough on the SABER 11 test to do so. I drop observations with SABER 11 or SISBEN scores lower than the 1st percentile or greater than the 99th percentile (since this causes issues in estimating the running variables' density).

Figure 4: Histograms of Running Variables

(a) Normalized SABER 11 Score



(b) Normalized SISBEN Wealth Index



4.2 Average Treatment Effects at the SABER 11 and SISBEN Thresholds

Following LRS, my MRD analysis studies the effect of the SPP on enrollment patterns for the first cohort potentially eligible for the SPP. An advantage of focusing on this cohort is that the program was announced two months after they took the SABER 11, so there is little scope for test score manipulation.

Tables 3 and 4 show the estimated effects of the program on enrollment in various types of college, for students with eligible SISBEN scores and with SABER 11 scores close to the threshold, and vice versa, respectively. Panel A in these tables shows the MRD estimates of the treatment effects, whereas Panel B shows the original estimates from LRS, which were obtained by applying single-dimensional RD following the CCT method.

We observe that the MRD point estimates and the original point estimates from LRS are quite similar qualitatively. Eligibility for the SPP (based on SABER 11 and SISBEN scores) increases overall enrollment in college, and this effect is driven by high quality private institutions – in fact, eligibility for SPP decreases enrollment in low quality college. This pattern can be explained by the fact that the SPP applies only for institutions with High Quality Accreditation. We also see that the effects on enrollment (in any college, any high quality college, or any high quality private college) tend to be larger for students at the SABER 11 threshold than for students at the SISBEN threshold. This is because the analysis at the SABER 11 threshold focuses on students with qualifying SISBEN scores, who are on average much poorer than students at the SISBEN threshold. Hence, credit constraints may be more binding for the former set of students, which explains the larger effects on enrollment.

While the MRD point estimates are not very different from the original LRS estimates, the standard errors are rather different. Comparing Panel A in the two tables, we see that there are some efficiency gains for the estimates on the SABER 11 threshold, with the MRD standard errors being 10 to 30 percent smaller. Looking at Panel B in the two tables, we observe even more dramatic efficiency gains on the SISBEN threshold, where the MRD standard errors tend to be about 30 to 70 percent shorter. This is because the SISBEN threshold is set relatively low, so that when LRS restricts the sample to those with qualifying SISBEN scores for the analysis at the SABER 11 threshold, they do not lose much data. However, the SABER 11 threshold is much higher, so LRS lose most of their sample when restricting the sample to students with eligible SABER 11 scores for the analysis at the SISBEN threshold. This can be seen clearly by comparing the number of observations in Panel B of Tables 3 and 4 – the number of observations used for the LRS estimates on the SABER 11 threshold are more than 10 times the number of observations used for their estimates on the SISBEN threshold. On the other hand, the MRD estimates does not suffer from this loss of precision since it uses all of the data for estimation simultaneously.

Table 3: Average Treatment Effect Estimates for Students with Eligible SISBEN Scores and with SABER 11 Scores Close to the Threshold

Panel A. MRD Estimates							
		High Quality Institutions			Low Quality Institutions		
	Any	Any	Private	Public	Any	Private	Public
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Effect Estimate	0.329	0.464	0.469	-0.007	-0.133	-0.061	-0.073
	(0.008)	(0.009)	(0.008)	(0.005)	(0.006)	(0.004)	(0.005)
Number of Observations	349,015	349,015	349,015	349,015	349,015	349,015	349,015
Panel B. Original Estimates from LRS							
		High Quality Institutions			Low Quality Institutions		
	Any	Any	Private	Public	Any	Private	Public
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Effect Estimate	0.320	0.465	0.466	0.000	-0.154	-0.063	-0.087
	(0.012)	(0.012)	(0.011)	(0.007)	(0.011)	(0.007)	(0.009)
Number of Observations	299,475	299,475	299,475	299,475	299,475	299,475	299,475

Notes: Standard errors are shown in parentheses.

Table 4: Average Treatment Effect Estimates for Students with Eligible SABER 11 Scores and with SISBEN Scores Close to the Threshold

Panel A. MRD Estimates							
		High Quality Institutions			Low Quality Institutions		
	Any	Any	Private	Public	Any	Private	Public
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Effect Estimate	0.296	0.438	0.482	-0.035	-0.152	-0.077	-0.075
	(0.013)	(0.016)	(0.014)	(0.008)	(0.009)	(0.005)	(0.008)
Number of Observations	349,015	349,015	349,015	349,015	349,015	349,015	349,015
Panel B. Original Estimates from LRS							
		High Quality Institutions			Low Quality Institutions		
	Any	Any	Private	Public	Any	Private	Public
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Treatment Effect Estimate	0.274	0.396	0.477	-0.079	-0.12	-0.052	-0.076
	(0.027)	(0.024)	(0.020)	(0.018)	(0.022)	(0.015)	(0.016)
Number of Observations	23,132	23,132	23,132	23,132	23,132	23,132	23,132

Notes: Standard errors are shown in parentheses.

4.3 Treatment Effect Heterogeneity

In addition to the precision gains evident in the previous subsection, estimation based on the MRD framework also allows us to test for heterogeneous treatment effects. Specifically, in this subsection I show MRD estimates of whether

the effect of SPP eligibility on enrollment varies by academic ability or family wealth.

First, I present evidence on treatment effect heterogeneity as a function of academic ability, as proxied by SABER 11 test scores. A priori, one might expect that the treatment effect would be increasing in SABER 11 test scores, since the returns to higher education is likely greater for high-ability students. However, the MRD estimates in Figure 5 show the exact opposite, i.e., the effect of SPP eligibility on enrollment is decreasing in the SABER 11 test score.

A possible explanation for this pattern is a "ceiling effect", in that the students who are strongest academically tend to enroll in college regardless of whether they are eligible for financial aid. To shed light on whether this explanation is plausible, Figure 6 shows bin scatterplots of the enrollment rate (in any college), as well as a second degree local polynomial fit and 95 percent CIs, separately for students with eligible and ineligible SISBEN scores. We see here that the enrollment rate is increasing in SABER 11 test scores for both groups, as economic theory would predict, if higher ability students having greater returns to education. However, we also observe that this increase is much steeper for students with SISBEN scores that do not meet the SPP criteria, consistent with the "ceiling effect".

Figure 5: Effect of the SPP on Enrollment as a Function of SABER 11 Test Scores: Any College

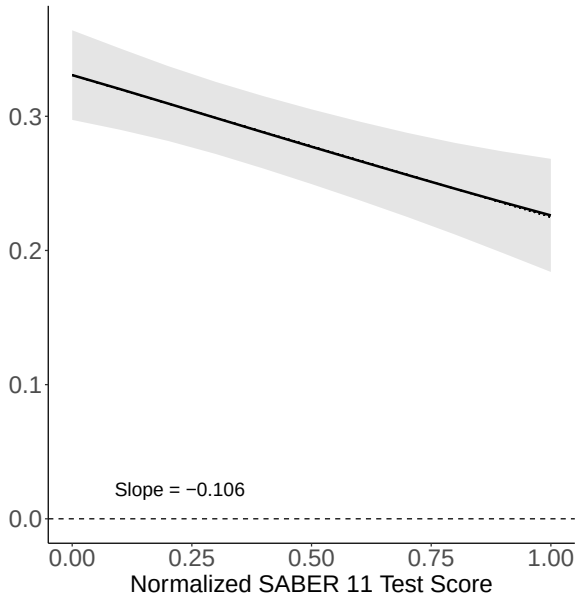


Figure 6: Enrollment Rates as a Function of SABER 11 Test Scores

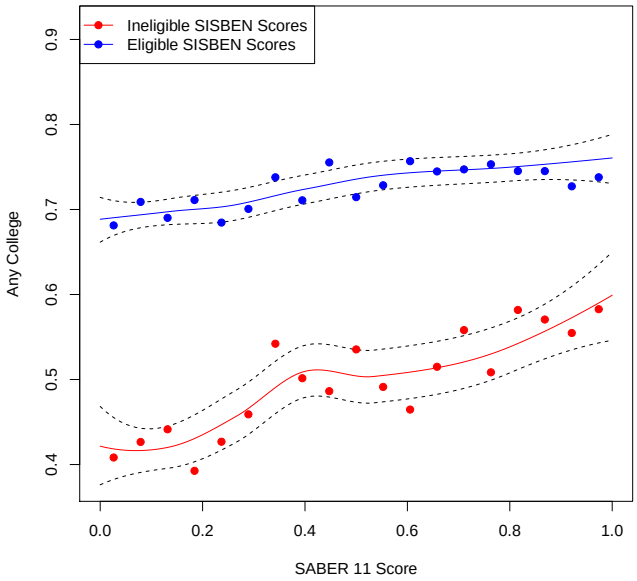
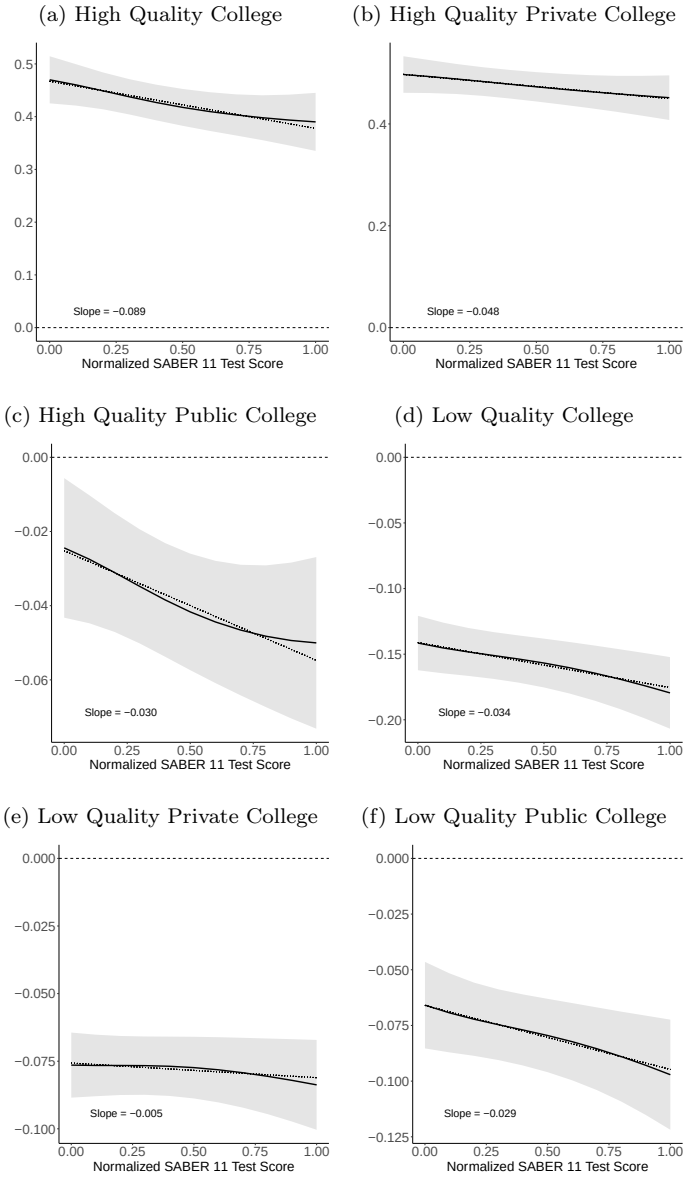


Figure 7 shows estimates of treatment effect heterogeneity where enrollment in different types of colleges are used as the dependent variable. For the two outcomes where the analysis on average effects yield similar results as the main outcome of enrollment in any college (specifically, enrollment in any high quality college, or enrollment in any high quality private college), we observe the same pattern of decreasing effect as a function of SABER 11 scores.

Figure 7: Effect of the SPP on Enrollment as a Function of SABER 11 Test Scores: Different Types of Colleges



Next, I consider treatment effect heterogeneity as a function of the SISBEN wealth index. A priori, we might expect the treatment effect to be increasing in the SISBEN score, since the SPP probably eases credit constraints to a greater degree for poorer families. On the other hand, as we saw in Figure 4, the cutoff

for SABER 11 scores is rather high, so the marginal student is quite strong academically. It might then be the case that the labor market returns to a college degree are so high for these students that they are willing to borrow to go to college even at high interest rates.

Figure 8 shows that the effect of the SPP program is roughly constant as a function of students' household wealth, consistent with the second hypothesis above. In Figure 9, we see that enrollment rates are increasing in household wealth, both for students with eligible or ineligible SABER 11 test scores. However, the magnitude of this increase is the roughly the same for these two groups, so that the treatment effect is approximately constant. Finally, Figure 10 shows that the effect of the SPP program on enrollment in high quality colleges or enrollment in high quality private colleges show the same qualitative patterns as the results for enrollment in any college.

Figure 8: Effect of the SPP on Enrollment as a Function of SISBEN Scores: Any College

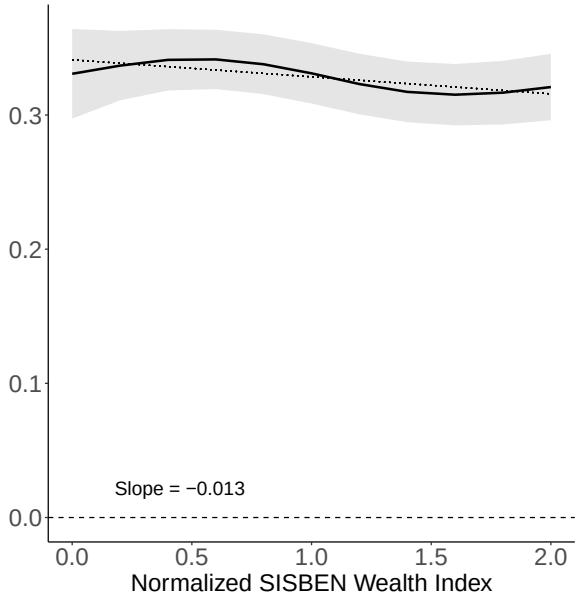


Figure 9: Enrollment Rates as a Function of SISBEN Scores

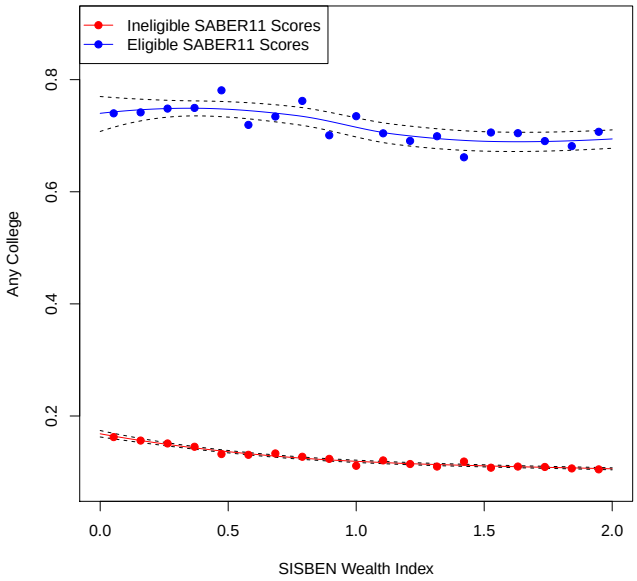
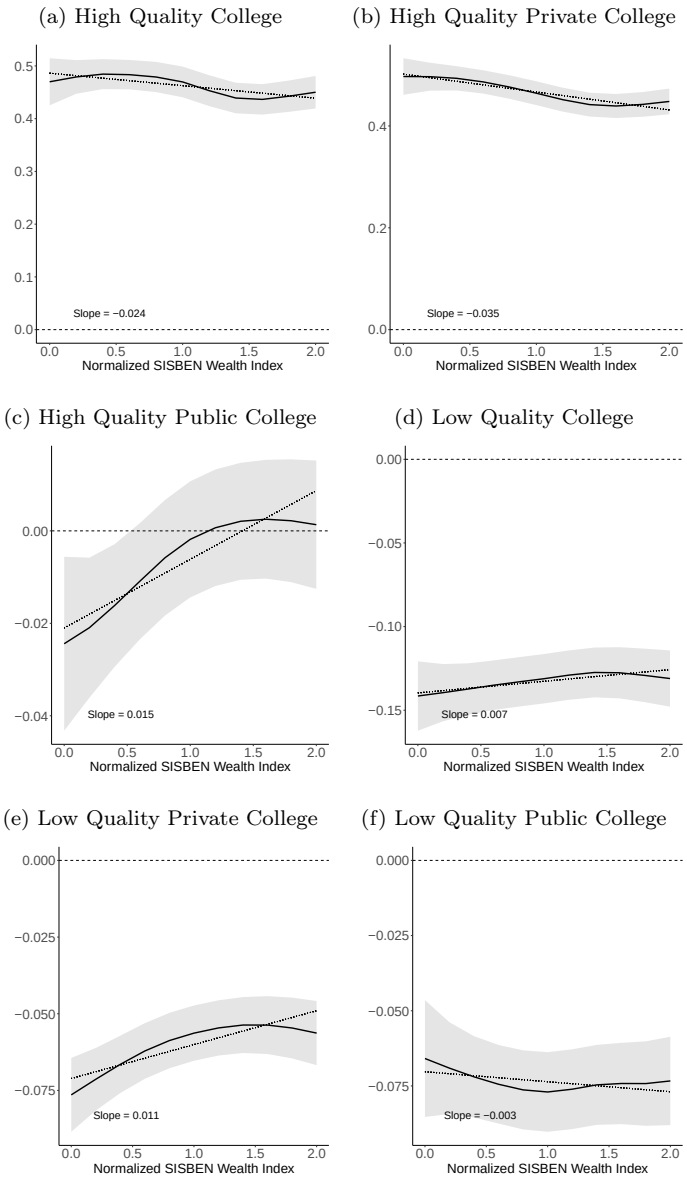


Figure 10: Effect of the SPP on Enrollment as a Function of SISBEN Scores: Different Types of Colleges



5 Conclusion

In this paper, I introduce a new method for estimating RD designs with multiple running variables. This estimator achieves efficiency gains relative to the common empirical approach of analyzing each running variable separately, and in addition it allows the researcher to estimate heterogeneous treatment effects. I verify the performance of my estimator in simulations. Finally, in an empirical application based on a large financial aid program in Colombia, I show that my estimator yields substantially more precise estimates of the average treatment effects for the marginal student, and that my estimates of heterogeneous treatment effects provide some economically relevant insights.

References

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

[2] Dell, Mellisa, 2010. "The Persistent Effects of Peru's Mining *Mita*." *Econometrica*, 78(6): 1863-1903.

[3] Duchon, Jean, 1977. "Splines minimizing rotation-invariant semi-norms in Sobolev spaces." *Constructive theory of functions of several variables.* Springer, Berlin, Heidelberg: 85-100.

[4] Gelman, Andrew, and Guido Imbens, 2019. "Why high-order polynomials should not be used in regression discontinuity designs". *Journal of Business & Economic Statistics*, 37(3): 447-456.

[5] Imbens, Guido, and Karthik Kalyanaraman, 2012. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies*, 79(3): 933-959.

[6] Imbens, Guido, and Stefan Wager, 2018. "Optimized Regression Discontinuity Designs." Working paper.

[7] Kolesár, Michal, and Chistoph Rothe, 2018. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review*, 108(8): 2277-2304.

[8] Londoño-Vélez, Juliana, Catherine Rodríguez, and Fabio Sánchez, 2020. "Upstream and Downstream Impacts of College Merit-Based Financial Aid for Low-Income Students: Ser Pilo Paga in Colombia." *American Economic Journal: Economic Policy.*

[9] Matsudaira, Jordan D., 2008. "Mandatory summer school and student achievement." *Journal of Econometrics,* 142(2): 829-850.

[10] Papay, John P., John B. Willett, and Richard J. Murnane, 2011. "Extending the regression-discontinuity approach to multiple assignment variables." *Journal of Econometrics,* 161(2): 203-207.

[11] Snider, Connan, and Jonathan W. Williams, 2015. "Barriers to Entry in the Airline Industry: A Multidimensional Regression-Discontinuity Analysis of AIR-21." *Review of Economics and Statistics*, 97(5): 1002-1022.

[12] Thistlethwaite, Donald L., and Donald T. Campbell, 1960. "Regression-discontinuity analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology*, 51(6): 309-317.

[13] Wahba, Grace, 1990. "Spline models for observational data." Vol. 59, Siam.

[14] Wood, 2003. "Thin plate regression splines." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95-114.

[15] Wood, Simon, 2006. *Generalized Additive Models: An Introduction with R.* CRC Press.

[16] Zajonc, Tristan, 2012. "Essays on Causal Inference for Public Policy." Doctoral dissertation.

Appendix

A1. Details on the Standard Error Calculations

My estimator of the conditional average treatment effect (CATE) is based on the difference between two fitted surfaces:

$$\hat{\tau}(x) = \hat{g}_1(x) - \hat{g}_0(x),$$

where $\hat{g}_1$ and $\hat{g}_0$ were estimated based on observations in the treated and untreated regions respectively. Generally, these functions can be written in the following form:

$$\hat{g}_r(x_1,x_2) = \sum_{k=1}^{K} \hat{\beta}_{(r),k} s_{(r),k}(x),$$

where $s_{(r),k}(x)$ is the kth basis function of the TPRS that is fit to the region where individuals receive treatment r. Adopting a Bayesian perspective with a normal prior, the posterior covariances $\Sigma_{(1)}$ and $\Sigma_{(0)}$ of the parameter vectors $\left(\hat{\beta}_{(1),k}\right)$ and $\left(\hat{\beta}_{(0),k}\right)$ are both distributed normal.[14] Also, if we assume i.i.d. error terms, then the fact that we fit $\hat{g}_1$ and $\hat{g}_0$ using separate data implies that the two parameter vectors are independent from each other, so that the entire posterior covariance matrix Σ is block diagonal with diagonal blocks $\Sigma_{(1)}$ and $\Sigma_{(0)}$. Standard software provides estimates $\hat{\Sigma}_{(1)}$ and $\hat{\Sigma}_{(0)}$ of these posterior covariances.[15]

Based on the discussion above, standard errors for the treatment effect at various points of the treatment frontier can be computed. Specifically, for $x \in \mathbb{F}$, we have:

$$\begin{aligned} Var\left(\hat{\tau}(x)\right) &= Var\left(\sum_{k=1}^{K} \hat{\beta}_{(1),k} s_{(1),k}(x) - \sum_{k=1}^{K} \hat{\beta}_{(0),k} s_{(0),k}(x)\right) \\ &= Var\left(\sum_{k=1}^{K} \hat{\beta}_{(1),k} s_{(1),k}(x)\right) + Var\left(\sum_{k=1}^{K} \hat{\beta}_{(0),k} s_{(0),k}(x)\right) \\ &= s_{(1)}(x)'\hat{\Sigma}_{(1)} s_{(1)}(x) + s_{(0)}(x)'\hat{\Sigma}_{(0)} s_{(0)}(x). \end{aligned}$$

Moreover, for any vector $(x^1, ..., x^L) \subset \mathbb{F}$, we can compute an estimate of the covariance matrix of this treatment effect estimate. In particular, we have:

$$\hat{Var}\left(\left(\hat{\tau}(x^1), ..., \hat{\tau}(x^L)\right)\right)_{l,p} = s_{(1)}(x^l)'\hat{\Sigma}_{(1)} s_{(1)}(x^p) + s_{(0)}(x^l)'\hat{\Sigma}_{(0)} s_{(0)}(x^p).$$

This can be written in matrix form:

$$\hat{Var}\left(\left(\hat{\tau}(x^1), ..., \hat{\tau}(x^L)\right)\right) = S'_{(1)}\hat{\Sigma}_{(1)} S_{(1)} + S'_{(0)}\hat{\Sigma}_{(0)} S_{(0)},$$

[14] Details on the Bayesian approach that leads to these posterior distributions can be found in Wood (2006).

[15] In particular, the "mgcv" package in R (which I use) provides these estimates of the posterior covariance matrices.

where $S_{(r)}$ is the $K \times L$ matrix with (k,l)th element equal to $s_{(r),k}(x^l)$.

Suppose we want to compute the average treatment effect over a subset of the treatment frontier $\mathbb{F}$. This can be done via numerical integration based on a discrete grid of points $(x^1, ..., x^L) \subset \mathbb{F}$. If we want to average the treatment effect using deterministic weights (e.g. based on some known counterfactual population of interest) $w_1, ..., w_L$ that are positive and sum to one, then we can simply compute the average of the estimate as $\sum_{l=1}^{L} w_l \hat{\tau}(x^l)$, and estimate the variance as $w' \hat{Var}\left(\left(\hat{\tau}(x^1), ..., \hat{\tau}(x^L)\right)\right) w$.

Suppose however, that we want to compute the average effect over the population (from which the estimation sample is randomly drawn from), but the distribution of the running variables for this population is unknown. In this case, we need to estimate the density. Many methods for density estimation yield estimates $\hat{f}(x^1), ..., \hat{f}(x^L)$ that are asymptotically normal, with some covariance matrix $\hat{\Sigma}^f$. To use these as weights, we need to scale them so that they sum to one. Hence, the estimate of the average treatment effect is:

$$\sum_{l=1}^{L} \frac{\hat{f}(x^l)}{\sum_{p=1}^{L} \hat{f}(x^p)} \hat{\tau}(x^l).$$

Denoting $\hat{\Sigma}^\tau \equiv \hat{Var}\left(\left(\hat{\tau}(x^1), ..., \hat{\tau}(x^L)\right)\right)$, and assuming independence between the estimates $\left\{\hat{\tau}(x^l)\right\}_{l=1}^{L}$ and $\left\{\hat{f}(x^l)\right\}_{l=1}^{L}$, we can obtain an estimate of the asymptotic variance of average treatment effect estimate using the delta method as follows.

Consider the function:

$$t(f, \tau) \equiv \sum_{l=1}^{L} \frac{f^l}{\sum_{p=1}^{L} f^p} \tau^l,$$

for $f \in \mathbb{R}_+^L$, $\tau \in \mathbb{R}^L$. The gradient of this function is:

$$\nabla t(\tau, f) = \begin{pmatrix} D_\tau t(\tau, f) \\ D_f t(\tau, f) \end{pmatrix},$$

where:

$$D_\tau t(\tau, f) = \begin{pmatrix} \dfrac{f^1}{\sum_{p=1}^{L} f^p} \\ \vdots \\ \dfrac{f^L}{\sum_{p=1}^{L} f^p} \end{pmatrix}, \; D_f t(\tau, f) = \begin{pmatrix} \sum_{l=1}^{L} \dfrac{f^l}{\left(\sum_{p=1}^{L} f^p\right)^2} \left(\tau^1 - \tau^l\right) \\ \vdots \\ \sum_{l=1}^{L} \dfrac{f^l}{\left(\sum_{p=1}^{L} f^p\right)^2} \left(\tau^L - \tau^l\right) \end{pmatrix}$$

So, we can estimate the asymptotic variance of the average treatment effect

estimate via the delta method using the following formula:

$$\nabla t(\hat{\tau}, \hat{f})' \begin{pmatrix} \hat{\Sigma}^{\tau} & 0 \\ 0 & \hat{\Sigma}^{f} \end{pmatrix} \nabla t(\hat{\tau}, \hat{f}) = \nabla t(\hat{\tau}, \hat{f})' \begin{pmatrix} \hat{\Sigma}^{\tau} D_{\tau} t(\hat{\tau}, \hat{f}) \\ \hat{\Sigma}^{f} D_{f} t(\tau, f) \end{pmatrix}$$
$$= D_{f} t(\hat{\tau}, \hat{f})' \hat{\Sigma}^{f} D_{f} t(\hat{\tau}, \hat{f}) + D_{\tau} t(\hat{\tau}, \hat{f})' \hat{\Sigma}^{\tau} D_{\tau} t(\hat{\tau}, \hat{f}).$$

More generally, this same methodology can be used to estimate average treatment effects over other subsets of the treatment frontier, and to obtain standard errors for these estimates.

A2. Estimation for MRD with Multiple Treatment Arms

The estimator described in this paper extends straightforwardly to the case with multiple treatment arms. In this setting, the running variable space is partitioned into more than two regions, with individuals in different regions receiving different treatments. In this case, one would simply fit a separate thin plate regression spline over each region of the running variable space, and take the difference between the fitted splines at the treatment frontiers to obtain an estimate of the relative effects between two different treatment types.