Regression Discontinuity Designs with Multiple Assignment Variables *

Yizhuang Alden Cheng

April 11, 2016

Abstract

In this paper, I extend current research on regression discontinuity (RD) designs with multiple assignment variables. I discuss the assumptions underlying the validity of such RD designs, and introduce graphical methods that check for violations of these assumptions. Instead of estimating a scalar treatment effect as in RD designs with a single assignment variable, I propose estimating treatment functions that are defined on the boundaries that separate different treatment groups in the assignment variable space. I generalize a local linear regression method currently used for estimation of said treatment functions, and separately develop a novel estimation method using thin plate regression splines. The performances of these estimation methods are assessed through an extensive simulation study based closely on real data.

1 Background

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

In practice however, there are many instances where treatment is determined by more than one assignment variable, such as when numerous criteria determine eligibility for a benefit, or when students need to achieve a minimum score for each component of an exam in order to move onto the next grade level. Multiple-assignment variable RD (MRD) designs estimate treatment effects for

*I would like to thank my thesis adviser, Professor David Card, for all of his support and insights. All errors in this paper are mine alone.

such instances by considering observations near the boundaries separating different treatment groups in the multidimensional assignment variable space. 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). For the latter category, treatment effects are only defined for pairwise comparisons of treatment conditions, so there is no natural "control" group. To avoid confusion, I adopt the following taxonomy for different categories of RDs throughout the rest of this paper:

Conventional RD: One assignment variable/Dichotomous treatment;

MDRD: Multiple assignment variables/Dichotomous treatment;

MMRD: Multiple assignment variables/Multiple mutually exclusive treatment conditions.

The next example illustrates the difference between MDRD and MMRD.

Consider first an exam that has a math and reading component, both of which students must pass in order to move onto the next grade level. There are two assignment variables (math and reading test scores) and dichotomous treatment (whether or not the student moves onto the next grade), making this a MDRD. Next, consider a slightly modified scenario, in that students who only fail one of the components are merely required to attend remedial classes. This is now a MMRD with two assignment variables and multiple mutually exclusive treatment conditions (grade retention, remedial classes, or moving onto the next grade).

While various papers have used MRD for analysis, the estimands of interest and estimation methods used are extremely varied. This is because, unlike for conventional RD, there is scant literature on methods for MRD, the following papers being rare exceptions. Wong, Steiner and Cook (2013) compare four different estimation methods for MDRD. Papay, Willett and Murnane (2011a) define MMRD, and propose a specific method of estimation. Reardon and Robinson (2012) briefly discuss both types of MRD.

This paper extends the nascent literature on both categories of MRD in several ways. First, I justify the estimation of treatment effect *functions* rather than *scalar* treatment effects, and introduce key assumptions underlying the validity of MRD. Then, I develop graphical analysis methods specifically tailored for MRD. Such graphical presentations can enhance transparency of the research design and check for violations of the identifying assumptions. Next, I modify an existing estimation method to address its limitations, and separately develop a novel estimation method for MRD using thin plate regression splines. In an extensive simulation study, estimation via thin plate regression splines outperforms other conventional estimation methods.

The rest of this paper is organized as follows. Section 2 briefly reviews conventional RD, defines both types of MRD formally, and outlines the assumptions that underpin these RD designs. Section 3 discusses graphical analysis methods for MRD. Section 4 describes a generalization of an existing estimation

approach, and proposes a novel estimation method which uses thin plate regression splines. Section 5 covers a simulation study modeled closely on data from Kane (2003). Section 6 discusses additional issues such as fuzzy MRD and section 7 concludes.

2 Regression Discontinuity Designs

This section covers the basics of conventional RD and MRD, introducing notation that will be used throughout this paper. Key assumptions underlying the validity of these designs, as well as estimands of interest are stated.

2.1 Conventional RD

I begin this section by briefly reviewing the conventional RD design, since it motivates most concepts in MRD. Given a sample of size N, I denote the treatment indicator for observation i by W_i. A sharp RD design is assumed, so W_i is completely determined by the value of a one-dimensional assignment variable X_i, which I assume to be continuous for now. To simplify notation, I assume without loss of generality that X_i has been centered about its cutoff, and that the assignment rule is:

$$W_i \equiv \mathbb{I}[X_i >= 0].$$

I write the potential outcomes as $Y_i(0)$ and $Y_i(1)$, where the observed outcomes are $Y_i = Y_i(0)$ for observations with $X_i < 0$, and $Y_i = Y_i(1)$ for observations with $X_i \geq 0$. The key identification assumption for the conventional RD design is that the conditional expectation functions of potential outcomes are continuous at the cutoff for the assignment variable X_i:

$$\lim_{x \to 0^+} \mathbb{E}[Y_i(0)|X_i = x] = \lim_{x \to 0^-} \mathbb{E}[Y_i(0)|X_i = x], \text{ and}$$

$$\lim_{x \to 0^+} \mathbb{E}[Y_i(1)|X_i = x] = \lim_{x \to 0^-} \mathbb{E}[Y_i(1)|X_i = x].$$

A stronger assumption – that the conditional expectation functions are continuous over their domains of definition – is often used in practice, since it is hard to imagine the weaker assumption being met but with discontinuities occurring at non-cutoff points in a well-formulated problem.

Under this setup, the estimand of interest is

$$\tau_{RD} = \lim_{x \to 0^+} \mathbb{E}[Y_i|X_i = x] - \lim_{x \to 0^-} \mathbb{E}[Y_i|X_i = x]. \quad (1)$$

2.2 MDRD

This subsection introduces RD designs with multiple assignment variables and dichotomous treatment, which I abbreviate as MDRD throughout this paper. For most of the paper, I focus on the case with $d = 2$ assignment variables for

notational simplicity. However, most of the discussion can easily be generalized to instances where $d > 2$. Again, I assume that the assignment variables have been centered about their cutoffs and denote the assignment variables for observation i by X_{1i} and X_{2i}. Occasionally, I use the notation $\boldsymbol{X_i}$ to denote the vector of assignment variables in a MRD design.

Retaining the notation for potential outcomes, I assume without loss of generality that the assignment rule is of the "AND" type (i.e. in order to qualify for treatment, the cutoffs for both assignment variables must be met)[1]:

$$W_i \equiv D_{1i} \times D_{2i} = \mathbb{I}[X_{1i} \geq 0 \text{ and } X_{2i} \geq 0], \text{ where}$$

$$D_{1i} \equiv \mathbb{I}[X_{1i} \geq 0] \text{ and } D_{2i} \equiv \mathbb{I}[X_{2i} \geq 0].$$

Here, in contrast to the conventional case where there is a single scalar cutoff, there are two thresholds, one for each assignment variable. Moreover, the assignment variable space in MRD is of dimension $d = 2$, so that the boundaries separating different treatment groups have dimension $d - 1 = 1$. The treatment frontiers are defined as:

$$F_1 \equiv \{(x_1, x_2) | x_1 = 0 \text{ and } x_2 \geq 0\}, \text{ and}$$

$$F_2 \equiv \{(x_1, x_2) | x_1 \geq 0 \text{ and } x_2 = 0\}.$$

In order to estimate treatment effects using these frontiers, certain continuity assumptions for the conditional expectation functions of the potential outcomes are needed. The assumptions for $Y(1)$ are that

$$\lim_{x_1 \to 0^+} \mathbb{E}[Y_i(1) | X_{1i} = x_1, X_{2i} = x_2] = \lim_{x_1 \to 0^-} \mathbb{E}[Y_i(1) | X_{1i} = x_1, X_{2i} = x_2] \text{ for } x_2 \geq 0,$$

$$\lim_{x_2 \to 0^+} \mathbb{E}[Y_i(1) | X_{1i} = x_1, X_{2i} = x_2] = \lim_{x_2 \to 0^-} \mathbb{E}[Y_i(1) | X_{1i} = x_1, X_{2i} = x_2] \text{ for } x_1 \geq 0,$$

and similarly for $Y(0)$, as Wong et al. (2013) note. Once again, it may be reasonable to assume that these conditional expectation functions are continuous

[1] As an example of why it suffices to consider "AND" assignment rule, consider the following "OR" assignment rule:

$$W_i = \mathbb{I}[X_{1i} > 0 \text{ or } X_{2i} > 0].$$

In this case, I can simply redefine the treatment indicator and assignment variables by

$$\tilde{W}_i \equiv 1 - W_i, \quad \tilde{X}_{1i} \equiv -X_{1i}, \quad \tilde{X}_{2i} \equiv -X_{2i}.$$

This yields

$$\tilde{W}_i = \mathbb{I}[X_{1i} \leq 0 \text{ and } X_{2i} \leq 0] = \mathbb{I}[\tilde{X}_{1i} \geq 0 \text{ and } \tilde{X}_{2i} \geq 0],$$

which is in the form of an "AND" assignment rule.

over their domains of definitions, although this is a stronger assumption than necessary.

Unlike in conventional RD, there is some ambiguity over the estimand of interest in MRD. Let $G_i = Y_i(1) - Y_i(0)$ be the difference in potential outcomes for observation i, $g(x_1, x_2)$ be the difference in expected potential outcomes as a function of the assignment variables, and $f(x_1, x_2)$ be the joint density function for (X_1, X_2).

Wong et al. (2013) consider the two frontier-specific treatment effects and the overall treatment effect to be the estimands of interest, and introduce a "frontier approach" that estimates these three quantities. Using this paper's notation, the treatment effects specific to F_1 and F_2 are

$$\tau_1^{Wong} \equiv \mathbb{E}[G_i|(X_{1i}, X_{2i}) \in F_1] = \frac{\int_{x_2 \geq 0} g(0, x_2) f(0, x_2) dx_2}{\int_{x_2 \geq 0} f(0, x_2) dx_2}, \text{ and} \tag{2}$$

$$\tau_2^{Wong} \equiv \mathbb{E}[G_i|(X_{1i}, X_{2i}) \in F_2] = \frac{\int_{x_1 \geq 0} g(x_1, 0) f(x_1, 0) dx_1}{\int_{x_1 \geq 0} f(x_1, 0) dx_1} \tag{3}$$

respectively. The overall treatment effect

$$\tau_{MDRD}^{Wong} \equiv \mathbb{E}[G_i|((X_{1i}, X_{2i}) \in F_1 \cup F_2)] = w_1 \tau_1^{Wong} + w_2 \tau_2^{Wong}, \tag{4}$$

is a weighted average of the frontier-specific treatment effects, with the weights respectively reflecting the probability of an observation being in each of the frontiers (conditional on being on one of the frontiers):

$$w_1 \equiv \frac{\int_{x_2 \geq 0} f(0, x_2) dx_2}{\int_{x_1 \geq 0} f(x_1, 0) dx_1 + \int_{x_2 \geq 0} f(0, x_2) dx_2}, \text{ and} \tag{5}$$

$$w_2 \equiv \frac{\int_{x_1 \geq 0} f(x_1, 0) dx_1}{\int_{x_1 \geq 0} f(x_1, 0) dx_1 + \int_{x_2 \geq 0} f(0, x_2) dx_2}. \tag{6}$$

However, there are several aspects to this approach that are undesirable.

First, as Wong et al. (2013) note, the estimated overall treatment effect is not invariant to rescaling of the assignment variables. This problem is less serious when the assignment variables are measured in comparable scales, for instance when they represent scores on different components of a test, as in Jacob and Lefgren (2004) and Matsudaira (2008). However, there are also cases where the units of measurement for assignment variables are not aligned, such as with parental income and high school GPA, which are the assignment variables used in Kane (2003). There does not seem to be a natural scaling of GPA and income that would make the units of measurement "comparable", and the fact that the overall treatment effect estimate depends on such an arbitrary scaling decision diminishes the credibility of this estimate.

Second, potentially interesting heterogeneities in the treatment effect may be lost in summarizing the effects as scalar quantities. For example, Kane (2003)

investigates how financial aid affects the college going decision, where the assignment variables determining aid eligibility are high school GPA and parental income. The treatment effect at the GPA threshold τ_{GPA}^{Wong} is a weighted average over students with different family incomes (below the income threshold). Yet, one may suspect that at the GPA frontier, the treatment effect would be greater for students with lower family income, since financial constraints are presumably a greater barrier to college going for these students. It would not be possible to test the validity of this conjecture using the scalar treatment effect estimates proposed by Wong et al. (2013).

Finally, Wong et al. (2013) admit that their approach requires the strong assumption that the response surface $g(x_1, x_2)$ is correctly specified. Moreover, estimation of the joint density $f(x_1, x_2)$ and numerical integration requires large amounts of data and is computationally expensive.

This paper proposes estimating treatment functions instead of scalar treatment effects. Specifically, I estimate the functions

$$\tau_1(x_2) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_{1i} = 0, X_{2i} = x_2], \tag{7}$$

$$\tau_2(x_1) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_{1i} = x_1, X_{2i} = 0] \tag{8}$$

for $x_2 \geq 0$ and $x_1 \geq 0$ respectively. Assuming that the expectations of potential outcomes conditional on the assignment variables are continuous, it follows that $\tau_1(x_2)$ and $\tau_2(x_1)$ are continuous functions and that their values coincide at the intersection of the treatment frontiers, i.e. $\tau_1(0) = \tau_2(0)$.

This approach is both simpler and circumvents most shortcomings of the "frontier approach" proposed by Wong et al. (2013). In particular, estimation of treatment effect functions does not require density estimation or numerical integration, and these functions capture variations in treatment effects over different subpopulations near the treatment frontiers.

2.3 MMRD

With more than one assignment variable, it is not uncommon for there to be multiple mutually exclusive treatments, as exemplified by the numerous examples given by Papay et al. (2011a). This necessitates additional notation, and the assignment rule from MDRD needs to be modified for MMRD[2].

[2]I assume here for simplicity of exposition, that there are four treatment conditions, with both assignment variables centered at their cutoffs. Papay et al. (2011a) note that in general with d assignment variables, there are 2^d different treatments. In fact, there may be more or less than 2^d different treatments. An example with two assignment variables and three treatments is given in the introduction – the treatments being (i) forced to stay back a grade if both tests are failed; (ii) summer remedial classes if only one test is failed; (iii) moving onto the next grade if both tests are passed. Cases with two assignment variables and more than four possible treatments are also imaginable when there are multiple cutoffs.

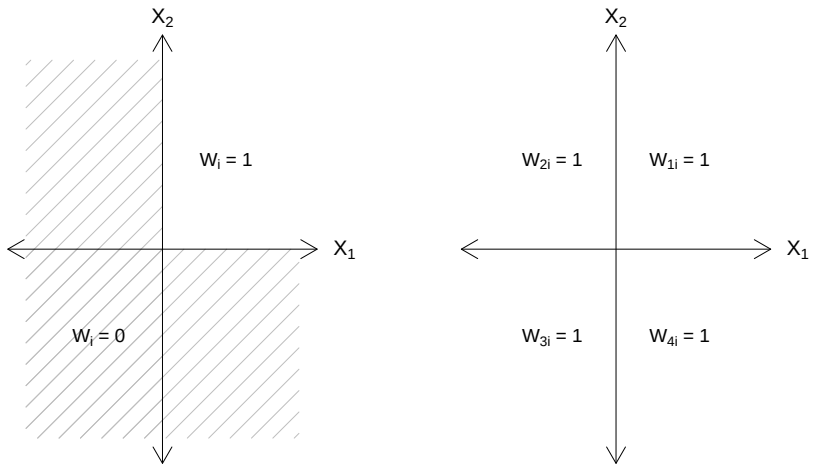

Figure 1: An illustration of the difference between MDRD (left panel) and MMRD (right panel)

To simplify notation, I denote the four quadrants of the assignment variable space by

$$R_1 \equiv \{(x_1, x_2)|x_1 \geq 0, x_2 \geq 0\},$$
$$R_2 \equiv \{(x_1, x_2)|x_1 < 0, x_2 \geq 0\},$$
$$R_3 \equiv \{(x_1, x_2)|x_1 < 0, x_2 < 0\},$$
$$R_4 \equiv \{(x_1, x_2)|x_1 \geq 0, x_2 < 0\}.$$

Indicators for the four treatments can thus written as

$$\text{Treatment 1: } W_{1i} \equiv D_{1i} \times D_{2i} = \mathbb{I}[\boldsymbol{X_i} \in R_1],$$

$$\text{Treatment 2: } W_{2i} \equiv (1 - D_{1i}) \times D_{2i} = \mathbb{I}[\boldsymbol{X_i} \in R_2],$$

$$\text{Treatment 3: } W_{3i} \equiv (1 - D_{1i}) \times (1 - D_{2i}) = \mathbb{I}[\boldsymbol{X_i} \in R_3],$$

$$\text{Treatment 4: } W_{4i} \equiv D_{1i} \times (1 - D_{2i}) = \mathbb{I}[\boldsymbol{X_i} \in R_4].$$

Similarly, there are four different treatment frontiers which coincide with the non-negative and negative x_1- and x_2-axes. A treatment effect function is estimated along each of these frontiers. The frontier separating treatments 1 and 2 is

$$F_{12} \equiv \{(x_1, x_2)\ x_1 = 0 \text{ and } x_2 \geq 0\}$$

and the treatment effect function (moving from treatment 1 to 2) is denoted

$$(9) \quad \tau_{12}(x_2) \equiv \mathbb{E}[Y_i(2) - Y_i(1)|X_{1i} = 0 \text{ and } X_{2i} = x_2].$$

The other frontiers F_{23}, F_{34} and F_{14}, and treatment effect functions $\tau_{23}(x_1)$, $\tau_{34}(x_2)$ and $\tau_{14}(x_1)$ are defined analogously. These treatment effect functions represent the effect of moving from treatment condition 3 to 2, 3 to 4, and 4 to 1 respectively[3].

Validity of the MMRD design once again relies on the conditional expectation of the potential outcomes obeying certain continuity conditions along the treatment frontiers. For instance, denoting the four potential outcomes by $Y(1)$, $Y(2)$, $Y(3)$ and $Y(4)$, estimation of $\tau_{12}(x_2)$ requires

$$\lim_{x_1 \to 0^+} \mathbb{E}[Y_i(1)|X_{1i} = x_1, X_{2i} = x_2] = \lim_{x_1 \to 0^-} \mathbb{E}[Y_i(1)|X_{1i} = x_1, X_{2i} = x_2],$$

$$\lim_{x_1 \to 0^+} \mathbb{E}[Y_i(2)|X_{1i} = x_1, X_{2i} = x_2] = \lim_{x_1 \to 0^-} \mathbb{E}[Y_i(2)|X_{1i} = x_1, X_{2i} = x_2],$$

for $x_2 \geq 0$. Analogous continuity assumptions are necessary for the estimation of $\tau_{23}(x_1)$, $\tau_{34}(x_2)$ and $\tau_{14}(x_1)$.

As before, it is not unreasonable to ask that the conditional expectation functions for the four potential outcomes be continuous over the entire assignment variable space.

However, it is worth mentioning that in general for $(i_1, j_1) \neq (i_2, j_2)$, the equality $\tau_{i_1 j_1}(0) = \tau_{i_2 j_2}(0)$ may not hold (unlike in MDRD)[4], which is due to there being more than two treatment functions under consideration. To explain why this is reasonable, consider an example with math and reading test scores as assignment variables determining four possible treatments – grade retention, math remedial classes over summer, reading remedial classes over summer, and moving onto the next grade level. It is clear in this case that there is no reason to expect the treatment effect which compares grade retention and summer math class, to be similar to the treatment effects comparing reading or math class to no remedial, even for pairs of test scores that are relatively close in Euclidean distance.

Another noteworthy issue is the fact that although there are $\binom{4}{2} = 6$ different pairs of treatment conditions that one may wish to compare, this paper only focuses on comparing treatments that are separated by a one-dimensional boundary in the assignment variable space (i.e. the non-negative or negative x_1 or x_2-axis). I omit comparisons of treatments 1 and 3, as well as 2 and 4, which have treatment boundaries that contain only a single point – the origin. In practice, this often implies that there is an insufficient number of observations (near the boundary) to estimate these treatment effects precisely.

[3]Strictly speaking, there are two treatment effects that can be estimated along each frontier, corresponding to a movement from one treatment to the other, and the movement in the opposite direction. Since these two treatment effect estimates will simply be of opposite signs, I only choose one treatment effect function to consider for each frontier.

[4]For a simple illustration of this, consider the case where the potential outcomes are deterministic, so that $Y_i(k) = k$ for $k \in \{1, 2, 3, 4\}$. Under this setup, $\tau_{12}(0) = -1 \neq -3 = \tau_{14}(0)$.

3 Graphical Analysis

Graphical analysis is critical in establishing the credibility of conventional RD designs. Typically, one plots the outcome variable as a function of the assignment variable, then examines the graph for the presence of a discontinuity at the cutoff value of the assignment variable, which may be taken as visual evidence of a nonzero treatment effect. The same is sometimes done for the treatment indicator variable. Diagnostic plots are also useful for checking whether the identifying assumptions of the RD design are being violated. A common diagnostic plot graphs predetermined characteristics as functions of the assignment variable. Another frequently used diagnostic plot is a histogram of the assignment variable. Graphical analysis methods for conventional RD are well-documented in survey papers such as Imbens and Lemieux (2008) and Lee and Lemieux (2010).

While relatively easy to implement for conventional RD, greater dimensionality in the MRD setting presents additional challenges. Most relevant among these is the fact that the visual impact of plots in more than two dimensions is greatly diminished. Perhaps due to this difficulty, there has been no discussion on graphical analysis in the current MRD literature that I am aware of. In this section, I review the common plots used in conventional RD and discusses how they may be generalized appropriately for the MRD setting.

3.1 Discontinuity in Outcomes

The most common graph used in conventional RD plots the outcome variable as a function of the assignment variable. This graph is designed to provide visual evidence of a discontinuity in outcomes at the cutoff, which would suggest a nonzero treatment effect. The procedure for conventional RD typically involves partitioning the assignment variable space into disjoint bins (or intervals) of constant width, computing the average outcomes within each bin, and plotting these average outcomes against midpoints of the bins. Care must be taken in defining the intervals so that the cutoff does not lie in the interior of any interval; otherwise, this would result in a point that aggregates outcomes for observations from both treatment conditions, making it hard to interpret. A polynomial regression line is often fitted to points on each side of the cutoff to improve the visual impact of the plot.

More formally, given a bandwidth h, one first constructs K intervals

$$\begin{aligned}
&[b_0, b_1), [b_1, b_2), ..., [b_{K-1}, b_K), \\
&\qquad \text{s.t. } 0 \in \{b_1, ..., b_{K-1}\}, \\
&\qquad\quad b_k - b_{k-1} = h \text{ for all } k \in \{1, ..., K\}, \\
&\qquad\quad \min\{x_i\} \in [b_0, b_1) \text{ and } \max\{x_i\} \in [b_{K-1}, b_K).
\end{aligned}$$

Then, one calculates the average outcomes within each interval,

$$\bar{Y}_k \equiv \frac{1}{|\{i|x_i \in [b_{k-1}, b_k)\}|} \sum_{i=1}^{N} Y_i \cdot \mathbb{I}[x_i \in [b_{k-1}, b_k)] \text{ for } k \in \{1, ..., K\},$$

and plots $\bar{Y}_k$ against $\frac{b_{k-1} + b_k}{2}$.

Various approaches have been suggested for the choice of h. Lee and Lemieux (2010) propose two – one based on cross-validation, and the other based on visual inspection coupled with an F-test to determine whether h is small enough[5]. The guiding principle behind these approaches is to choose h sufficiently small so that the plot does not "over-smooth" the data, but large enough so that the bin estimates are still reasonably precise.

Here, I consider how a similar type of graphical analysis might work in the MRD setting with two assignment variables. I first describe several straightforward extensions of the procedure for conventional RD, which I argue are either inappropriate or suboptimal. Then, I introduce the "slicing" and "sliding window" approaches, which create plots that are better suited for MRD.

3.1.1 Straightforward Extensions of the Conventional RD Plot that are Suboptimal

Perhaps the most straightforward extension of the procedure for conventional RD is to create an analogous graph in three dimensions, using the two assignment variables and the outcome variable. Since the assignment variable space is two-dimensional, two widths need to be chosen (which I denote by h_1 and h_2 for X_1 and X_2 respectively)[6]. One would then construct P one-dimensional intervals for X_1 and separately, Q intervals for X_2, using the same procedure as for conventional RD. This yields PQ two-dimensional intervals (or bins) which I denote by

$$I_{p,q} \equiv \{(x_1, x_2)|b_{1,p-1} \leq x_1 < b_{1,p} \text{ and } b_{2,q-1} \leq x_2 < b_{2,q}\}.$$

For non-empty intervals, average outcomes within each bin are computed as before using the formula

$$\bar{Y}_{p,q} \equiv \frac{1}{|\{i|(x_{1i}, x_{2i}) \in I_{p,q}|} \sum_{i=1}^{N} Y_i \cdot \mathbb{I}[(x_{1i}, x_{2i}) \in I_{p,q}],$$

[5]Specifically, they consider the regression with K bin dummies (indicating whether an observation is in a given bin) for bins of width h, as well as an alternative specification with $2K$ bin dummies for bins of width $h/2$. Since the first model is nested within the second, an F-test can be used to compare the models. A rejection of the null hypothesis would suggest that the choice of h is too large and that the data is being "over-smoothed".

[6]One may be able to get away with choosing a single bin width h for both assignment variables if the scales are comparable, e.g. scores for different components of a test. However, I focus on the more general case, since there are many instances where this does not apply, such as when the assignment variables are GPA and income.

and plotted over the centers of the bins, $\left(\frac{b_{1,p-1}+b_{1,p}}{2}, \frac{b_{2,q-1}+b_{2,q}}{2}\right)$ in a three-dimensional graph, possibly with a smoothing surface fitted to each treatment region.

The main problem with this graph is that the height (which represents the value of the outcome variable) in three dimensional graphs tends to be distorted. In particular, perceptions about the sizes of discontinuities tend to depend on the "viewpoint" chosen for the plot. It will be especially challenging to choose a "viewpoint" that accurately portrays the discontinuities along all four frontiers in a MMRD. Hence, it may be more appropriate instead to summarize this information using two-dimensional graphs.

For MDRD, an easy way to create a single plot illustrating the discontinuity near the boundary makes use of the "centering" approach of Wong et al. (2013)[7]. The motivating idea is that instead of considering the two assignment variables separately, one focuses instead on the distance of an observation from the boundary separating the two treatments. Assuming that the scales of the two centered assignment variables are comparable (otherwise, one may standardize the variables to have equal variance), one creates a new variable

$$Z_i \equiv \inf\{d(\boldsymbol{X_i}, \boldsymbol{X}) | \boldsymbol{X} \in F_1 \cup F_2\},$$

where d denotes a distance function of the user's choice[8]. Now, the variable Z_i is used as the new assignment variable for graphical purposes, so the approach for conventional RD can be used.

A similar method may be employed for MMRD, although the two "adjacent" treatments that are being compared need to be specified. This yields four plots in total (one for each pair of treatments being compared). For instance, to compare treatments 1 and 2 (which are separated in the assignment variable space by F_{12}) graphically, one would compute

$$Z_i \equiv \inf\{d(\boldsymbol{X_i}, \boldsymbol{X}) | \boldsymbol{X} \in F_{12}\} \text{ for } i \text{ s.t. } \boldsymbol{X_i} \in R_1 \cup R_2$$

and create the conventional RD plot described earlier, using only the subset of observations receiving either treatments 1 or 2.

However, this "centering" approach has a major limitation in the dichotomous treatment case, in that it does not show frontier-specific treatment effects. By using a single assignment variable Z_i instead of the two original assignment variables, individuals close to the frontier F_1 are treated essentially the same as those close to the F_2. This is especially undesirable when the original assignment variables are qualitatively different, so that frontier-specific effects are of

[7] Although Wong et al. (2013) originally described this as a method for estimation, the graphical method I describe here is motivated by the same ideas.

[8] Wong et al. (2013) use the distance function induced by the L_∞-norm, although one may prefer distance functions induced by other norms such as L_2 (Euclidean) or L_1 depending on the context.

interest (e.g. whether children around the income threshold respond to financial aid differently from children around GPA threshold)[9].

A simple method for MDRD that seems to address this shortcoming is to plot two graphs using the procedure for conventional RD. Specifically, for the first graph, one takes the graphical approach for conventional RD, treating the first assignment variable as if it were the only assignment variable (and likewise for the second graph using the second assignment variable).

At first glance, this method seems to capture frontier-specific effects (or more precisely, estimates of τ_1^{Wong} and τ_2^{Wong}). However, I argue that this approach yields misleading graphical estimates. To elaborate, consider the plot with X_1 on the horizontal axis, which should display approximately the average discontinuity along F_1, i.e. τ_1^{Wong}. An outcome Y_{0+} for a point at (or just above) the cutoff of zero on this graph *should* represent an average of the points near F_1 that receive the treatment, i.e.

$$Y_{0+} \approx \mathbb{E}[Y_i | X_i \in F_1] \approx \int_0^{\infty} \int_0^{\epsilon} y \cdot f(x_1, x_2) dx_1 dx_2$$

for some small $\epsilon > 0$. However, the actual quantity represented by Y_{0+} is

$$Y_{0+} \approx \mathbb{E}[Y_i | X_{1i} = 0] \approx \int_{-\infty}^{\infty} \int_0^{\epsilon} y \cdot f(x_1, x_2) dx_1 dx_2.$$

The problem is that this method does not take into account the value of the other assignment variable X_2, and so the expectation is taken over the entire x_2-axis rather than over F_1 (the non-negative x_2-axis). As a result, observations with $X_{2i} < 0$ (which are ineligible for treatment and may be far from F_1) are involved in the computation of Y_{0+}, which is supposed to represent the average outcomes for individuals just qualifying for treatment (along the F_1 frontier). Hence, graphs produced by this approach are essentially uninterpretable.

This problem can be fixed with a simple modification – when plotting the graph using X_1 as the assignment variable, instead of using all the data, one only uses observations with $X_2 \geq 0$ (and vice versa when constructing the graph for X_2). This approach is similar in spirit to the "univariate" estimation method discussed in Wong et al. (2013). For MMRD, there is no difference between this method and the "centering" approach described earlier.

Nonetheless, it is possible to do better, since the estimands of interest in this paper are the treatment functions $\tau_1(x_2)$ and $\tau_2(x_1)$, rather than the scalar treatment effects τ_1^{Wong} and τ_2^{Wong}. The remainder of this subsection presents two graphical methods that are more appropriate for displaying non-constant treatment effects – the "slicing" and "sliding window" plots.

[9] This drawback is also the main reason this paper does not generally recommend using the "centering" approach for estimation despite the method's simplicity.

3.1.2 "Slicing" and "Sliding Window" Plots

The following method, which this paper calls the "slicing" approach, represents an extension of the "univariate" approach that enables an approximate visual representation of the treatment effect functions[10]. Here, I focus on the dichotomous treatment case and describe the procedure for producing graphs associated with $\tau_1(x_2)$. The procedures for $\tau_2(x_1)$, as well as for MMRD are completely analogous.

One considers only the subset of observations $\{(Y_i, \boldsymbol{X_i})|X_{2i} \geq 0\}$, and partitions (or "slices") this data into disjoint subsets (or "slices")

$$S_k \equiv \{(Y_i, \boldsymbol{X_i})|X_{2i} \in [s_{k-1}, s_k)\}$$

for $k = 1, ..., K$, where $0 = s_0 < s_1 < ... < s_{K-1} < s_K < \infty$.

Next, for each of these partitions S_k, one treats X_1 as the assignment variable in a conventional RD design and creates a two-dimensional plot using a suitable bandwidth. This results in a series of K "slicing" plots, one corresponding to each interval of (non-negative) X_2. An estimate of the treatment effect corresponding to each subset S_k of observations can be obtained using the polynomial regressions that are fit to points on each side of the cutoff. This yields a crude point estimate of $\tau_1(x_2)$ at a number of values for x_2.

The "slicing" method allows one to examine discontinuities in the outcome variable at the cutoff for one assignment variable (say X_1) for different subsets of observations (grouped based on values of the other assignment variable, X_2). These plots also contain some information on whether the size of these discontinuities vary with X_2 (i.e. whether the treatment effect function $\tau_1(x_2)$ is constant). However, there are typically only a limited number of these plots to compare since disjoint subsets of observations are used to create each plot, so the visual evidence on whether $\tau_1(x_2)$ is constant is still rather scant. Hence, I next describe the "sliding window" plot, which essentially summarizes a series of "slicing" plots to provide a better visual summary of the treatment effect function over its domain of definition.

For concreteness, I once again focus on $\tau_1(x_2)$ in my description of the "sliding window" plot (as before, the procedures for creating the plot for $\tau_2(x_1)$, as well as for MMRD are completely analogous). This graph summarizes the discontinuities occurring at the cutoff for X_1 over different intervals of X_2. One of the main differences in the calculations required to construct this plot versus the "slicing" plots, is that instead of partitioning the data into disjoint subsets according to values of X_2, one creates subsets of data by "sliding" the interval of X_2 under consideration (hence the name of the graph). In particular, the intervals of X_2 used for different subsets are no longer disjoint.

There are a few parameters in the "sliding window" plot that the user may choose to provide the best visual impact – w (the width of intervals for X_2),

[10] In fact, one may think of the "univariate" approach as a special case of the "slicing" approach with $K = 1$, using the notation introduced below.

c (the constant that the interval for X_2 is shifted by each time) and h (the bandwidth for X_1, controlling the amount of data just above and below the cutoff of X_1 to use). The following subsets of $\{(Y_i, \boldsymbol{X_i})|X_{2i} \geq 0\}$ are created:

$$W_k \equiv \{(Y_i, \boldsymbol{X_i})|X_{2i} \in [(k-1)c, (k-1)c + w)\}$$

for $k = 1, ..., K'$, where

$$K' \equiv \inf\{k \in \mathbb{Z}|(k-1)c + w > \max\{X_{2i})\}.$$

For each W_k, let W_k^l and W_k^u be the subsets of observations in W_k below and above the cutoff for X_1 that are within h of this cutoff respectively. One then computes the average values of the outcome variable Y for the subsets W_k^l and W_k^u, which I denote by $\bar{Y}_k^l$ and $\bar{Y}_k^u$ respectively.

The final step in creating the "sliding window" graph is to plot $\bar{Y}_k^l$ and $\bar{Y}_k^u$, against the midpoints of X_2 in W_k, which are given by

$$X_{2,k}^{mid} = \frac{2(k-1)c + w}{2}.$$

The vertical distance between $\bar{Y}_k^u$ and $\bar{Y}_k^l$ is an approximation of $\tau_1(X_{2,k}^{mid})$.

The procedure just described for $\tau_1(x_2)$ uses a rectangular kernel of width w for X_2 in the calculation of $\bar{Y}_k^l$ and $\bar{Y}_k^u$. However, one may reasonably argue against this uniform weighting of points within W_k by insisting that observations with values of X_2 closer to $X_{2,k}^{mid}$ should have a greater impact in the determination of $\bar{Y}_k^l$ and $\bar{Y}_k^u$. The method for constructing "sliding window" plots I described can easily be modified to accommodate other choices of kernels with compact support, e.g. triangular or Epanechnikov. In particular, the only change required is to use weighted averages for $\bar{Y}_k^l$ and $\bar{Y}_k^u$. The formulae are

$$\bar{Y}_k^l = \frac{1}{\sum_{i=1}^N w_{i,k}^l} \sum_{i=1}^N w_{i,k}^l Y_i \quad \text{and} \quad \bar{Y}_k^u = \frac{1}{\sum_{i=1}^N w_{i,k}^u} \sum_{i=1}^N w_{i,k}^u Y_i,$$

where $w_{i,k}^l$ and $w_{i,k}^u$ represent the (relative) weights for observation i, which depends on the sets W_k^l and W_k^u respectively, as well as on the choice of kernel. In particular, for the three kernels mentioned above, the (relative) weights may be written as[11]

$$\begin{aligned}
\textbf{Rectangular}: \quad & w_{i,k}^l = \mathbb{I}[(Y_i, \boldsymbol{X_i}) \in W_k^l] \\
& w_{i,k}^u = \mathbb{I}[(Y_i, \boldsymbol{X_i}) \in W_k^u]; \\
\textbf{Triangular}: \quad & w_{i,k}^l = \left(\frac{w}{2} - |X_2 - X_{2,k}^{mid}|\right) \cdot \mathbb{I}[(Y_i, \boldsymbol{X_i}) \in W_k^l] \\
& w_{i,k}^u = \left(\frac{w}{2} - |X_2 - X_{2,k}^{mid}|\right) \cdot \mathbb{I}[(Y_i, \boldsymbol{X_i}) \in W_k^u];
\end{aligned}$$

[11] In the formulae I give for the (relative) weights, I omit the constant of integration since my formulae for $\bar{Y}_k^l$ and $\bar{Y}_k^u$ include the sum of the weights in their denominators.

$$\textbf{Epanechnikov: } w_{i,k}^{l} = \left[\frac{w^2}{4} - (X_2 - X_{2,k}^{mid})^2\right] \cdot \mathbb{I}[(Y_i, \boldsymbol{X_i}) \in W_k^l]$$
$$w_{i,k}^{u} = \left[\frac{w^2}{4} - (X_2 - X_{2,k}^{mid})^2\right] \cdot \mathbb{I}[(Y_i, \boldsymbol{X_i}) \in W_k^u].$$

To summarize, in this subsection I described various plots that one may think are suitable for displaying the discontinuities in outcome for MRD. I discussed why several of these plots are unsuitable, and recommended the "slicing" and "sliding window" plots. The "slicing" plots allow one to separately examine the discontinuities in outcome for different regions along each treatment frontier. The "sliding window" plot provides additional visual evidence on whether these discontinuities are constant along each treatment frontier. Examples of both these two plots based on a realistic simulated dataset are presented in section 5 of this paper.

3.2 Diagnostic Plots

The crucial assumption underpinning validity of conventional RD and MRD designs requires that the conditional expectation functions of potential outcomes be continuous at the cutoffs. Unfortunately, this assumption cannot be tested directly, since potential outcomes for any particular treatment are only observed on one side of the threshold (assuming that the RD or MRD is sharp). Nonetheless, two types of diagnostic plots are often used in conventional RD to test this assumption indirectly. This subsection briefly reviews the diagnostic plots for conventional RD, and describes how they can be adapted for use in MRD.

3.2.1 Plotting a Predetermined Outcome as a Function of the Assignment Variables

The first type of diagnostic graph often used in conventional RD plots a predetermined outcome against the assignment variable. For example, if the outcome is college going and scholarship eligibility is determined by whether one's SAT math score passes a certain threshold, the diagnostic graph may plot a predetermined outcome such as the mother's education level or the student's high school GPA against the student's SAT math score. The motivation is that if the values of observations' assignment variables near the cutoff are genuinely as-good-as-random, then the baseline characteristics of observations just below or above the threshold should not differ systematically.

The discussion in the preceding subsection for plotting outcomes against assignment variables in MRD apply to this diagnostic plot as well. This paper proposes using the "slicing" approach to create these diagnostic plots, with the predetermined variable (instead of the outcome variable of interest) on the vertical axis.

3.2.2 Examining the Density of the Assignment Variables

The second type of diagnostic plot common in conventional RD is a histogram of the assignment variable (where no bin contains the cutoff in its interior). The rectangles immediately to the left and right of the threshold are examined to get a sense of whether the density of the assignment variable is continuous at the cutoff. An obvious discontinuity would indicate that the locations of observations' assignment variable values near the threshold are not as-good-as-random (or in other words, that the assignment variable is manipulable), which calls into question the validity of the RD design.

As an example, suppose that admission to the most prestigious university in a (fictional) country depended on whether the applicant's admission test score passed a certain threshold. The goal is to determine the "value-added" of attending this university, using future earnings as the outcome variable. Suppose (rather pessimistically) also that corruption is rampant in this country, and parents of rich kids anywhere near the threshold are able to bribe officials grading the test into giving their kids a score above the cutoff. In such a case, only kids in relatively poor families will have test scores that fall just under the cutoff, so that the characteristics of applicants just above and below the cutoff are different on average. Hence, the treatment effect estimate may be biased upwards if rich kids tend to have higher earnings on average independent of ability, due to better connections and so forth. The validity of the RD design is clearly violated, and this will show up in the histogram in the form of the rectangle immediately to the right of the cutoff being significantly higher than the one to its left.

While the concept of examining the density of assignment variables at the treatment frontiers in MRD follows straightforwardly from the motivation for the histogram in conventional RD, implementation is much trickier. A three-dimensional histogram encounters the difficulty of the height (which approximately represents the joint density of the assignment variables) being distorted, a problem mentioned in the previous section for graphs showing discontinuities in outcome. This paper suggests two alternatives for displaying the assignment variables' density in MRD with minimal distortion.

The first approach is to display the three-dimensional histogram as a contour plot. The reader can then examine whether the colors of bins on either side of each frontier change drastically (which would indicate a likely discontinuity). The main challenge in creating an informative contour plot is the selection of a suitable color gradient, which is arguably much simpler than the problem of choosing an appropriate "viewpoint" for a three-dimensional histogram.

The second approach is similar in spirit to the "slicing" approach described earlier in this section, with frequency on the vertical axis. The following is a more detailed description of this procedure. Suppose that I intervals are used for X_1, and J intervals for X_2, so that there is a total of at most IJ bins. For each of the I intervals for X_1, one can create a two-dimensional histogram with the J intervals for X_2 on the horizontal axis, and rectangle heights that represent the number of observations with X_1 and X_2 values within the respective intervals.

The same can be done for each unique interval of X_2. For a MDRD design, not all of these $I+J$ two-dimensional histograms are relevant, since only histograms that contain bins for observations receiving different treatments are relevant[12].

Regardless of whether one chooses to use the contour plot or a series of "slicing" histograms to examine the density of the assignment variables, care must be taken in defining the bins so that no point on the treatment frontiers falls in the interior of any bin.

3.2.3 Demonstration of Diagnostic Plots using NCHS Dataset

This example demonstrates the recommended diagnostic plots described in the previous subsection for a MMRD, with neonate birthweights and gestation ages as the assignment variables. Neonates with birthweights or gestation ages below the cutoffs of 1500g and 37 weeks are respectively classified as "very low birth weight" and "premature", and tend to receive extra medical attention as a result. The design for this example is motivated by Almond et al. (2010), who use a conventional RD design with birthweight as the only assignment variable. Almond et al. do not use gestation age as an assignment variable due to worries that it is manipulable. The purpose of this graphical exercise is to determine whether the inclusion of gestation age as a second assignment variable (in addition to birthweight) violates the assumptions underpinning the validity of the MMRD design[13]. The dataset used for these plots is the 2008 Cohort Linked

[12] One may notice that this "slicing" approach is designed only to detect discontinuities in the density as the frontier is approached from a direction parallel to one of the treatment frontiers in the assignment variable space. Hence, it is possible that a discontinuity at some point of the frontier may exist that would not be detectable by this method, as illustrated by the following example.

Suppose that a discontinuity existed at a point $(0, x_2^*)$ along the frontier that is the positive x_2-axis, where $x_2^* > 0$. Further assume that

$$\lim_{x_1 \to 0^+} f(x_1, x_2^*) = \lim_{x_1 \to 0^-} f(x_1, x_2^*),$$

but that

$$\lim_{t_1 \to 0^+} f(-t_1, x_2^* + t_1) \neq \lim_{t_2 \to 0^+} f(t_2, x_2^* + t_2),$$

where $f(x_1, x_2)$ is the joint density function of the assignment variables. This would represent a discontinuity in the joint density function that is not detectable by the "slicing" approach, which can only test whether the limits $\lim_{x_1 \to 0^+} f(x_1, x_2^*)$ and $\lim_{x_1 \to 0^-} f(x_1, x_2^*)$ are equal.

Nonetheless, while it is easy to construct counterexamples in theory, it is difficult to imagine manipulation of the assignment variables in practice that would result in such discontinuities in practice. For instance, if teachers were manipulating reading and math test scores, they would have to do so in a way that for each given reading test score, the proportion of students with math test scores just above and below the threshold remained roughly the same, and vice versa, switching the roles of reading and math. It is hard to think of why agents might be motivated to engage in such contrived manipulation.

[13] One may realize that although I had been assuming so far that the assignment variables have continuous support, the measurement of gestation age in weeks is rather coarse, so that the support of the gestation age variable "is more discrete than continuous". Assignment variables that have discrete rather than continuous support present issues for treatment effect estimation, as Lee and Card (2008) point out. I revisit this issue in greater detail when discussing estimation for MRD in the next section.

Birth/Infant Death Data Set, from the National Center for Health Statistics (NCHS).

The first type of diagnostic graph shown here plots a predetermined outcome – mother's age – against each assignment variable in turn using the "slicing" approach. Linear regression lines are added to the points on each side of the cutoff. Several bandwidths were considered for birthweight and gestation age (25g 50g, 100g and 200g for birthweight and one to five weeks for gestation age). None of these choices resulted in a plot that displayed an obvious discontinuity at the cutoff, which is consistent with a valid MMRD design. The graphs shown in the paper use a bandwidth of 100g for birthweight and one week for gestation age.

The contour plot and the "slicing" histograms, shown next, examine the density of the assignment variables near the cutoff. In both plots, signs of a significant discontinuity at the treatment frontiers are absent. Again, this result is to be expected if the MMRD assumptions are not violated.

While no single one of these diagnostic plots can by itself guarantee the validity of the MMRD design, taken together, they provide some degree of assurance that there are no obvious violations of the MMRD assumptions.

4 Estimation

Most estimands in economics seek to capture a global relationship between covariates and the outcome variable. By contrast, the estimand of interest in RD designs – being the difference between point estimates of two regression functions at their boundaries – is highly local and far more uncertain. As a result, issues such as specification error and boundary effects are particularly relevant for treatment effect estimation in RD designs.

This has led to a substantial body of literature investigating a variety of parametric and nonparametric estimation methods for conventional RD. As survey papers on RD such as Imbens and Lemieux (2008) and Lee and Lemieux (2010) document, some degree of consensus on estimation procedures has developed over time for conventional RD. By contrast, there has been scant research on methods for MRD, and thus, little by way of consensus on MRD estimation.

This section begins by introducing a few of the most common estimation approaches for conventional RD, and discusses their advantages and disadvantages. Then, I describe an estimation method for MMRD proposed by Papay et al. (2011a). I propose a modification of their method for MDRD estimation and suggest a generalization of the cross-validation procedure they define. Finally, I develop a novel estimation method that addresses some of the shortcomings of popular RD estimation approaches, and can be easily implemented for MDRD as well as MMRD.

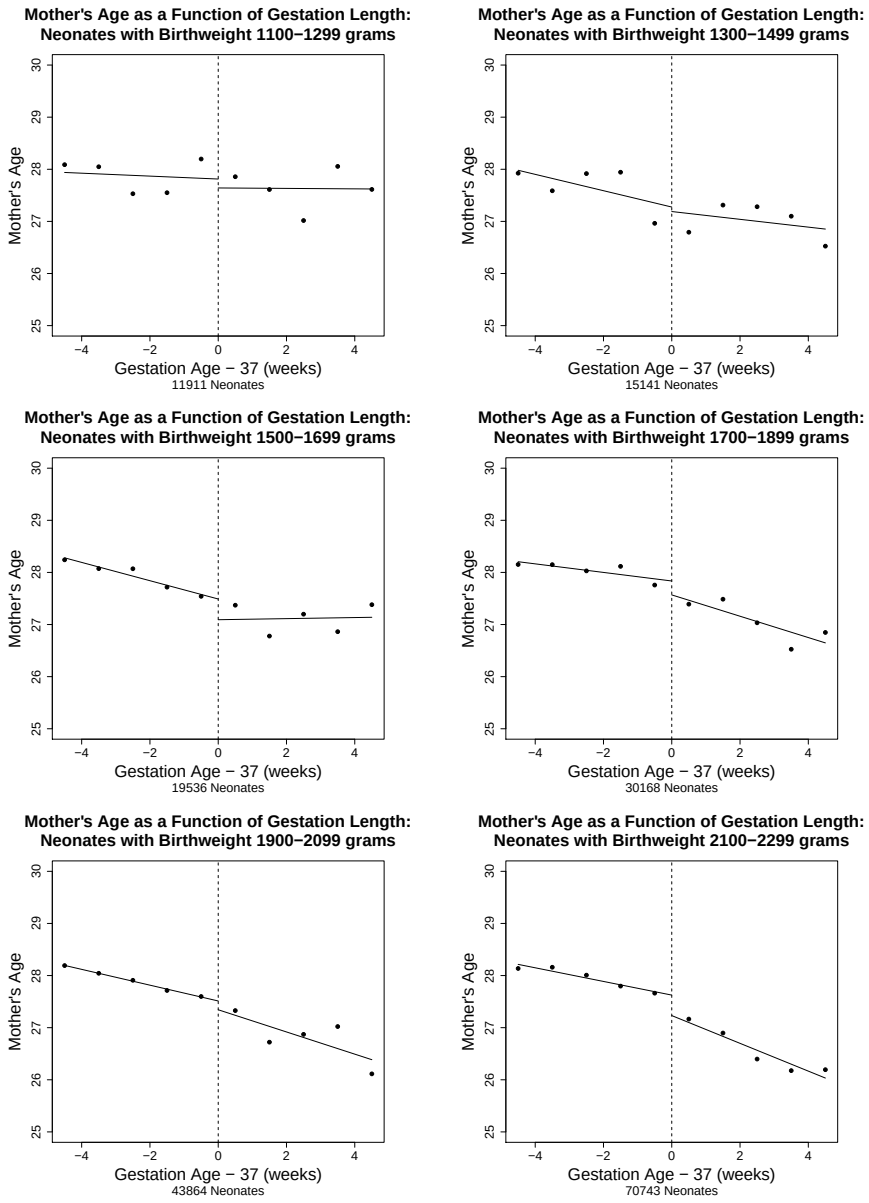

Figure 2: Plots of a predetermined outcome as a function of an assignment variable, created using the "slicing" approach. Each of these graphs considers only neonates with birthweights within a certain interval, and plots mother's age (the predetermined outcome) as a function of gestation age (the assignment variable). A linear regression line is fitted to the points on each side of the cutoff. *Source: National Center for Health Statistics (2008).*

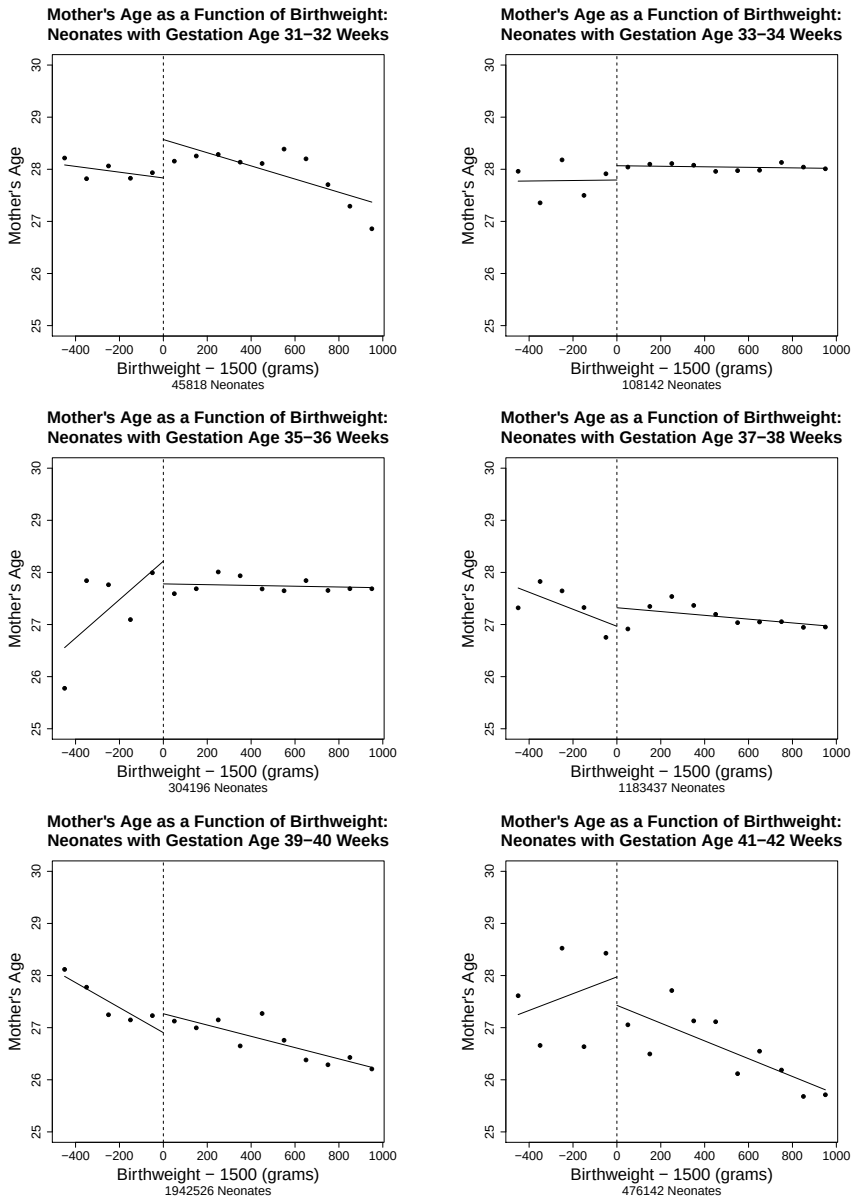

Figure 3: Plots of a predetermined outcome as a function of an assignment variable, created using the "slicing" approach. Each of these graphs considers only neonates with gestation ages within a certain interval, and plots mother's age (the predetermined outcome) as a function of birthweight (the assignment variable). A linear regression line is fitted to the points on each side of the cutoff. *Source: National Center for Health Statistics (2008).*

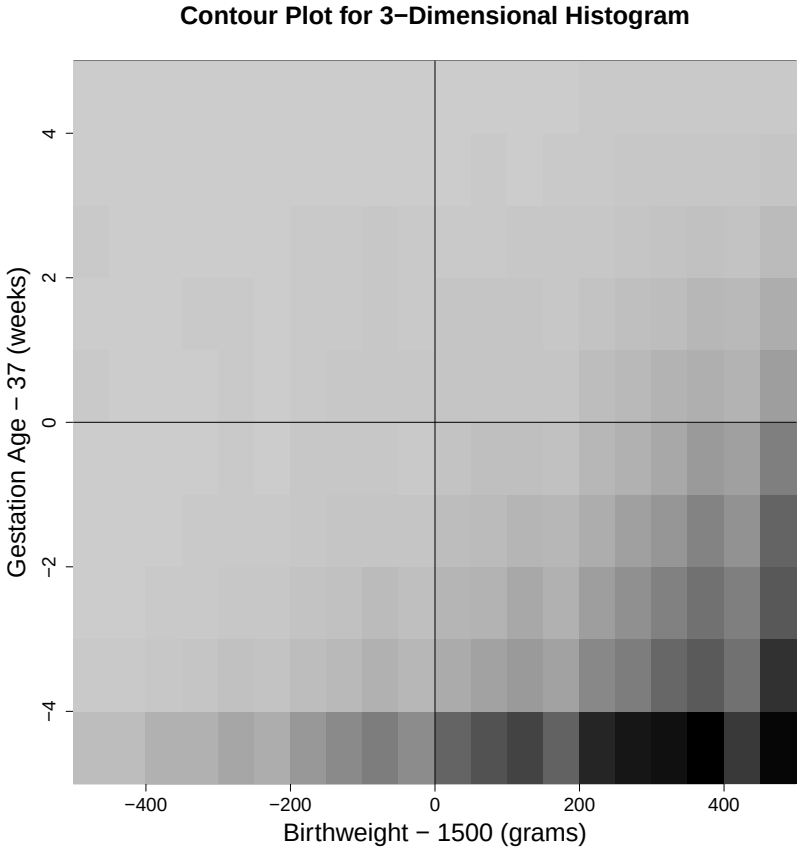

Figure 4: Contour Plot to Examine the Density of the Assignment Variables. Darker regions represent higher frequencies of observations. *Source: National Center for Health Statistics (2008).*

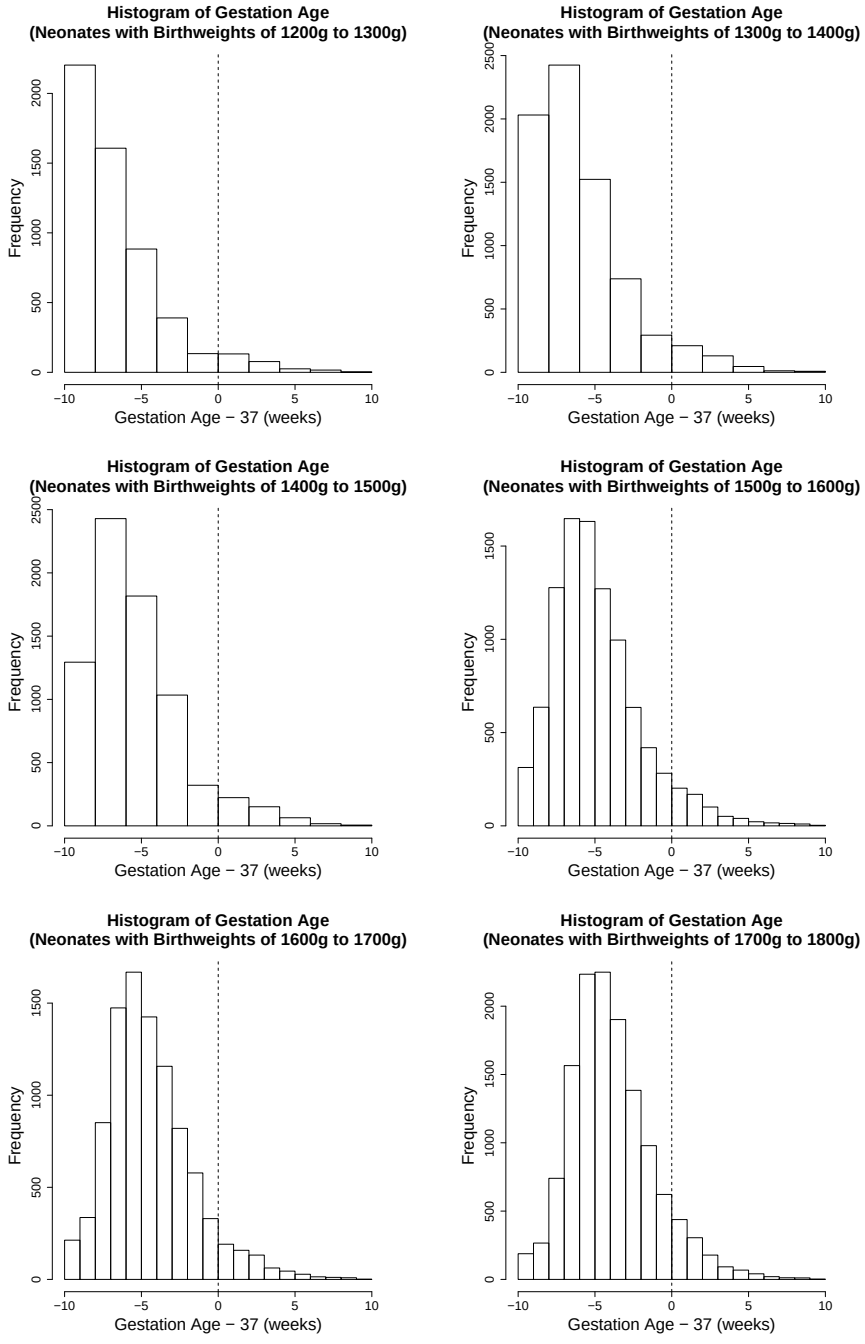

Figure 5: "Slicing Histograms" of Gestation Age for Neonates with Birthweights in Various Intervals. *Source: National Center for Health Statistics (2008).*

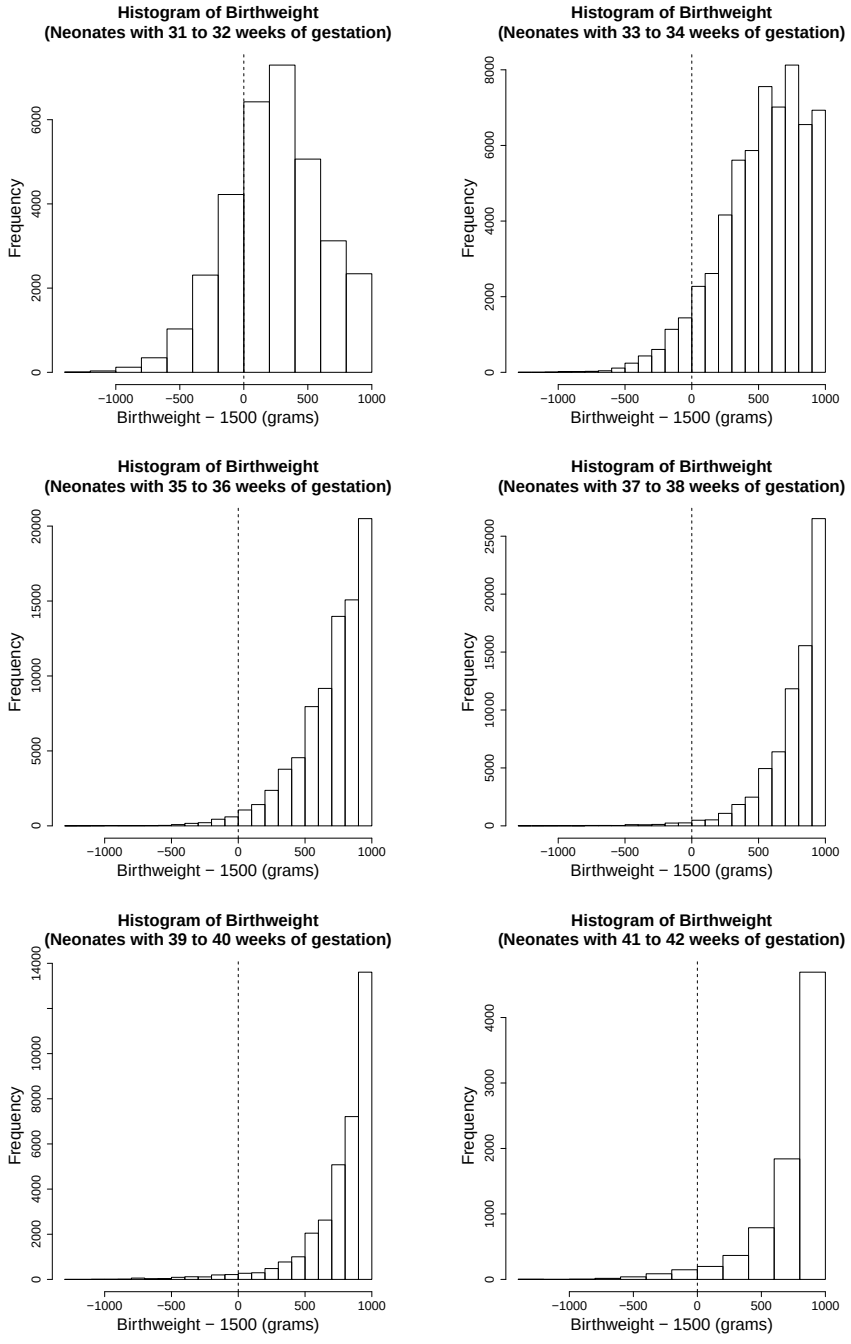

Figure 6: "Slicing Histograms" of Birthweight for Neonates with Various Gestation Ages. *Source: National Center for Health Statistics (2008).*

4.1 Estimation in Conventional RD

A common approach for conventional RD estimation in the past was to fit a global polynomial to each side of the cutoff (with the outcome variable as a function of the assignment variable). This approach has the advantage of being easy to implement, but has come under increasing criticism for various reasons, including sensitivity of the treatment estimate to the order of polynomial, and the boundary effects of higher-order polynomial fits[14].

RD estimation via local linear regression has become increasingly widespread of late, based on properties proved by Fan and Gijbels (1992) and Porter (2003). Yet, while the local linear estimator does not have boundary effects, implementation requires specification of a bandwidth. Generally, choosing an optimal bandwidth involves a variance-bias trade-off. In particular, both the estimator's squared bias and its variance contribute to the expected mean squared error (MSE) of the treatment effect, which one seeks to minimize. Choosing a smaller bandwidth leads to lower bias, but also higher variance due to the smaller number of observations available for estimation, and vice versa for choices of larger bandwidths.

Several methods have been proposed for a systematic way to choose an optimal bandwidth, and one of the more popular approaches that have emerged is that taken by Imbens and Kalyanaraman (2011), henceforth IK. The IK algorithm involves estimating the density of the assignment variable, as well as several orders of derivatives for the mean regression function. The optimal bandwidth implied by the IK algorithm can be written as $h_{opt} = C_K \cdot \Sigma^{1/5} \cdot N^{-1/5}$, where C_K is a constant which depends on the choice of kernel, and Σ is a function of the assignment variable's density as well as the mean regression function. Provided certain assumptions are satisfied, the bias in the local linear treatment effect estimate using this bandwidth tends to zero at a rate of $O_p(N^{-2/5})$.

Asymptotic properties of the IK bandwidth selection algorithm are based on the assumption that the assignment variable is continuous. Yet, most real world datasets only contain variables that are measured in discrete units. Hence, as Lee and Card (2008) note, the bandwidth cannot be made arbitrarily small even as the sample size tends to infinity. Essentially, discrete measurement of the assignment variable implies that there exists a neighborhood around the cutoff with no observations, thus resulting in an "irreducible gap"[15]. In practice, the discrete nature of the assignment variable is unlikely to be a serious issue if the assignment variable takes many possible values (for instance, birthweight measured in grams). However, there are also many RD designs with assignment variables that take relatively few unique values or are measured in coarse intervals (such as age, which is often reported in years). In such cases, it is not clear whether the IK algorithm will result in an optimal bandwidth choice.

[14] For a more detailed discussion about the pitfalls of using global polynomials for RD estimation, see for instance, Gelman and Imbens (2014).

[15] Strictly speaking, it is possible that some observations may have assignment variable values that are exactly equal to the cutoff. However, this does not help with the "irreducible gap" problem since observations on *both* sides of the cutoff are required for estimation.

Moreover, discreteness of assignment variables will likely remain an issue even as increasingly large datasets become available in the age of big data, since the precision of measurements will still be limited by a number of factors, including privacy concerns[16].

4.2 Local Linear Regression for MRD

Most MRD applications in the literature have focused on estimating scalar quantities, such as those described in Wong et al. (2013). Section 2 of this paper argued that important heterogeneities in the treatment effect may be lost when summarizing treatment effects as scalar quantities, and proposed estimating treatment effect functions instead. The only estimation of MRD treatment functions that I am aware of uses a local linear regression approach for MMRD. This method is explained in Papay, Willett and Murnane (2011a), who also implement this estimation in Papay, Willett and Murnane (2011b).

This subsection will begin by describing the MMRD estimation method of Papay et al. (2011a), before introducing a modified version of their method that can be used for MDRD. Then, I will discuss advantages and disadvantages of this estimation approach, and propose a generalization of their bandwidth selection procedure that addresses some (but not all) of its shortcomings.

4.2.1 MMRD Estimation via Local Linear Regression, as described in Papay et al. (2011a)

Papay et al. (2011a) consider estimation of MMRD via local linear regression, with optimal bandwidth (for the two assignment variables) chosen using a generalization of LM CV, the cross-validation (CV) procedure described in Ludwig and Miller (2005), whom I abbreviate as LM. Using the notation introduced in section 2 of this paper, the local linear regression proposed by Papay et al. can be written as:

$$
\begin{aligned}
\mathbb{E}[Y_i|X_{1i}, X_{2i}] =& \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) \\
&+ \beta_4 X_{1i} + \beta_5 X_{2i} + \beta_6 (X_{1i} \times X_{2i}) \\
&+ \beta_7 (X_{1i} \times D_{1i}) + \beta_8 (X_{2i} \times D_{2i}) \\
&+ \beta_9 (X_{1i} \times D_{2i}) + \beta_{10} (X_{2i} \times D_{1i}) \\
&+ \beta_{11} (X_{1i} \times X_{2i} \times D_{1i}) + \beta_{12} (X_{1i} \times X_{2i} \times D_{2i}) \\
&+ \beta_{13} (X_{1i} \times D_{1i} \times D_{2i}) + \beta_{14} (X_{2i} \times D_{1i} \times D_{2i}) \\
&+ \beta_{15} (X_{1i} \times X_{2i} \times D_{1i} \times D_{2i}).
\end{aligned} \tag{10}
$$

This regression equation results in the following four discontinuous linear surfaces, each defined over a quadrant of the assignment variable space:

[16] For example, it is unlikely that date of birth or precise geographical location will be made freely available to researchers, which will be an issue for RD designs that use age or proximity to geographic boundaries as assignment variables.

$$\mathbb{E}[Y_i|D_{1i}=1,D_{2i}=1,X_{1i},X_{2i}] = (\beta_0+\beta_1+\beta_2+\beta_3)+(\beta_4+\beta_7+\beta_9+\beta_{13})X_{1i} + (\beta_5+\beta_8+\beta_{10}+\beta_{14})X_{2i} + (\beta_6+\beta_{11}+\beta_{12}+\beta_{15})(X_{1i}\times X_{2i}),$$

$$\mathbb{E}[Y_i|D_{1i}=0,D_{2i}=1,X_{1i},X_{2i}] = (\beta_0+\beta_2) + (\beta_4+\beta_9)X_{1i} + (\beta_5+\beta_8)X_{2i} + (\beta_6+\beta_{12})(X_{1i}\times X_{2i}),$$

$$\mathbb{E}[Y_i|D_{1i}=0,D_{2i}=0,X_{1i},X_{2i}] = \beta_0+\beta_4X_{1i}+\beta_5X_{2i}+\beta_6(X_{1i}\times X_{2i}),$$

$$\mathbb{E}[Y_i|D_{1i}=1,D_{2i}=0,X_{1i},X_{2i}] = (\beta_0+\beta_1) + (\beta_4+\beta_7)X_{1i} + (\beta_5+\beta_{10})X_{2i} + (\beta_6+\beta_{11})(X_{1i}\times X_{2i}).$$

The treatment functions are obtained by taking the differences between these surfaces along the treatment frontiers F_{12}, F_{23}, F_{34} and F_{14}[17]. Hence, the treatment functions are given by:

$$\begin{aligned}
&\tau_{12}(x_2) = (\beta_1+\beta_3)+(\beta_{10}+\beta_{14})x_2 \quad \text{for } x_2 \geq 0,\\
&\tau_{23}(x_1) = \beta_2+\beta_9x_1 \quad \text{for } x_1 \leq 0,\\
&\tau_{34}(x_2) = \beta_1+\beta_{10}x_2 \quad \text{for } x_2 \leq 0,\\
&\tau_{14}(x_1) = (\beta_2+\beta_3)+(\beta_9+\beta_{13})x_1 \quad \text{for } x_1 \geq 0.
\end{aligned} \tag{11}$$

As is the case for conventional RD, the performance of local linear regression for MRD estimation depends on appropriate bandwidth choice. Papay et al. (2011a) recommend bandwidth selection using a two-dimensional generalization of LM CV, rather than to generalize the IK algorithm to higher dimensions. The authors cite concerns over unknown properties of the IK algorithm when the assignment variable has discrete support when explaining their choice to use LM CV.

LM CV is a type of CV that is specifically designed for estimation of a boundary point. The motivation for LM CV is the fact that estimation of mean potential outcomes at the treatment frontiers only uses observations on one side of the frontier. Since LM CV was originally designed for conventional RD, Papay et al. (2011a) use a two-dimensional generalization to jointly select the bandwidths h_1^* and h_2^* for the assignment variables X_1 and X_2 respectively.

To elaborate, denote the candidate (joint) bandwidth under consideration in a MMRD by (h_1, h_2), and suppose that the fitted value $\hat{Y}_{i^*}(h_1,h_2)$ for an observation $(Y_{i^*}, \boldsymbol{X_{i^*}})$ is desired. Also, assume that $\boldsymbol{X_{i^*}}$ lies in the first quadrant, R_1 of the assignment variable space, so that the frontiers relevant to $(Y_{i^*}, \boldsymbol{X_{i^*}})$ are F_{12} and F_{14}. Estimation of the mean potential outcome for a point in R_1

[17]Strictly speaking, in order to compute these differences, the surfaces need to be extended continuously so that their domains of definition include the treatment frontiers.

arbitrarily close to these treatment frontiers will typically only use points to the "north" or "east" of it. Therefore, in order to mimic the estimation of a boundary point, instead of using all points that are "close" to $(Y_{i^*}, \boldsymbol{X_{i^*}})$ in the assignment variable space,

$$\{(Y_j, \boldsymbol{X_j}) : |X_{1j} - X_{1i^*}| \leq h_1 \text{ and } |X_{2j} - X_{2i^*}| \leq h_2\} \quad \setminus \quad \{(Y_{i^*}, \boldsymbol{X_{i^*}})\}$$

to estimate $\hat{Y}_{i^*}(h_1, h_2)$ as one might do for standard CV, only points that are "close" to $(Y_{i^*}, \boldsymbol{X_{i^*}})$ *and* to the "northeast" of $(Y_{i^*}, \boldsymbol{X_{i^*}})$ are used, i.e.

$$\{(Y_j, \boldsymbol{X_j}) : 0 \leq X_{1j} - X_{1i^*} \leq h_1 \text{ and } 0 \leq X_{2j} - X_{2i^*} \leq h_2\} \quad \setminus \quad \{(Y_{i^*}, \boldsymbol{X_{i^*}})\}.$$

To simplify notation, denote this set by $S_{i^*}(h_1, h_2)$. The following plot clarifies this concept by considering four points – one in each quadrant of the assignment variable space – and the regions determining $S_i(h_1, h_2)$ for each point according to this bandwidth selection procedure.

After determining $S_{i^*}(h_1, h_2)$ for observation $(Y_{i^*}, \boldsymbol{X_{i^*}})$ for a given bandwidth, the fitted value for this point, $\hat{Y}_{i^*}(h_1, h_2)$, is obtained via a linear regression that only uses points in $S_{i^*}(h_1, h_2)$[18]. To be explicit, one first estimates the OLS regression

$$Y_i = \hat{\gamma_0} + \hat{\gamma_1} X_{1i} + \hat{\gamma_2} X_{2i} + \hat{\gamma_3}(X_{1i} \times X_{2i}) + \hat{\epsilon} \quad \text{for } (Y_i, \boldsymbol{X_i}) \in S_{i^*}(h_1, h_2), \tag{12}$$

and then obtains the fitted value for $(Y_{i^*}, \boldsymbol{X_{i^*}})$ using the formula

$$\hat{Y}_{i^*}(h_1, h_2) = \hat{\gamma_0} + \hat{\gamma_1} X_{1i^*} + \hat{\gamma_2} X_{2i^*} + \hat{\gamma_3}(X_{1i^*} \times X_{2i^*}). \tag{13}$$

The MSE for each candidate bandwidth (h_1, h_2)

$$MSE(h_1, h_2) = \frac{1}{N} \sum_{i=1}^{N} (\hat{Y}_i(h_1, h_2) - Y_i)^2 \tag{14}$$

is calculated, and the bandwidth resulting in the lowest MSE is chosen as the optimal bandwidth (h_1^*, h_2^*).

Finally, the local linear regression is estimated using the subset of observations

$$\{(Y_i, \boldsymbol{X_i}) : |X_{1i}| \leq h_1^* \text{ or } |X_{2i}| \leq h_2^*\}.$$

[18]The procedure described in this section uses a rectangular kernel (i.e. OLS regression), as Papay et al. do, for simplicity of exposition. The estimation can easily be modified to accommodate other kernel choices, by using a weighted least squares regression with weights that depend on the choice of kernel. Two other popular choices of kernel (triangular and Epanechnikov) were discussed in the previous section on graphical analysis. An application of local linear estimation for MMRD that does not use a rectangular kernel can be found in Snider and Williams (2015). Incidentally, Snider and Williams mention in a footnote that their attempt at bandwidth selection via cross-validation was unsuccessful, without providing details about their implementation method.

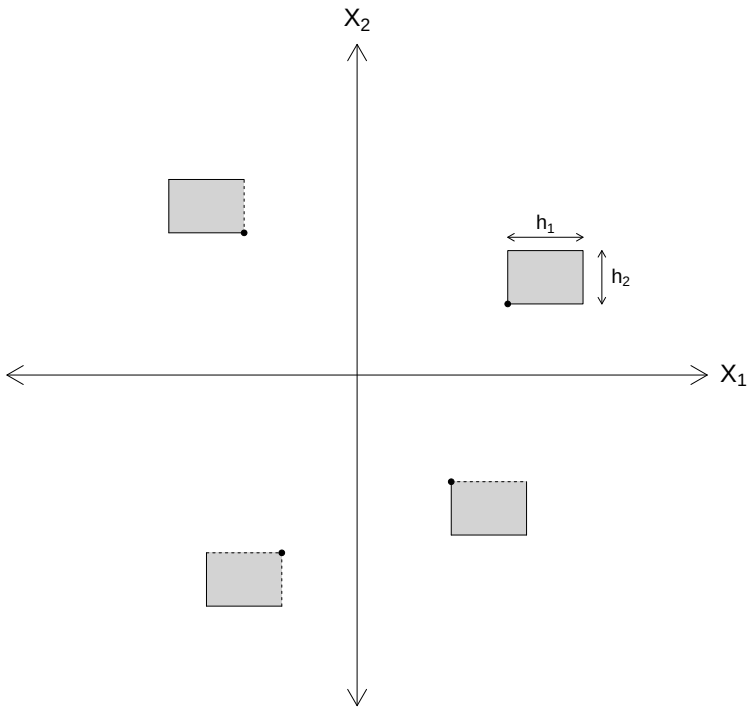

Figure 7: This figure shows the regions (shaded rectangles) determining $S_i(h_1, h_2)$ for four (solid black) points, as defined by the bandwidth selection procedure described in Papay et al. (2011a). In addition to the candidate bandwidth (h_1, h_2), the set of observations that are used to estimate the fitted value of a point is also determined by the specific treatment region that the point lies in. The determination of the inclusion or exclusion of a side of the rectangle in the relevant region reflects how the treatment conditions are defined along the treatment frontiers.

4.2.2 MDRD Estimation via Local Linear Regression

While the estimation approach by Papay et al. (2011a) that I just described is meant for MMRD (with four treatments), it can easily be modified for MDRD. In particular, there is no reason to revise the bandwidth selection procedure for MDRD, so the only change needed is to tweak the local linear regression function appropriately[19].

I retain notation introduced earlier in the text, so that the dummy variable for receiving treatment in MDRD is $W_i = D_{1i} \times D_{2i}$. The local linear regression equation for MDRD is thus:

$$\begin{aligned}(15)\quad \mathbb{E}[Y_i|X_{1i}, X_{2i}] &= \beta_0 + \beta_1 W_i + \beta_2 X_{1i} + \beta_3 X_{2i} \\ &+ \beta_4(X_{1i} \times X_{2i}) + \beta_5(W_i \times X_{1i}) + \beta_6(W_i \times X_{2i}) + \beta_7(W_i \times X_{1i} \times X_{2i}).\end{aligned}$$

This regression equation results in the following two discontinuous linear surfaces, the first being defined over the non-negative quadrant R_1, and the second being defined over the rest of the assignment variable space $R_2 \cup R_3 \cup R_4$:

$$\begin{aligned}\mathbb{E}[Y_i|W_i = 1, X_{1i}, X_{2i}] &= (\beta_0 + \beta_1) \\ &+ (\beta_2 + \beta_5)X_{1i} + (\beta_3 + \beta_6)X_{2i} + (\beta_4 + \beta_7)(X_{1i} \times X_{2i}),\end{aligned}$$

$$\mathbb{E}[Y_i|W_i = 0, X_{1i}, X_{2i}] = \beta_0 + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4(X_{1i} \times X_{2i}).$$

The treatment functions are obtained by taking the differences between these surfaces along the treatment frontiers F_1 and F_2, i.e. the non-negative x_1- and x_2-axes[20]. Hence, the treatment functions are given by:

$$\begin{aligned}(16)\quad &\tau_1(x_2) = \beta_1 + \beta_6 x_2 \text{ for } x_2 \geq 0, \\ &\tau_2(x_1) = \beta_1 + \beta_5 x_1, \text{ for } x_1 \geq 0.\end{aligned}$$

The subset of points used for this local linear regression is

$$\{(Y_i, \boldsymbol{X_i}) : |X_{1i}| \leq h_1^* \text{ or } |X_{2i}| \leq h_2^*\} \cap \{(Y_i, \boldsymbol{X_i}) : X_{1i} \geq -h_1^* \text{ and } X_{2i} \geq -h_2^*\}.$$

4.2.3 Advantages and Disadvantages of Local Linear Regression for MRD Estimation

Some advantages of local linear regression for boundary estimation (based on the estimator's asymptotic properties) were mentioned earlier in this paper.

[19] This change in the regression function is required due to differences in the treatment frontiers for MDRD and MMRD. Specifically, the union of the treatment frontiers ($F_{12} \cup F_{23} \cup F_{34} \cup F_{14}$) for the latter comprises of the entire x_1- and x_2-axes, so the regression function allows for discontinuities along all of the two axes. By contrast, the union of the treatment frontiers for the latter ($F_1 \cup F_2$) comprises of only the non-negative part of the two axes, so it would not make sense for the regression function to be discontinuous along the negative parts of the axes (since there is no treatment effect to be estimated there). The local linear regression function that I introduce for MDRD ensures that the regression function is only allowed to be discontinuous along the treatment frontiers F_1 and F_2.

[20] As in the case for MMRD, in order to compute these differences, the surfaces need to be extended continuously so that their domains of definition include the treatment frontiers.

Another advantage of this approach is that the standard regression outputs – the estimated coefficients and their covariance matrix – are very convenient for hypothesis testing.

For instance, consider a MDRD estimation, and denote the vector of coefficient estimates and its covariance matrix respectively by $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}$ (with rows and columns indexed from 0 to 7, in order to match the indices of the coefficients). For simplicity of exposition, assume that the error term is normally distributed. To obtain point-wise confidence intervals for the treatment function, I first denote, for a given value of $x_2 \geq 0$, the random variable representing the treatment effect estimate at the point by T, and express the estimated treatment effect as

$$\hat{\tau}_1(x_2) = \boldsymbol{c}'\hat{\boldsymbol{\beta}}, \text{ where } \boldsymbol{c}' = [0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad x_2 \quad 0].$$

The (approximate) distribution of T is thus given by $T \sim N(\boldsymbol{c}'\hat{\boldsymbol{\beta}}, \boldsymbol{c}'\hat{\boldsymbol{\Sigma}}\boldsymbol{c})$[21], which allows for easy hypothesis testing of whether the estimated treatment function at a given point is statistically different from zero (at a specified significance level).

Another hypothesis test of interest is whether there is statistical evidence of a non-constant treatment effect. In this example, the hypothesis test amounts to whether the confidence interval for $\hat{\beta}_6$ contains zero, which is trivial since the estimated coefficient is (approximately) normally distributed and its variance is given in the regression output as $\hat{\boldsymbol{\Sigma}}_{6,6}$.

However, the attractive theoretical properties of estimation via local linear regression are predicated on appropriate bandwidth choice. In practice, the task of selecting a suitable bandwidth has been a real difficulty which has not even been fully resolved for conventional RD. The uncertainty over bandwidth choice is exacerbated in MRD since the bandwidths for multiple assignment variables must be jointly selected, making this a multidimensional problem.

The bandwidth selection procedure just described is unsatisfactory in various ways. As LM (2005) themselves point out, CV estimates of the loss function are typically relatively flat. LM take this as an indication of a more general problem, that asymptotic properties of CV methods imply extremely slow rates of convergence. Moreover, the CV procedure documented by Papay et al. (2011a) has several other shortcomings, which I discuss below.

The first concerns a technical issue that the earlier description glosses over. The problem is that for a candidate bandwidth (h_1, h_2), the set of observations $S_{i^*}(h_1, h_2)$ that are used to obtain a fitted value for $(Y_{i^*}, \boldsymbol{X_{i^*}})$ may not contain enough observations with unique combinations of the $\boldsymbol{X_i}$ to fit the linear regression, so that predicted values for these points may be undefined[22]. While

[21] This distributional result is only an approximation because $\hat{\boldsymbol{\beta}}$ follows a t-distribution, rather than a normal distribution. However, this approximation is likely to be good even for moderate sample sizes.

[22] In fact, this will always be the case for points. For instance, consider the point $\boldsymbol{X_i}$ in the non-negative quadrant R_1 with the largest values of X_{1i} and X_{2i}. By definition, there are no points to the northeast of this particular point, and thus, $S_i(h_1, h_2)$ is empty.

Papay et al. (2011a) do not mention this problem in their discussion, I take the approach of discarding these points and computing the MSE over the remaining points[23].

Second, as pointed out by IK (2011), this procedure implicitly selects a bandwidth that is optimal for fitting the mean regression function over the entire assignment variable space, rather than simply close to the treatment frontiers. To see why this may be problematic, consider an example where there is a greater density of observations near the treatment frontiers than further away, and denote the true optimal bandwidth by (h_1^{opt}, h_2^{opt}). The sparse points that are far away from the treatment frontier may in fact cause the procedure described by Papay et al. to select a bandwidth (h_1^*, h_2^*) that is larger than (h_1^{opt}, h_2^{opt}). This is because the true optimal bandwidth (h_1^{opt}, h_2^{opt}) for estimation at the treatment frontiers is too small for these points (due to the sparseness of points in their neighborhoods), leading to the selection of a larger bandwidth as a compromise. Heteroskedasticity may also result in a biased bandwidth choice.

Third, the bandwidth selection procedure is computationally expensive. For each candidate bandwidth (h_1, h_2), the algorithm involves a loop over all observations in the dataset. Within this loop, for each observation i, the algorithm must determine $S_i(h_1, h_2)$, fit a local linear regression using points in this set and obtain the fitted value $\hat{Y}_i(h_1, h_2)$. Moreover, the number of potential bandwidth choices is large, since the search for an optimal bandwidth is being conducted on a multidimensional grid.

4.2.4 Generalization of Bandwidth Selection Procedure described in Papay et al. (2011a)

In order to mitigate some of these problems, I propose a modification of the CV procedure described in Papay et al. (2011a). In fact, the method I propose is a more general version of the CV procedure just discussed, and is closer in spirit to the original LM CV approach for conventional RD.

The original method implemented in LM (2005) does not compute the MSE over all points, as Papay et al. (2011a) do. Instead, only observations within five percentage points of either side of the cutoff are used for bandwidth selection. IK (2011) consider a slight generalization of LM CV by introducing an additional

[23] It is not obvious that discarding points for which fitted values cannot be obtained (for a given candidate bandwidth) is the "right" thing to do. Consider a point that does not have "extreme" values of X_{1i} and X_{2i} (e.g. not in the northeast corner of R_1, the northwest corner of R_2, and so forth), and suppose that there are insufficiently many points in $S_i(h_1, h_2)$ to estimate its fitted value. This is in fact a sign that for this point at least, the candidate bandwidth is "too small". Hence, by ignoring such points when computing MSE, useful information for bandwidth choice is lost.

One possible way to deal with this issue is to incorporate a penalty term for points which have undefined fitted values, and to modify the criterion function (currently the MSE), to be a linear combinations of the sum of squared errors and the penalty term. However, this introduces another layer of complexity into a bandwidth selection procedure that is already rather computationally burdensome, and does not address the method's other shortcomings mentioned in the text.

parameter δ which specifies the percentage of points to use on either side of the cutoff for bandwidth selection.

This idea of using only a proportion of points close to the threshold for bandwidth selection does not extend neatly to MRD. For instance, suppose one decides to use δ percent of the points in R_1 that are close to the treatment frontiers. Since there are two relevant frontiers for this region (the non-negative x_1- and x_2- axes), it is not obvious how to devise an objective method that allocates this limited quota of points "fairly" between regions in R_1 close to each frontier, as well as along each frontier.

Therefore, I take the approach of using the quantiles for each assignment variable in each of the four quadrants of the assignment variable space. Specifically, for a chosen $\delta \in (0, 1]$, I define the following quantiles:

- Let p_1 and q_1 be the δth quantiles of X_{1i} and X_{2i} respectively, for observations in R_1 (i.e. observations with $D_{1i} = 1$ and $D_{2i} = 1$).
- Let p_2 and q_2 be the $(1-\delta)$th and δth quantiles of X_{1i} and X_{2i} respectively, for observations in R_2 (i.e. observations with $D_{1i} = 0$ and $D_{2i} = 1$).
- Let p_3 and q_3 be the $(1 - \delta)$th quantiles of X_{1i} and X_{2i} respectively, for observations in R_3 (i.e. observations with $D_{1i} = 0$ and $D_{2i} = 0$).
- Let p_4 and q_4 be the δth and $(1-\delta)$th quantiles of X_{1i} and X_{2i} respectively, for observations in R_4 (i.e. observations with $D_{1i} = 1$ and $D_{2i} = 0$).

For MMRD, the set of observations used for bandwidth selection is

$$\bigcup_{k=1}^{4} \Big(\{(Y_i, \boldsymbol{X_i}) : |X_{1i}| \leq |p_k| \text{ and } |X_{2i}| \leq |q_k|\} \cap R_k \Big).$$

For MDRD, the precise definition is slightly messier, although it is also the case that only observations close to the treatment frontiers (as defined by p_k and q_k) are used:

$$\begin{aligned} &\Big(\{(Y_i, \boldsymbol{X_i}) : |X_{1i}| \leq |p_1| \text{ or } |X_{2i}| \leq |q_1|\} \cap R_1 \Big) \\ &\quad \bigcup \Big(\{(Y_i, \boldsymbol{X_i}) : X_{1i} \geq p_2\} \cap R_2 \Big) \\ &\quad \bigcup \Big(\{(Y_i, \boldsymbol{X_i}) : |X_{1i}| \leq |p_3| \text{ and } |X_{2i}| \leq |q_3|\} \cap R_3 \Big) \\ &\quad \bigcup \Big(\{(Y_i, \boldsymbol{X_i}) : X_{2i} \geq q_3\} \cap R_4 \Big). \end{aligned}$$

While this parameter δ does not represent the proportion of points in each treatment region being used, it still controls the amount of data close to the treatment frontiers that is used for bandwidth selection. For instance, the procedure described in Papay et al. (2011a) that uses all observations corresponds to the special case of $\delta = 1$.

By only using points that are "close" ("closeness" being controlled by δ) to the treatment frontiers for estimation, the problems mentioned above are ameliorated to a certain extent. For suitably small choices of δ, observations that "too far" from the treatment frontiers are no longer used for bandwidth selection. Moreover, the computational burden for bandwidth selection is eased slightly since this method uses fewer observations.

However, this modification does not solve other problems with CV mentioned earlier. Furthermore, this method raises an additional question concerning the choice of δ. Both LM and IK use subjective judgment to determine the proportion of data near the threshold to use. The lack of an objective method to determine δ is especially problematic if the bandwidth selected by this approach turns out to be sensitive to the choice of δ.[24]

The discussion in this subsection reveals that, despite the attractive properties of local linear regressions for boundary estimates, performance in practice can be hampered by inappropriate bandwidth choice. Procedures for bandwidth selection suffer from various problems, which tend to be exacerbated when these methods are extended from conventional RD to MRD.

In light of these difficulties, this paper proposes a non-parametric method for MRD estimation that does not require bandwidth selection – thin plate regression splines. The method is closely related to thin plate splines, which is the multidimensional analogue of a penalized spline method that Rau (2011) proposed for estimation in conventional RD. The next subsection covers MRD estimation via thin plate regression splines.

4.3 Thin Plate Regression Splines

A thin plate spline has various properties that make it attractive for MRD estimation. First, it is a local estimator (unlike global polynomials), which makes it less susceptible to boundary effects. Second, the flexibility of this method allows for easy estimation of non-constant treatment effects. In particular, one does not need to specify a functional form for the treatment function estimate. Finally, it is far more convenient to fit than other popular local smoothing methods. For instance, local linear regression and cubic splines require bandwidth and knot selection respectively, in order to control the smoothness of the fitted surface (i.e. to avoid over-fitting or over-smoothing). In the context of MRD, both bandwidth and knot selection require a grid search in multiple dimensions,

[24] One possible way to choose δ objectively is to treat δ as an additional tuning parameter, and jointly determine the optimal bandwidth as well as δ using the CV method just described. Unfortunately, in addition to imposing significant additional computational burden (since the inclusion of δ as a tuning parameter adds yet another dimension to the grid search for optimal parameter values), this approach may not yield a good choice of δ or bandwidth. For instance, suppose that "irrelevant" points that are far from the treatment frontiers come from a data-generating process that is much closer to linear than for points closer to the frontiers. Including these "irrelevant" points may reduce MSE significantly, so the resulting choice of δ is larger than optimal. Hence, even when δ is used as a tuning parameter, "irrelevant" observations far from the frontiers can have substantial influence on bandwidth selection.

which can be computationally expensive. By contrast, the smoothness of thin plate splines is controlled by a single scalar tuning parameter.

4.3.1 Thin Plate Splines

The concept of thin plates splines is first introduced in Duchon (1977), and is later revisited in Wood (2003) and Wood (2006). A thin plate spline seeks to estimate a function of multiple variables from noisy data, with the smoothness of this function being controlled by a penalty term. More formally, given N observations $(y_i, \boldsymbol{x_i})$, $\boldsymbol{x_i} \in \mathbb{R}^d$ from a data-generating process

$$y_i = g(\boldsymbol{x_i}) + \epsilon_i,$$

where g is a smooth function and ϵ_i are random errors, the smoothing spline $\hat{f}_i$ is the estimate of g that minimizes the following quantity:

$$\sum_{i=1}^{N}(y_i - f(\boldsymbol{x_i}))^2 + \lambda J_{md}(f). \tag{17}$$

This minimand is the sum of a loss function and a penalty term, with J_{md} in the penalty term defined by:

$$J_{md} \equiv \int ... \int_{\mathbb{R}^d} \sum_{\nu_1+...+\nu_d=m} \frac{m!}{\nu_1!...\nu_d!}\left(\frac{\partial^m f}{\partial x_1^{\nu_1}...x_d^{\nu_d}}\right)^2 dx_1...dx_d. \tag{18}$$

The choice of m is required to satisfy $2m > d$, although typically $2m > d + 1$ is also respected so that the estimated function is "visually smooth". If one is dealing with two assignment variables ($d = 2$), then $m = 2$ would be a good choice, in which case the penalty term is:

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

The coefficient on the penalty term, λ, is a tuning parameter that is typically chosen by generalized cross-validation (GCV), which is a modification to another popular choice – leave-out-out-cross-validation (LOOCV). GCV is often preferred to LOOCV because it is less computationally expensive, and is invariant to rotation of the outcome vector and basis matrix[25].

[25] In fact, the popularity of LOOCV (relative to k-fold cross-validation for instance) is also partly due to computational reasons. It turns out that instead of fitting N different models to find the LOOCV score for a particular choice of λ, there is an easy formula which only requires a single model fit,

$$LOOCV(\lambda) = \frac{1}{N}\sum_{i=1}^{N}\frac{[y_i - (Ay)_i]^2}{(1 - A_{ii})^2},$$

where $\boldsymbol{A}$ is the influence matrix for the model fit using λ (so that $\boldsymbol{A}$ is actually a function of λ, although the notation does not reflect this). The GCV score is given by a similar formula:

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

4.3.2 Thin Plate Regression Splines

Despite the advantages of thin plate splines outlined at the start of this subsection, a major disadvantage is its computational cost, which remains a major obstacle to its widespread use in practical statistical work. At present, thin plate splines are most commonly used in the special case of $d = 1$ (where the method is known by a number of different names such as "penalized splines" or "smoothing splines") since there exists an efficient algorithm for this special case[26]. Denoting the number of unique combinations of covariate values by n (which may be smaller than the total number of observations N due to discrete measurements of data), the computational cost for penalized splines is on the order of $O(n)$ as opposed to $O(n^3)$ for the general case. Hence, the primary motivation for the construction of thin plate regression splines (henceforth TPRS) is to provide a relatively computationally inexpensive alternative to thin plate splines.

A major reason for the computational expense in fitting thin plate splines is that the basis matrix is of rank n. Yet, this seems "wasteful" in the sense that the effective degrees of freedom[27] tends to be a small proportion of n. The idea behind TPRS is to approximate thin plate splines using a low rank basis matrix (of say, rank k), obtained by truncating the most "wiggly" components of the thin plate spline (which were penalized heavily anyway in the thin plate spline fit), while leaving those components with "zero wiggliness" untouched[28]. This modification results in significant computational cost savings – the cost of fitting a TPRS is at most $O(kn^2)$, as opposed to $O(n^3)$ for thin plate splines. Moreover, the computationally costly model selection algorithm is only $O(k^2n)$ for TPRS, compared to $O(n^3)$ for thin plate splines.

When fitting a TPRS, one must choose the basis dimension k, in addition to the smoothness parameter λ, so it may seem that tuning the model for TPRS is more tedious than for thin plate splines. However in practice, for a given value of k, the actual effective degrees of freedom is controlled by λ. So, the choice of k is not critical as long as it is chosen large enough so that the model is not overly restricted by the basis dimension (i.e. k should be chosen to be larger than the effective degrees of freedom believed to be required). Kim and Gu (2004) show that basis dimension should scale as $n^{2/9}$, and suggest $10n^{2/9}$ based on simulations. The appendix in this paper provides an example exploring the sensitivity of TPRS treatment effect estimates to the choice of k.

Standard errors for the model parameters in TPRS can be derived using a Bayesian approach to uncertainty estimation, assuming a fixed value of the smoothing parameter λ.[29] If the error term in the data-generating process is normal, then the posterior distribution of the model parameters is multivariate normal. Otherwise, approximate normality of the posterior distribution is

[26] Rau (2011) discusses an estimation approach for conventional RD using penalized splines (i.e. thin plate splines with $d = 1$).

[27] The effective degrees of freedom is defined as the trace of the influence matrix.

[28] Interested readers should refer to the appendix for additional details on the approximation of thin plate splines by TPRS.

[29] Readers may refer to Wood (2006) for a derivation of this result.

justified by large sample theory.

4.3.3 MRD Estimation via TPRS

MRD estimation via TPRS is easy to implement. First, one would fit a separate TPRS for observations in each treatment region in the assignment variable space. Point estimates of the treatment effect along each treatment frontier can thus be obtained by taking the difference between fitted values (from the two TPRS functions estimated using data in the two adjacent treatment regions).

To avoid any ambiguity, I explicitly state the formulae for the estimated treatment effect functions, starting with those for MDRD. Denote the fitted TPRS functions by $\hat{f}_1(x_1, x_2)$ and $\hat{f}_0(x_1, x_2)$, these being the estimates of

$$\mathbb{E}[Y(1)|X_1 = x_1, X_2 = x_2)] \text{ and } \mathbb{E}[Y(0)|X_1 = x_1, X_2 = x_2)],$$

using data from R_1 and $\mathbb{R}^2 \setminus R_1$ respectively. The MDRD treatment effect estimates are thus given by:

$$\begin{aligned} &\hat{\tau}_2(x_2) \equiv \hat{f}_1(0, x_2) - \hat{f}_0(0, x_2) \text{ for } x_2 \geq 0, \\ &\hat{\tau}_2(x_1) \equiv \hat{f}_1(x_1, 0) - \hat{f}_0(x_1, 0) \text{ for } x_1 \geq 0. \end{aligned} \tag{19}$$

Next for MMRD, I denote the fitted TPRS functions for observations in regions R_1, R_2, R_3 and R_4 respectively by $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$ and $\hat{f}_4$. Using this notation, the treatment effect estimates for MMRD can be expressed as:

$$\begin{aligned} &\hat{\tau}_{12}(x_2) \equiv \hat{f}_1(0, x_2) - \hat{f}_2(0, x_2) \text{ for } x_2 \geq 0, \\ &\hat{\tau}_{23}(x_1) \equiv \hat{f}_3(x_1, 0) - \hat{f}_2(x_1, 0) \text{ for } x_1 \leq 0, \\ &\hat{\tau}_{34}(x_2) \equiv \hat{f}_4(0, x_2) - \hat{f}_3(0, x_2) \text{ for } x_2 \leq 0, \\ &\hat{\tau}_{14}(x_1) \equiv \hat{f}_1(x_1, 0) - \hat{f}_4(x_1, 0) \text{ for } x_1 \geq 0. \end{aligned} \tag{20}$$

Approximate Bayesian confidence intervals for the point estimates of these treatment effect functions can be obtained using the posterior distributions of point estimates from the fitted TPRS functions. For concreteness, I take the estimation of $\tau_1(x_2)$ in MDRD as an example, although the procedures for $\tau_2(x_1)$ as well as for MMRD are completely analogous. Following the Bayesian framework, I consider the parameters

$$f_1(0, x_2) = \mathbb{E}[Y(1)|X_1 = 0, X_2 = x_2)] \text{ and } f_0(0, x_2) = \mathbb{E}[Y(0)|X_1 = 0, X_2 = x_2)]$$

as random variables, which I denote by V_1 and V_0 respectively. This yields

$$V_1 \sim N(\hat{f}_1(0, x_2), \sigma_1^2) \text{ and } V_0 \sim N(\hat{f}_0(0, x_2), \sigma_0^2),$$

where the moments of these normal distributions are known. It follows that

$$V_1 - V_0 \sim N(\hat{f}_1(0, x_2) - \hat{f}_0(0, x_2), \sigma^2),$$

where $\sigma^2 = \sigma_1^2 + \sigma_0^2 - 2Cov(V_1, V_0)$. Although the value of $Cov(V_1, V_2)$ is unknown, an upper bound for σ^2 can still be found by using the inequality $Cov(V_1, V_0) \geq -\sigma_1\sigma_0$, so that a conservative Bayesian confidence interval for $\hat{\tau}_1(x_2)$ may be obtained.

On the other hand, the assumption of maximum negative covariance between V_1 and V_0 may be too pessimistic. One may think that the correlation between expected potential outcomes at any point on the frontier should be positive. Hence, a more optimistic confidence interval may be obtained by assuming zero covariance (and hence independence, due to normality) between V_1 and V_0.[30]

Hence, to recap the discussion about point-wise confidence intervals for treatment effect estimates along the frontiers, one may either use a conservative Bayesian confidence interval which assumes, for $\hat{\tau}_1(x_2)$ in a MDRD, that

$$V_1 - V_0 \sim N(\hat{f}_1(0, x_2) - \hat{f}_0(0, x_2), (\sigma_1 + \sigma_0)^2),$$

or a more optimistic confidence interval which assumes

$$V_1 - V_0 \sim N(\hat{f}_1(0, x_2) - \hat{f}_0(0, x_2), \sigma_1^2 + \sigma_0^2).$$

5 Simulation

In this section, I conduct an extensive simulation study based closely on the NBER paper by Kane (2003). Methods for graphical analysis and estimation described earlier in this paper are demonstrated, and the strengths and weaknesses of each approach assessed.

5.1 Background on Original Paper

The design of this simulation study is based on the NBER paper by Kane (2003), titled "A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going". The paper investigates how the CalGrant program affected college going rates in California from 1998 to 1999 using a sharp regression-discontinuity design. Eligibility for the CalGrant A program required that the applicant's high school GPA, parental income and asset (excluding home equity) satisfy certain cutoff rules for each of these variables[31]. The number of awards each year was fixed, so the GPA cutoff (which ultimately depended on the number of eligible applicants) was unknown at the time of application. This makes high school GPA a particularly suitable assignment variable since it cannot be manipulated precisely, thus ensuring that students' GPAs near the threshold are as-good-as-random.

[30] In fact, the optimistic confidence interval occupies the middle ground between the most pessimistic and optimistic assumptions about $Cov(V_1, V_0)$. Strictly speaking, the most optimistic assumption is that V_1 and V_0 are perfectly positively correlated, but such heroic assumptions are not typically adopted for inference.

[31] There were actually two CalGrant programs (CalGrant A and B), but Kane focuses mainly on estimating the impact of CalGrant A on college going.

Since there are multiple assignment variables (GPA, income and assets) and two mutually exclusive treatment conditions (eligibility/ineligibility for the Cal-Grant program), this is a MDRD problem. Kane takes the "univariate" approach, estimating the treatment effects separately along each frontier. He focuses mostly on the GPA threshold, where he finds approximately a three percent increase in college going rate for applicants whose GPAs just pass the threshold, relative to those whose GPAs just miss in 1998. Kane also considers the income and asset thresholds, but for the most part, does not find statistically significant effects (possibly due to there being an insufficient number of observations near these thresholds, especially for the asset threshold).

5.2 Generating the Simulated Dataset

Unfortunately, the original data for Kane's paper is no longer available, so I create simulated dataset that is similar to the original one based on summary statistics and estimation results reported in the paper. Following the paper, I focus on estimating the impact that CalGrant A eligibility had on college going rates in 1998. My assignment variables are high school GPA and income, since the original paper hardly presented any results for the asset threshold.

First, I generate the assignment variables so that their distributions approximate those in the original paper. For high school GPA, I chose a scaled, shifted and truncated beta distribution which is similar to the empirical distribution in Kane's paper. The income threshold varies for families of different sizes, so I assume that the incomes for families of various sizes each follow a lognormal distribution with separate means, choosing the proportions of different family sizes so that the mean family size is close to that reported in the paper[32]. I then restrict the data to observations with GPA between 2.50 and 3.60 and exclude applicants who were eligible for both Cal Grants A and B, using this reduced sample for all subsequent analysis, following Kane's paper. The number of observations as well as the mean and median of assignment variables in the reduced sample are very similar to those reported for the original data. I then shift and rescale the variables so that they are centered at zero, with the non-negative quadrant of the assignment variable space, R_1 corresponding to eligibility for the program. Histograms for the simulated covariate values in the reduced sample are shown below.

Next, I select by trial-and-error, two data-generating processes (DGPs) to generate the outcome variable (college going). In particular, I made sure that the mean college going rate in the simulated data, as well as the regression results obtained by applying Kane's estimation approach for the GPA threshold to my data, are similar to those presented in the paper. I focus on replicating Kane's estimation results for the GPA threshold for two main reasons – first, Kane's regressions for the GPA and income thresholds are incomparable; second, his regression results for the GPA threshold are much stronger. To elaborate, Kane

[32] Clementi and Gallegati (2005) document that the empirical income distribution for the bottom 97 to 99 percent of the population (which encompasses the population of interest for Kane's paper) is consistent with lognormal distribution.

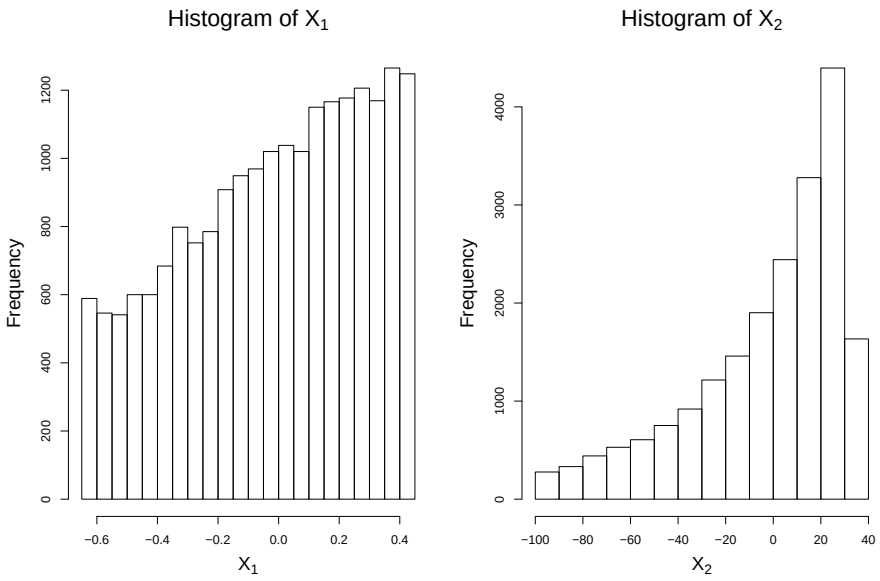

Figure 8: Histograms of the simulated assignment variables after restricting the data following Kane's approach.

estimates regressions for the GPA threshold with and without other covariates, using the subsamples of observations in 1998 and 1999 separately. For the income threshold, Kane instead estimates a regression using the pooled sample of observations from both years, with the inclusion of covariates. Yet, even using a pooled sample and including other covariates to increase precision, Kane's estimate for the income threshold is insignificant at the 5 percent level and has a higher p-value than the estimate at the GPA threshold using the 1998 sample alone without the inclusion of other covariates. Moreover, there is insufficient information about the covariates included in his income threshold regression to simulate these variables realistically. Hence, it likely would have been a fruitless task to attempt a reasonable replication of Kane's estimates on the income threshold.

Next, I describe the two DGPs I chose. Denoting the indicator for college going by

$$Y \equiv \mathbb{I}[\text{Entered College in the Next Year}],$$

the general form for the DGP is

$$\begin{aligned} Pr(Y = 1|X_1, X_2) = {} & \alpha_0 + \alpha_1 X_1 + \alpha_2 X_1^2 + \alpha_3 X_1^3 + \alpha_4 X_2 + \alpha_5 X_2^2 + \alpha_6 X_1^3 + \\ & \alpha_7 X_1 X_2 + \alpha_8 X_1^2 X_2 + \alpha_9 X_1 X_2^2 + \alpha_{10} sin(X_1) + \alpha_{11} sin(X_2) + \\ & W_i \cdot (g(X_1) + h(X_2)), \end{aligned}$$

where

$$X_1 \equiv (\text{High School GPA}) - (1998 \text{ GPA Cutoff}),$$
$$X_2 \equiv -\frac{1}{1000} \cdot (\text{Family Income} - \text{Income Cutoff}), \text{ and}$$
$$\alpha_0 = 0.82, \;\; \alpha_1 = 0.1, \;\; \alpha_2 = -0.01, \;\; \alpha_3 = -0.01, \;\; \alpha_4 = -2 \cdot 10^{-4},$$
$$\alpha_5 = -10^{-6}, \;\; \alpha_6 = -10^{-7}, \;\; \alpha_7 = 10^{-6}, \;\; \alpha_8 = -10^{-6}, \;\; \alpha_9 = 10^{-6},$$
$$\alpha_{10} = 0.01, \;\; \alpha_{11} = 0.01.$$

For the first DGP, I assume constant treatment effect along both frontiers, setting

$$g(X_1) = h(X_2) = 0.025 \text{ for all } X_1 \text{ and } X_2,$$

whereas for the second process, non-constant treatment effects are assumed, with

$$g(X_1) = 0.15 * \left[\Phi\left(\frac{X_1}{0.25}\right) - 0.5\right] \text{ and } h(X_2) = 0.06 * \left[\Phi\left(\frac{X_2}{15}\right) - 0.5\right],$$

where Φ denotes the cumulative distribution function of a standard normal random variable.

Using these probabilities, I generate the outcome variable Y corresponding to the simulated assignment variables. As with the assignment variables, the mean value of simulated outcomes is similar to the mean in the original data.

For estimation at the GPA threshold, Kane uses a probit regression with a cubic polynomial in high school GPA and a dummy variable indicating whether GPA passes the threshold. The DGPs were chosen so that running the same regression on the simulated datasets produces results similar to Kane's. The following regression table shows estimates for the marginal impact of crossing the GPA threshold on probability of college going. The point estimates and standard errors from the original paper are similar to those obtained using the simulated datasets, and the estimated treatment effect is significant at the 10 percent level in all three columns[33].

The next plots display the DGPs I chose for the constant and non-constant treatment effects respectively. Two features of these DGPs would seem to favor estimation via local linear regression over estimation via TPRS (although it turns out that TPRS still outperforms probit and local linear estimates on these simulated datasets).

First, the plots clearly show that the DGPs exhibit sinusoidal behavior, especially along the X_2 direction (which corresponds to income)[34]. These periodic

[33] Kane's paper does not include other statistics from the regression other than sample size.

[34] In fact, the DGP also includes periodic behavior in the X_1 direction (which corresponds to GPA). However, due to the scaling of the assignment variables, only the periodic behavior in the income direction is obvious. This is because the period of an oscillation in both the X_1 and X_2 direction is 2π, whereas the ranges of X_1 and X_2 in the plots are from -0.4 to 0.4 and -30 to 30 respectively.

Table 1: Probit Regression for GPA Threshold

	Dependent variable:		
	College Going		
	Kane's Data	Simulated Data (Constant Treat)	Simulated Data (Non-Constant Treat)
$\mathbb{I}[X_1 \geq 0]$	0.029*	0.033*	0.030*
	(0.015)	(0.017)	(0.016)
Observations	11,750	11,750	11,750
Order of Polynomial in GPA	3	3	3
Note:	*p<0.1; **p<0.05; ***p<0.01		

terms were included in the DGP to test the robustness of the TPRS treatment effect estimate. In particular, the true treatment effect along the GPA threshold – being the difference between the discontinuous surfaces along F_1 – is not periodic because the sinusoidal behavior at the boundaries of the two surfaces cancel out. However, the TPRS surfaces are unlikely to fit this periodic behavior perfectly, so that the TPRS estimate – being the difference between the two discontinuous TPRS surfaces – may still exhibit some periodic behavior which is not canceled out in the difference[35]. By contrast, the true DGP hardly exhibits any periodic behavior along the income threshold, so one may expect the TPRS estimate to perform better along this threshold.

Second, periodic terms aside, the DGP is a cubic polynomial in X_1 and X_2, with the quadratic, cubic and interaction coefficients largely dominated by the linear coefficients on X_1 and X_2. Hence, one would expect that local linear estimate to perform well given that the DGP is "close to linear".

5.3 Graphical Analysis with the Simulated Data

In this subsection, I demonstrate several graphical analysis methods described in section 3 using the simulated dataset with non-constant treatment effects. In particular, I show the outcome discontinuity plots for the GPA and income thresholds created using the "slicing" approach with a linear regression line fit to observations on each side of the cutoff, as well as the corresponding "sliding window" plots for the two cutoffs. In the "sliding window" graphs, I overlay the true DGP for comparison (using the true probabilities of college going arbitrarily close to each cutoff). Of course, lines representing the DGP would be absent in "sliding window" plots for a real dataset (where the DGP is unknown).

[35]This periodic behavior has a less pronounced effect on the probit and local linear estimates, since these methods are less flexible than the TPRS approach.

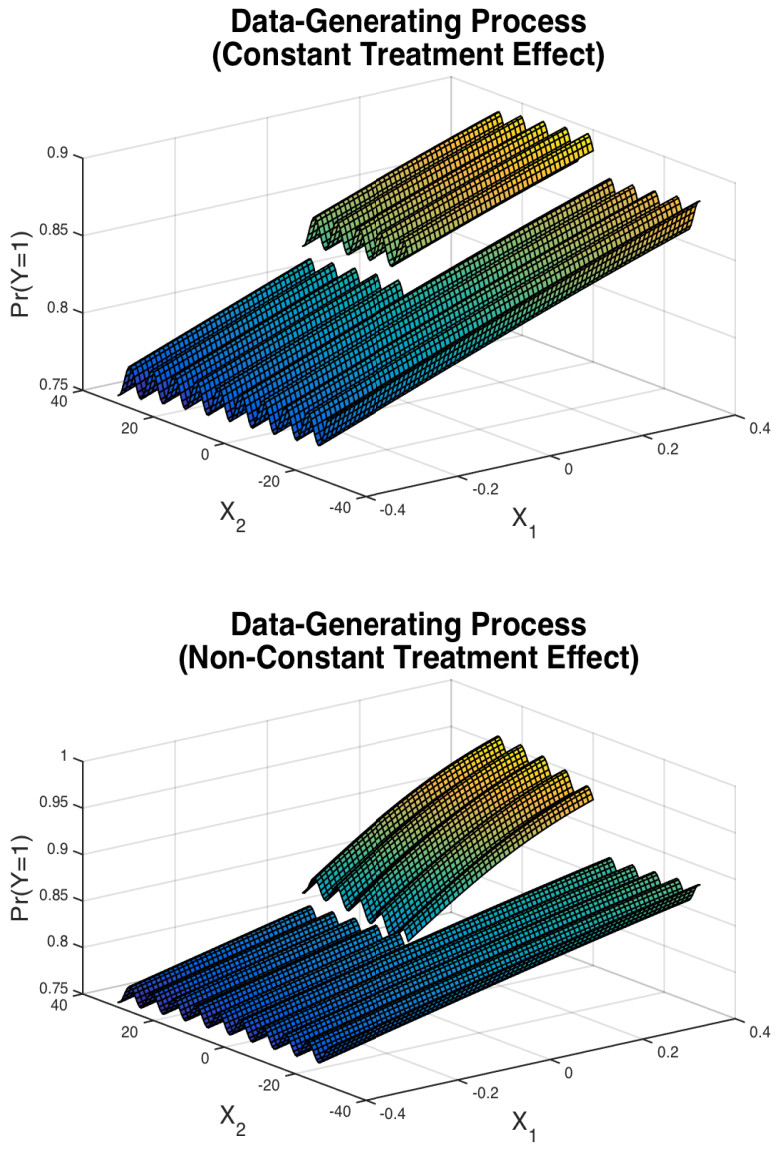

Figure 9: Data-generating processes for the simulated datasets.

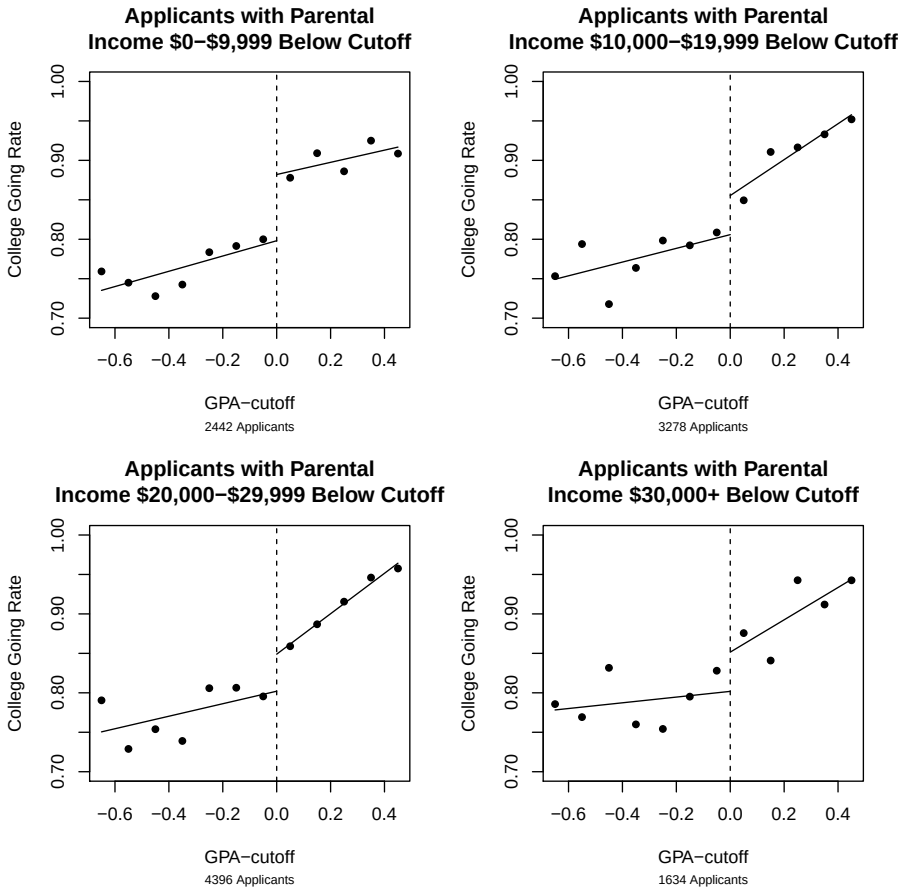

Figure 10: These "slicing" method plots show the discontinuity in college-going rate at the GPA threshold for applicants with different levels of parental income. These graphs were created using the simulated dataset with non-constant treatment effect.

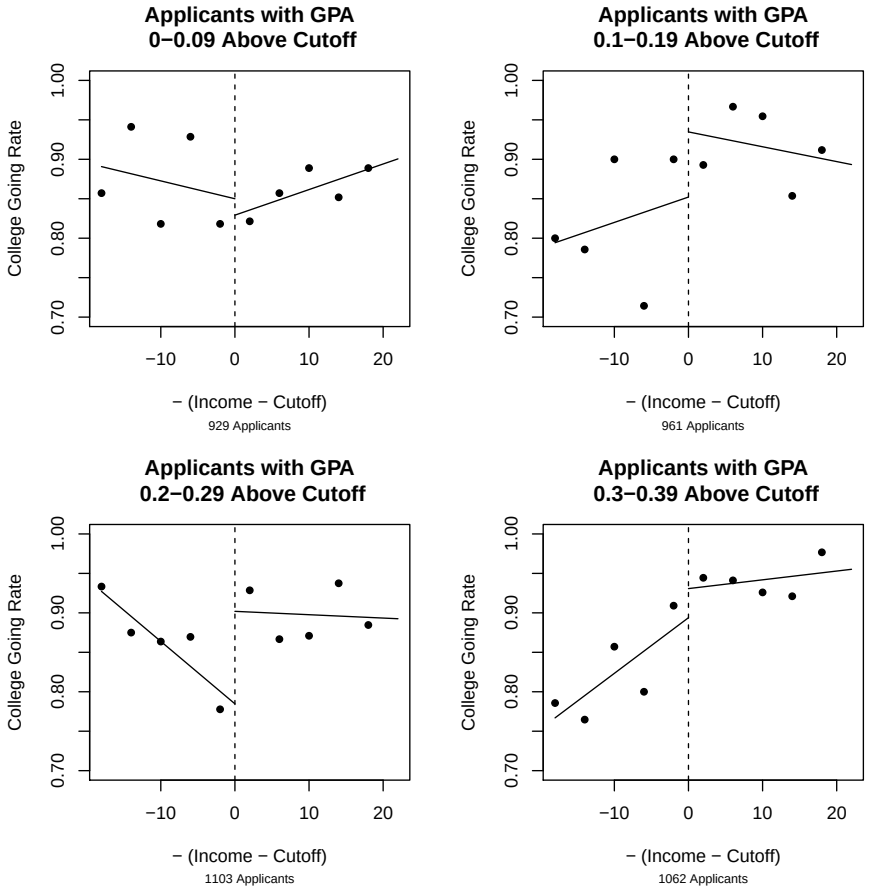

Figure 11: These "slicing" method plots show the discontinuity in college-going rate at the income threshold for applicants with different GPAs These graphs were created using the simulated dataset with non-constant treatment effect.

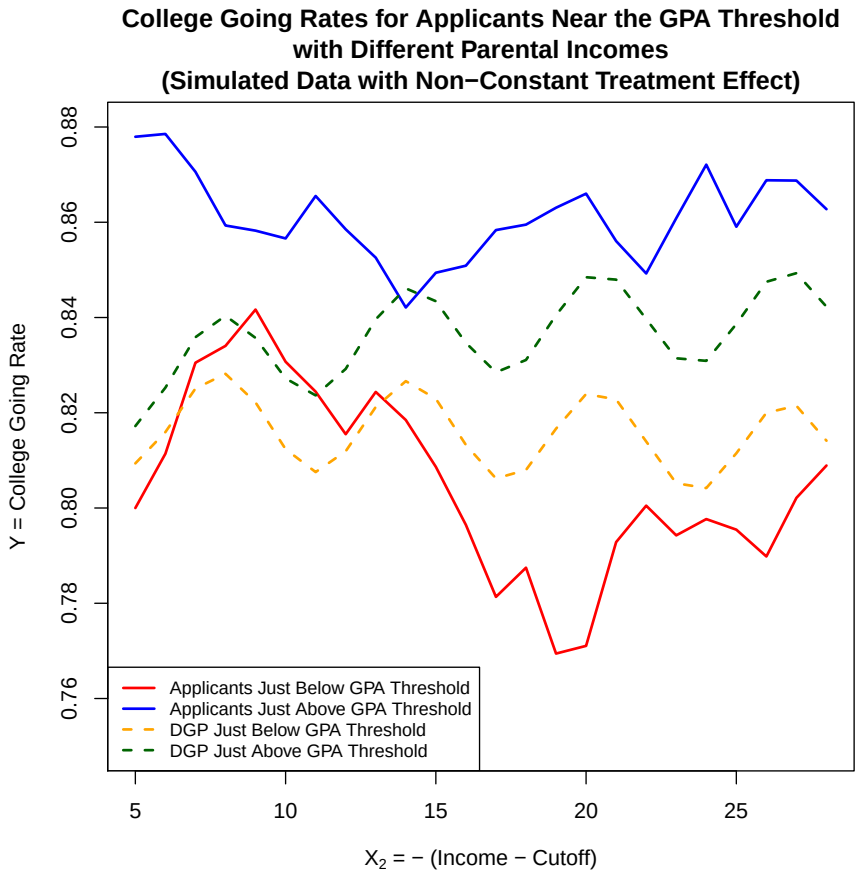

Figure 12: This "sliding window" plot shows the college-going rate for applicants with different levels of parental income who are just below and just above the GPA threshold. The solid lines represent the actual "sliding window" plot, and the dashed lines are the DGP for the simulated data, overlaid for comparison. Lines representing the DGP would be absent in "sliding window" plots for a real dataset (where the DGP is unknown). This graph was created using the simulated dataset with non-constant treatment effect.

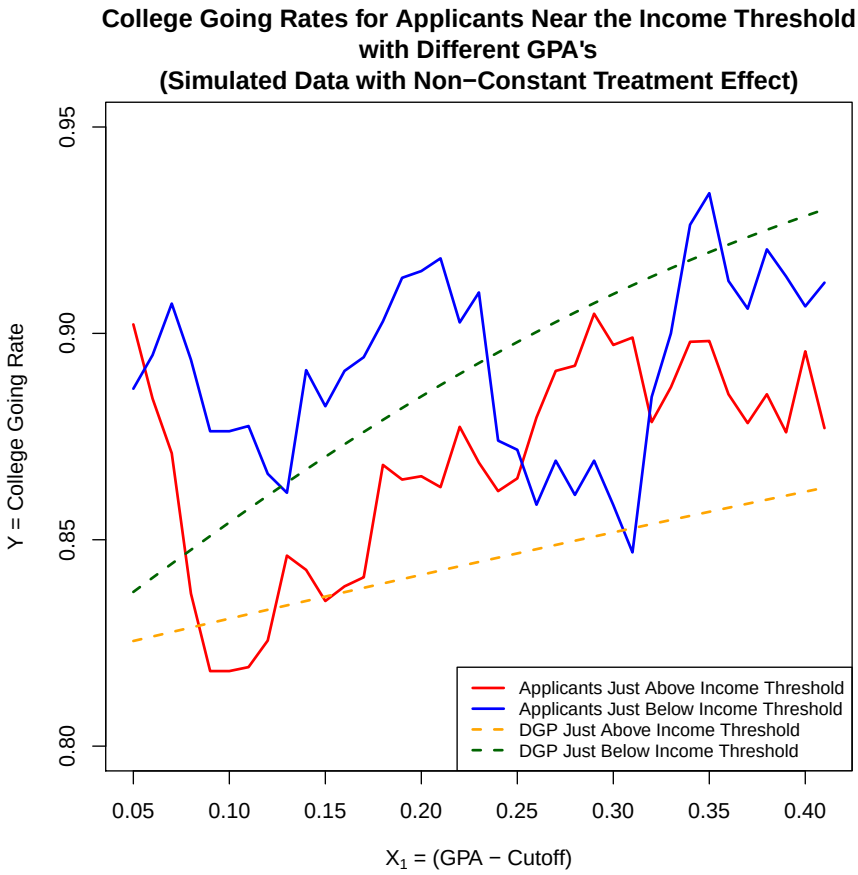

Figure 13: This "sliding window" plot shows the college-going rate for applicants with different GPAs who are just below and just above the income threshold. The solid lines represent the actual "sliding window" plot, and the dashed lines are the DGP for the simulated data, overlaid for comparison. Lines representing the DGP would be absent in "sliding window" plots for a real dataset (where the DGP is unknown). This graph was created using the simulated dataset with non-constant treatment effect.

The best fit lines for most of the "slicing" plots show a sizable discontinuity at the threshold, which provides visual evidence of a non-zero treatment effect for both frontiers. However, these plots fail to capture the fact that on both frontiers, the true treatment effect functions are increasing and concave.

By contrast, the "sliding window" plot for the GPA threshold provides some visual evidence that the treatment effect function $\tau_1(x_2)$ is increasing (even if the gap between the solid lines is larger than the gap between the dashed lines representing the DGP, i.e. the "graphical estimate" is biased upwards). However, the same plot for the income threshold fails to capture this pattern.

Therefore, although the "slicing" and "sliding window" plots provide some visual evidence on the discontinuity in outcome and of a non-constant treatment effect, they are insufficient for making inferences about the shape and magnitude of the treatment effect functions. A better understanding of the treatment effect along the two frontiers can only be gained through formal estimation, which is the topic of the remainder of this section.

5.4 Main Estimation Results

I estimate MDRD treatment effect functions for the two simulated datasets using Kane's probit approach, local linear regression and TPRS. For the local linear estimate, I conduct a grid search for an optimal bandwidth over 100 possible values of (h_1, h_2). Specifically, I allow h_1 and h_2 to range from 0.05 to 0.5 in increments of 0.05, and 5 to 50 in increments of 5 respectively. Several different values of δ (0.05, 0.10, 0.25, 0.50 and 1) are explored, although most of the subsequent discussion focuses on the results for $\delta = 0.05$.

Ideally, the estimated treatment function $\hat{\tau}_1(x_2)$ along the GPA frontier F_1 should have low bias and approximately follow the shape of the true treatment function, which is flat for the DGP with constant treatment, and increasing but concave for the DGP with non-constant treatment.

The performances of local linear regression and TPRS on the data with non-constant treatment is of particular interest, since the main reason for preferring estimation of treatment effect functions over scalar treatment effects is precisely to capture such non-constant treatment effects. Moreover, the heterogeneous treatment effect in this example has practical significance. Since X_2 is defined by how far *below* the income threshold an applicant's parental income is, financial constraints tend to represent a greater impediment to college going for applicants with greater values of X_2. Hence, it is reasonable to expect the CalGrant's impact to be increasing in X_2. On the other hand, the opportunity cost of lost wages may be more salient for very poor families (who have large values of X_2), which partially offsets the positive impact on college going, leading to an increasing but concave treatment effect function.

First, I show the surfaces fitted using local linear regression and TPRS for the two simulated datasets, which may be compared to the surfaces representing the true DGPs shown earlier.

Next, in order to compare the performances of different estimation approaches, I plot the local linear, TPRS, and probit treatment function estimates

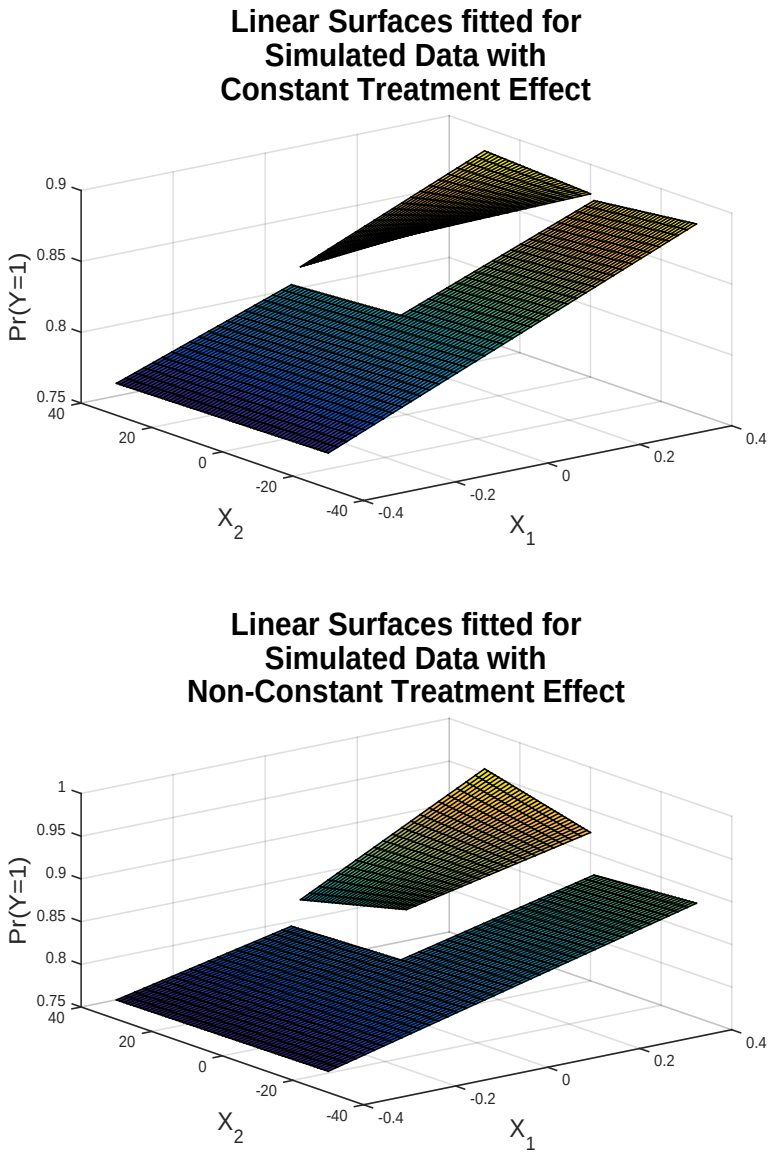

Figure 14: Linear surfaces fitted for the simulated datasets using local linear regression.

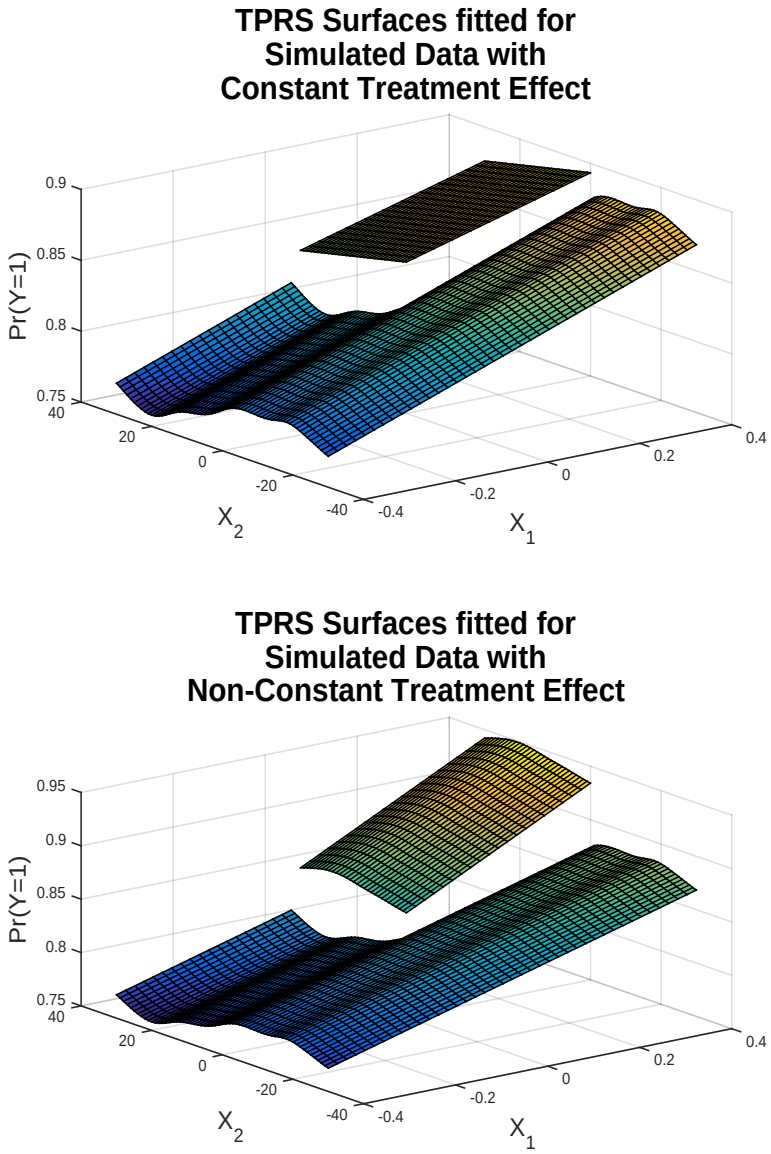

Figure 15: TPRS surfaces fitted for the simulated datasets.

for the GPA threshold on the same plots. Conservative 95 percent Bayesian confidence intervals (CIs) for the TPRS estimates, as well as the true treatment effects are also shown.

The plots for the estimated treatment functions along the GPA boundary F_1 show that the TPRS estimate works reasonably well in practice. Its performance on the GPA threshold is comparable to the probit regression, and far superior to the local linear estimate. This is despite the TPRS estimates being affected (for the worse) by the periodic behavior of the true DGPs near the GPA threshold (as evidenced by the plots showing the fitted TPRS surfaces).

In the plot for constant treatment effect, although the TPRS estimate is slightly wiggly, it does not exhibit an obvious increasing or decreasing trend, and its bias is similar to Kane's probit regression approach. For the plot with an increasing and concave treatment function, the TPRS estimate is also increasing and concave over most of F_1. While the TPRS estimate is far from perfect (for instance, it is biased upwards and its shape does not follow that of the true function exactly), its 95 percent CIs contain the true treatment function in both plots. Moreover, for the data with non-constant treatment effect, the TPRS CI includes zero for small values of X_2 and excludes zero for most larger values of X_2. This is in line with the true treatment function, which starts at zero and increases with X_2. The precision of the TPRS estimates on the GPA threshold is similar to that of the probit estimate – significant at the 10 percent level but not always at the 5 percent level.

By contrast, the local linear regression tends to exhibit higher bias in both plots, and its shape is totally uninformative about the shape of the true treatment function. In particular, as the linear surfaces and the treatment effect plots show, the local linear estimate is decreasing in X_2 for both datasets, even though the true treatment functions are constant for the first dataset and increasing for the second.

Next, I consider estimation of the treatment effect along the income frontier. As mentioned earlier, Kane's income threshold regressions are incomparable to his GPA threshold regressions. Nonetheless, I estimate probit regressions for the income threshold on the simulated datasets in the way that most closely follows his approach (i.e. with the same specification, except without other covariates, and using the 1998 subsample). The regression output and plots of the MDRD treatment function estimates are shown below.[36]

[36]Strictly speaking, the blue line in the plot should not be constant in X_1 because Kane's probit regression at the income threshold includes linear and quadratic terms in X_1. The constant blue line represents the marginal effect of crossing the income threshold on the probability of entering college when all the other covariates (including X_1 and X_1^2) are fixed at their mean. Hence, the blue line in a "correct" plot should be a non-constant function of X_1. Such a plot is shown in the appendix, where it is clear from the output that the varying treatment effect predicted using the probit model is still relatively constant, and in fact, performs slightly worse. Since the goal is to compare the TPRS (and local linear) estimate to the best benchmark of the probit regression, none of the results in this paper will be overstated by presenting a treatment effect for Kane's probit model that is better than it actually is. Hence, for simplicity sake (since standard errors for the marginal probit effect

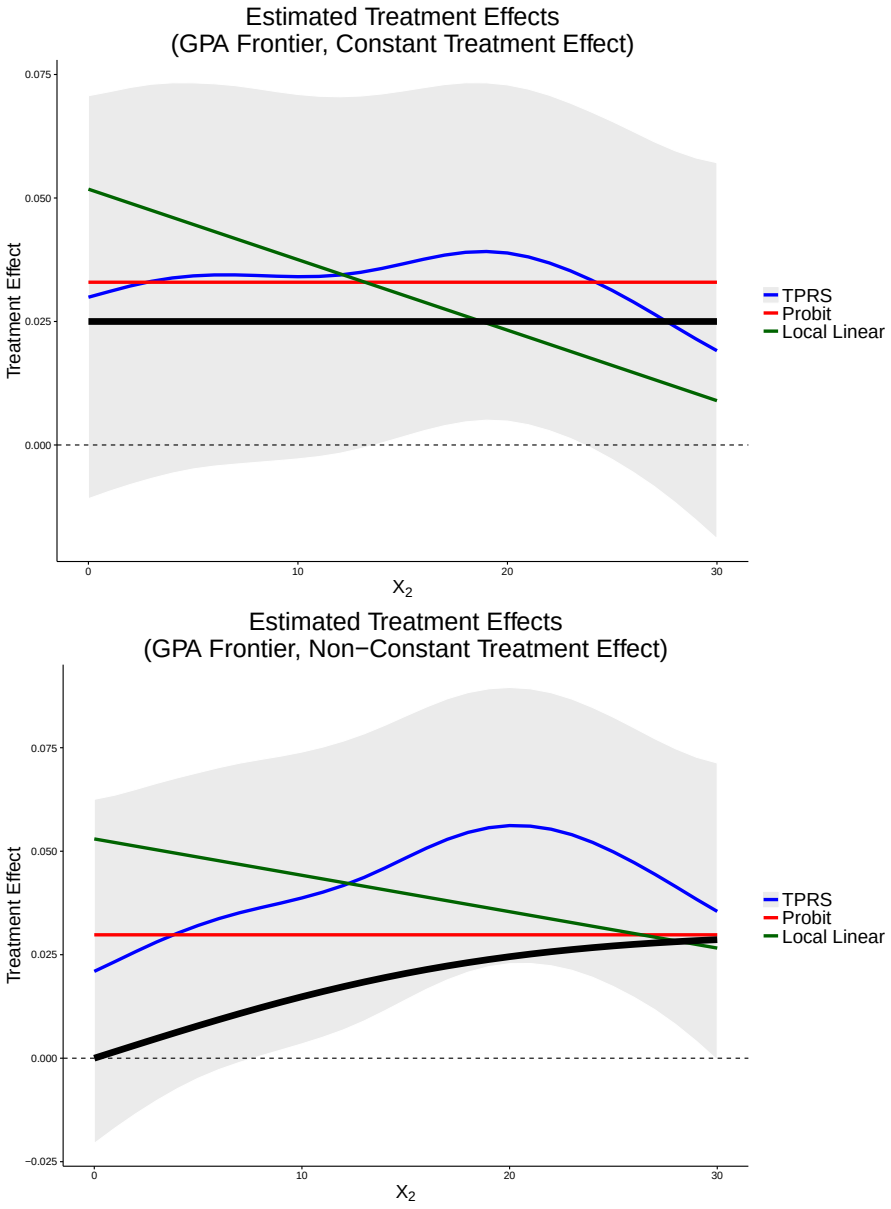

Figure 16: The thick black line in these plots represents the true treatment effect at the GPA threshold for the simulated data. The blue line is the MDRD estimate using TPRS, the red line is the probit estimate using Kane's approach and the dark green line is the local linear estimate with bandwidth selection via LM CV using $\delta = 0.05$. The shaded region represents the 95 percent conservative Bayesian confidence intervals for the TPRS point estimates of the treatment function.

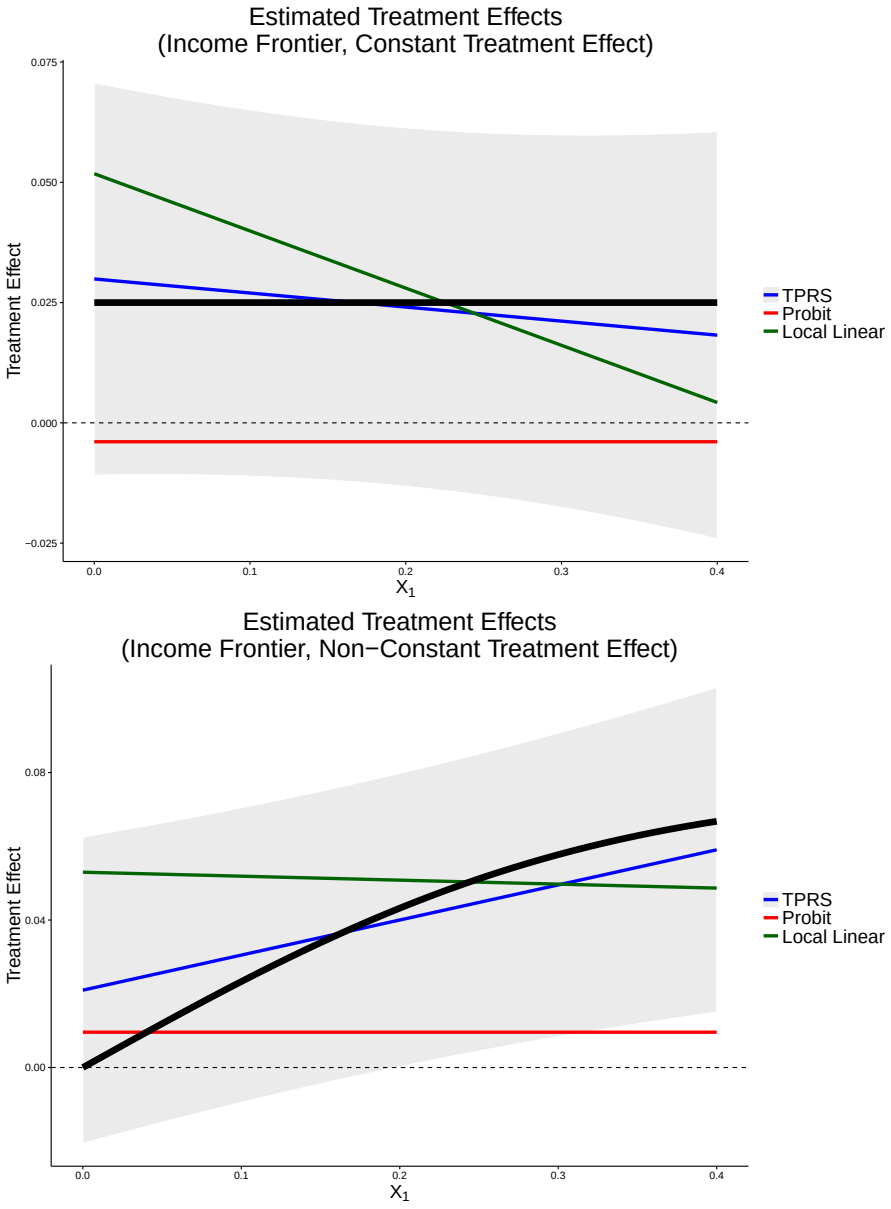

Figure 17: The thick black line in these plots represents the true treatment effect at the income threshold for the simulated data. The blue line is the MDRD estimate using TPRS, the red line is the probit estimate using Kane's approach and the dark green line is the local linear estimate with bandwidth selection via LM CV using $\delta = 0.05$. The shaded region represents the 95 percent conservative Bayesian confidence intervals for the TPRS point estimates of the treatment function.

Table 2: Probit Regression for Income Threshold

	Dependent variable:		
	College Going		
	Kane's Data	Simulated Data (Constant Treat)	Simulated Data (Non-Constant Treat)
$\mathbb{I}[X_2 \geq 0]$	0.031*	−0.004	0.010
	(0.018)	(0.027)	(0.024)
Observations	8,410	4,760	4,760
Order of Polynomial in Income	4	4	4
Order of Polynomial in GPA	2	2	2
Other Covariates Included	Yes	No	No
Year Used for Sample	1998 and 1999	1998	1998
Note:	*p<0.1; **p<0.05; ***p<0.01		

In the absence of notable periodic behavior in the DGP near the income threshold, the TPRS estimate significantly outperforms the probit and local linear estimates on the income threshold for both simulated datasets. The TPRS estimates approximate the shapes of both treatment functions well and exhibit low bias. As before, the 95 percent CI for the TPRS estimate for the non-constant treatment data includes zero at lower values of X_1 and excludes zero for larger values, which is in line with the increasing nature of the true treatment function.

By contrast, the probit estimate is badly biased – the estimate has the opposite sign in the constant treatment case, and is close to zero (with a p-value of 0.69) for the other simulated dataset.

The local linear estimate, while not as biased as the probit regression, still has a higher bias than the TPRS estimate. Furthermore, the plots once again show that the local linear estimates (which slope downwards for both datasets) are completely uninformative of the shape of the true treatment function.

5.5 Performance of Local Linear Regression Estimates

The previous subsection made clear that local linear regression estimates on the simulated datasets tended to be biased and failed to capture the shape of the

in probability at arbitrary covariate values are not easy to calculate), the rest of this paper treats the probit estimate at the income threshold as being constant in X_1.

true treatment function. This subsection investigates the causes of this poor performance.

As mentioned earlier in this paper, the primary issue with estimation via local linear regression is the selection of an appropriate bandwidth. The approach developed earlier in this paper for bandwidth selection requires a choice of δ which controls the amount of the data (close to the treatment frontiers) that is to be used for bandwidth selection. However, it was noted that there is no straightforward and objective way to determine a suitable value of δ. The following tables show the optimal bandwidths chosen when different values of δ are used.

Table 3: Optimal Bandwidth Choice for Different Values of δ (Simulated Data with Constant Treatment Effect)

δ	h_1	h_2
0.05	0.25	50
0.10	0.45	50
0.25	0.45	50
0.50	0.10	50
1.00	0.50	50

Table 4: Optimal Bandwidth Choice for Different Values of δ (Simulated Data with Non-Constant Treatment Effect)

δ	h_1	h_2
0.05	0.25	50
0.10	0.45	50
0.25	0.45	50
0.50	0.10	50
1.00	0.50	50

These tables show that the bandwidth choice varies with the value of δ. In particular, while the bandwidth chosen for X_2 stays constant at 50, the bandwidth chosen for X_1 varies from 0.1 to 0.5. The histograms for the simulated assignment variables presented earlier put these numbers in context – X_1 ranges from -0.65 to 0.45 (with first and third quartiles of -0.25 and 0.25 respectively), while X_2 ranges from -100 to 33 (with first and third quartiles of -20.2 and 22.6 respectively). Hence, the bandwidths chosen for X_1 under different δ's vary from a relatively small bandwidth to an extremely large one. On the other hand, the bandwidth chosen for X_2 covers most of its range, which may be reasonable considering that the DGP is "approximately linear".

This exercise also revealed that CV estimates of the loss function (MSE) are

relatively flat, which was the basis for a criticism leveled against CV methods by LM (2005), mentioned in the previous section. As an example, for the simulated dataset with non-constant treatment effect, the MSE estimates (for the 100 bandwidths considered) from my modified CV procedure with $\delta = 0.05$ had a minimum of 0.1341, with first, second and third quartiles of 0.1381, 0.1441 and 0.1670 respectively.

Still, this may not be a serious issue if these different bandwidths result in similar estimates. To determine whether this is the case, regression tables for the local linear regressions with bandwidths selected using different values of δ, as well as plots of the estimated treatment effect functions are shown below.

The regression tables and plots show that the estimated treatment effect functions are fairly similar for bandwidths chosen using different values of δ. This would have been a good sign, if not for the fact that all of these estimates perform poorly when compared to the true treatment effects underlying the DGPs, especially for the data with non-constant treatment effect.

In particular, the estimated coefficients for the slopes of the estimated treatment effect functions ($\hat{\beta}_5$ and $\hat{\beta}_6$) for the dataset with non-constant treatment effect are all negative (although they are statistically insignificant at the 10 percent level), in direct contradiction to the true treatment effect, which has a positive slope. Furthermore, the true treatment effect at the origin in this dataset, i.e. $\tau_1(0)$ and $\tau_2(0)$, given by β_1, is zero, but the local linear estimates for β_1 are all positive and statistically significant at the 5 percent level.

Given that the bandwidth selection procedure considered a relatively comprehensive grid of potential bandwidth values, it is unlikely that the poor performance of the local linear regression method with bandwidth selection via modified CV was caused by poor implementation alone.

In light of the poor results for MDRD estimation via local linear regression, the rest of this paper focuses on the TPRS estimation method, which has superior performance and is far less computationally expensive[37].

5.6 Performance of TPRS Estimate when Assignment Variables have Discrete Support

A criticism of the IK bandwidth choice algorithm for local linear regression in conventional RD is that the algorithm's asymptotic properties are unknown when assignment variables have discrete support. This same argument was made against extending the IK bandwidth choice algorithm to MRD in Papay et al. (2011a). Yet, one may also question whether the performance of MRD estimation using TPRS deteriorates when assignment variables have discrete support. While this paper does not offer theoretical results on this question, I investigate the issue empirically using the simulated datasets.

In fact, the exercise so far already has an assignment variable with discrete support – GPA, which is reported in hundredths. Yet, there are many possible

[37] While it may take more than three hours to run the CV bandwidth selection procedure for local linear regression using $\delta = 1$, a TPRS estimate may be obtained within five seconds.

Table 5: Local Linear Regressions Resulting from Different Values of δ (Simulated Data with Constant Treatment Effect)

	Dependent variable:				
	College Going				
	($\delta = 0.05$)	($\delta = 0.10$)	($\delta = 0.25$)	($\delta = 0.50$)	($\delta = 1$)
β_0	0.824^{***}	0.826^{***}	0.826^{***}	0.819^{***}	0.824^{***}
	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
β_1	0.052^{**}	0.050^{**}	0.050^{**}	0.057^{***}	0.052^{**}
	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)
β_2	0.127^{***}	0.144^{***}	0.144^{***}	0.168^{***}	0.138^{***}
	(0.028)	(0.019)	(0.019)	(0.039)	(0.018)
β_3	-0.0002	-0.0003	-0.0003	-0.0001	-0.0003
	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0002)
β_4	-0.0003	0.001	0.001	0.0004	0.0005
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
β_5	-0.119	-0.137	-0.137	-0.160^{*}	-0.130
	(0.084)	(0.084)	(0.084)	(0.086)	(0.084)
β_6	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
β_7	0.005	0.003	0.003	0.004	0.004
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Obs.	13,581	16,091	16,091	11,282	16,619
R^2	0.007	0.012	0.012	0.005	0.012

Notes: *p<0.1; **p<0.05; ***p<0.01

Treatment effect estimates along the GPA and income frontiers are respectively given by $\hat{\beta}_1 + \hat{\beta}_6 X_2$ *and* $\hat{\beta}_1 + \hat{\beta}_5 X_1$. *The treatment coefficients implied by the DGP with constant treatment effect are* $\beta_1 = 0.025$, $\beta_5 = 0$ *and* $\beta_6 = 0$.

Table 6: Local Linear Regressions Resulting from Different Values of δ (Simulated Data with Non-Constant Treatment Effect)

	Dependent variable:				
	College Going				
	($\delta = 0.05$)	($\delta = 0.10$)	($\delta = 0.25$)	($\delta = 0.50$)	($\delta = 1$)
β_0	0.824^{***}	0.826^{***}	0.826^{***}	0.819^{***}	0.824^{***}
	(0.005)	(0.005)	(0.005)	(0.006)	(0.005)
β_1	0.053^{***}	0.051^{**}	0.051^{**}	0.058^{***}	0.053^{**}
	(0.020)	(0.021)	(0.021)	(0.020)	(0.021)
β_2	0.127^{***}	0.144^{***}	0.144^{***}	0.168^{***}	0.138^{***}
	(0.026)	(0.018)	(0.018)	(0.036)	(0.017)
β_3	-0.0002	-0.0003	-0.0003	-0.0001	-0.0003
	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0002)
β_4	-0.0003	0.001	0.001	0.0004	0.0005
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
β_5	-0.011	-0.028	-0.028	-0.052	-0.022
	(0.079)	(0.080)	(0.080)	(0.080)	(0.080)
β_6	-0.001	-0.001	-0.001	-0.001	-0.001
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
β_7	0.006	0.004	0.004	0.005	0.005
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Obs.	13,581	16,091	16,091	11,282	16,619
R^2	0.020	0.025	0.025	0.018	0.026

Notes: *p<0.1; **p<0.05; ***p<0.01

Treatment effect estimates along the GPA and income frontiers are respectively given by $\hat{\beta}_1 + \hat{\beta}_6 X_2$ and $\hat{\beta}_1 + \hat{\beta}_5 X_1$. The DGP with non-constant treatment effect implies $\beta_1 = 0$. The coefficients β_5 and β_6 are unable to capture the non-linear treatment effect perfectly. Nonetheless, specification error aside, one would expect β_5 and β_6 to be positive, given that the treatment effect functions on both treatment frontiers are increasing.

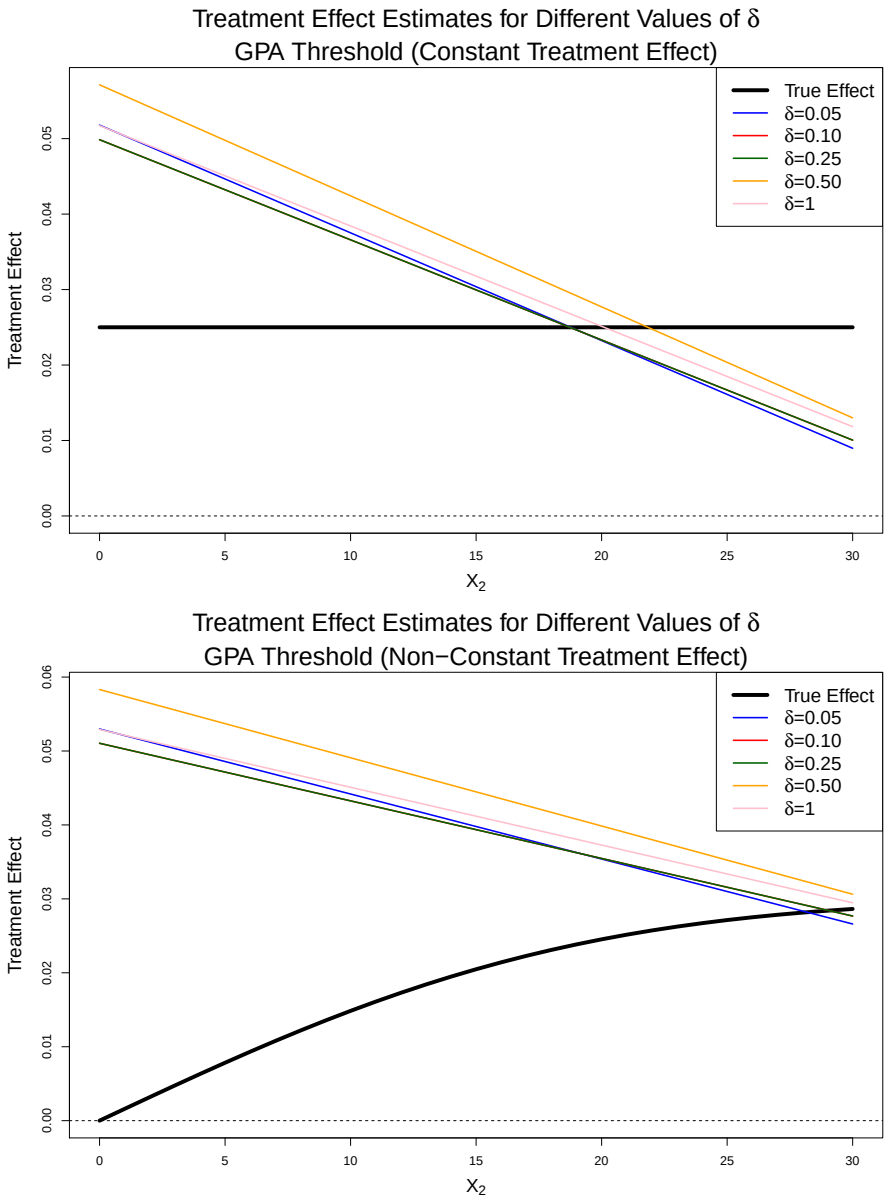

Figure 18: Local linear regression estimates of the MDRD treatment effect function at the GPA threshold using various values of δ. Note that the red line is hidden behind the dark green line due to identical bandwidth choice.

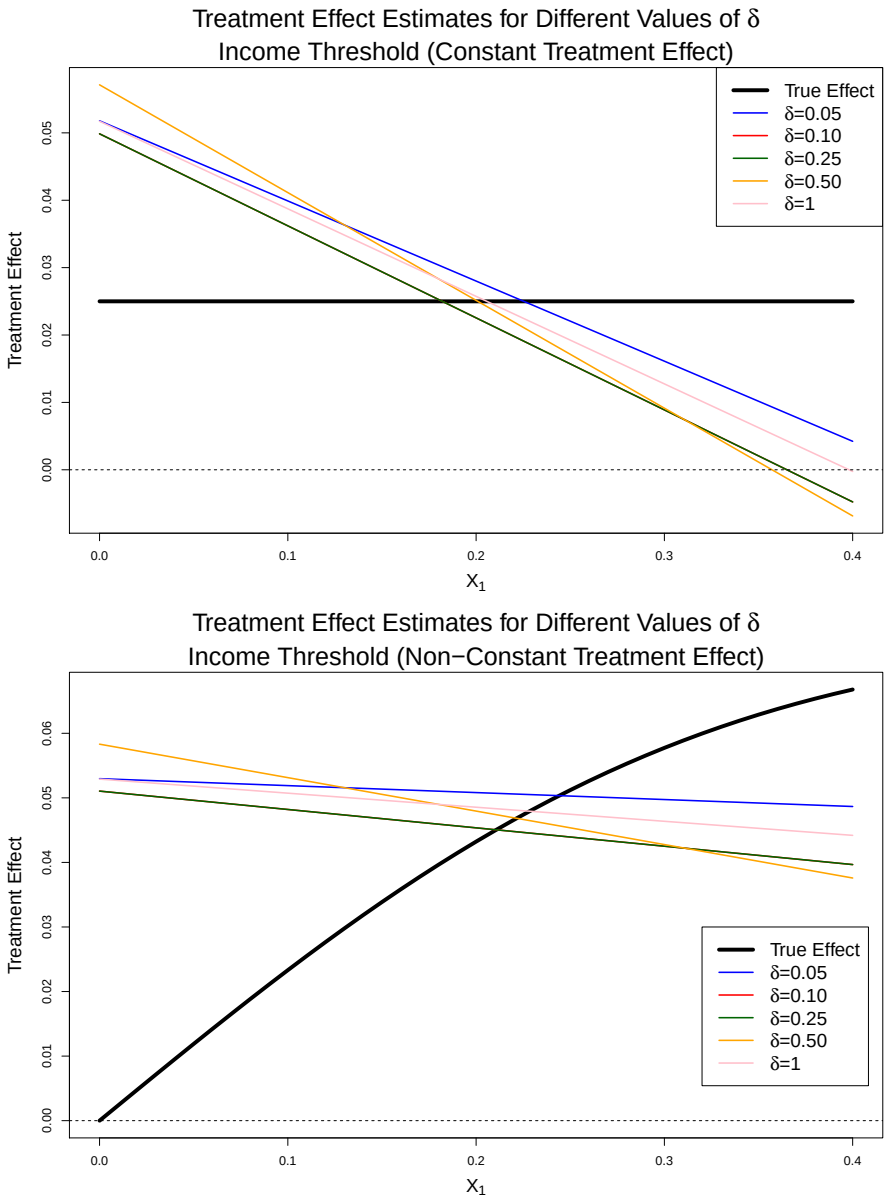

Figure 19: Local linear regression estimates of the MDRD treatment effect function at the income threshold using various values of δ. Note that the red line is hidden behind the dark green line due to identical bandwidth choice.

values for GPA in the sample used for analysis (111 in total). This large number of possible values makes GPA "similar in spirit" to a continuous variable. On the other hand, income – which the exercise has treated as a continuous variable so far – is often reported in round numbers, such as in $1,000's. In this subsection, I generate simulated datasets with income that is discretized to be in $1,000's, then estimate probit and TPRS treatment effects using this new data. The majority of observations take a relatively small number of (discretized) income values. For instance, Kane's regression for the income threshold uses only the subset of applicants with parental income within $20,000 of the threshold, which amounts to 41 possible income values.

Reassuringly, the figures below show that the estimated treatment effect functions using TPRS remain very similar even when income is discretized in $1,000's. The regression tables below show that the probit estimates are also hardly affected. In other words, all of the comments above about the performance of MDRD estimation using TPRS (relative to Kane's probit method) when income was a continuous variable still hold when income is discretized. Most importantly, this suggests that the performance of TPRS estimates is not adversely affected by assignment variables with discrete support.

Table 7: Probit Regressions using Simulated Data for GPA Threshold (with and without discretized income)

	Dependent variable:			
	College Going			
$\mathbb{I}[X_1 \geq 0]$	0.033*	0.033*	0.030*	0.030*
	(0.017)	(0.017)	(0.016)	(0.016)
Observations	11,750	11,750	11,750	11,750
Order of Polynomial in GPA	3	3	3	3
Discretized Income	No	Yes	No	Yes
Constant Treatment Effect	Yes	Yes	No	No
Note:	*p<0.1; **p<0.05; ***p<0.01			

5.7 Sensitivity of TPRS Estimates to Bandwidth Choice

One of the main advantages of using TPRS for estimation over local linear regressions is that one does not need to estimate an optimal bandwidth for TPRS. Instead, one can simply use most of the data for TPRS estimation (which is equivalent to choosing very large bandwidth), since data that is "too far" from the frontiers will not, in theory, significantly affect the treatment effect estimate due to TPRS being a local estimator. In this subsection, I reestimate the TPRS

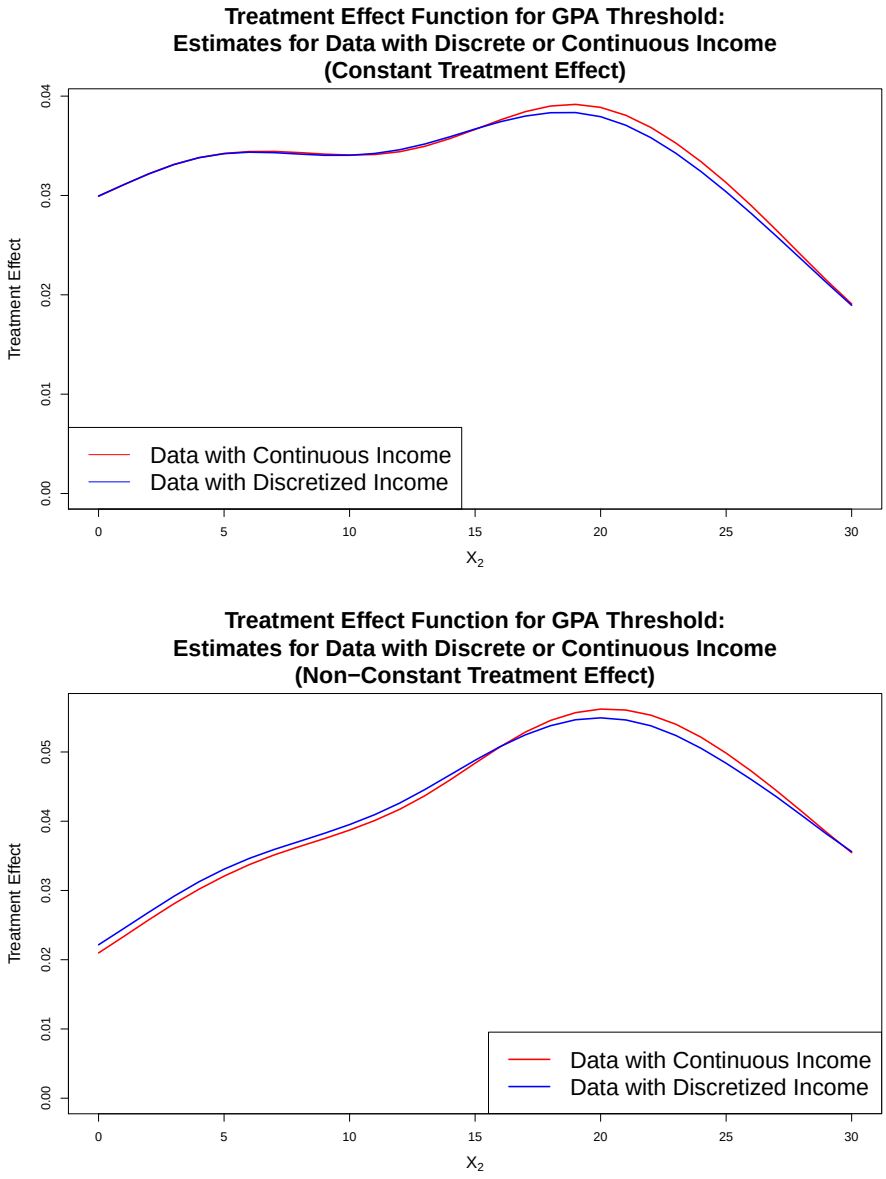

Figure 20: The blue and red lines represent the MDRD treatment effect estimates at the GPA threshold using TPRS for the dataset with discretized income, and with continuous income respectively.

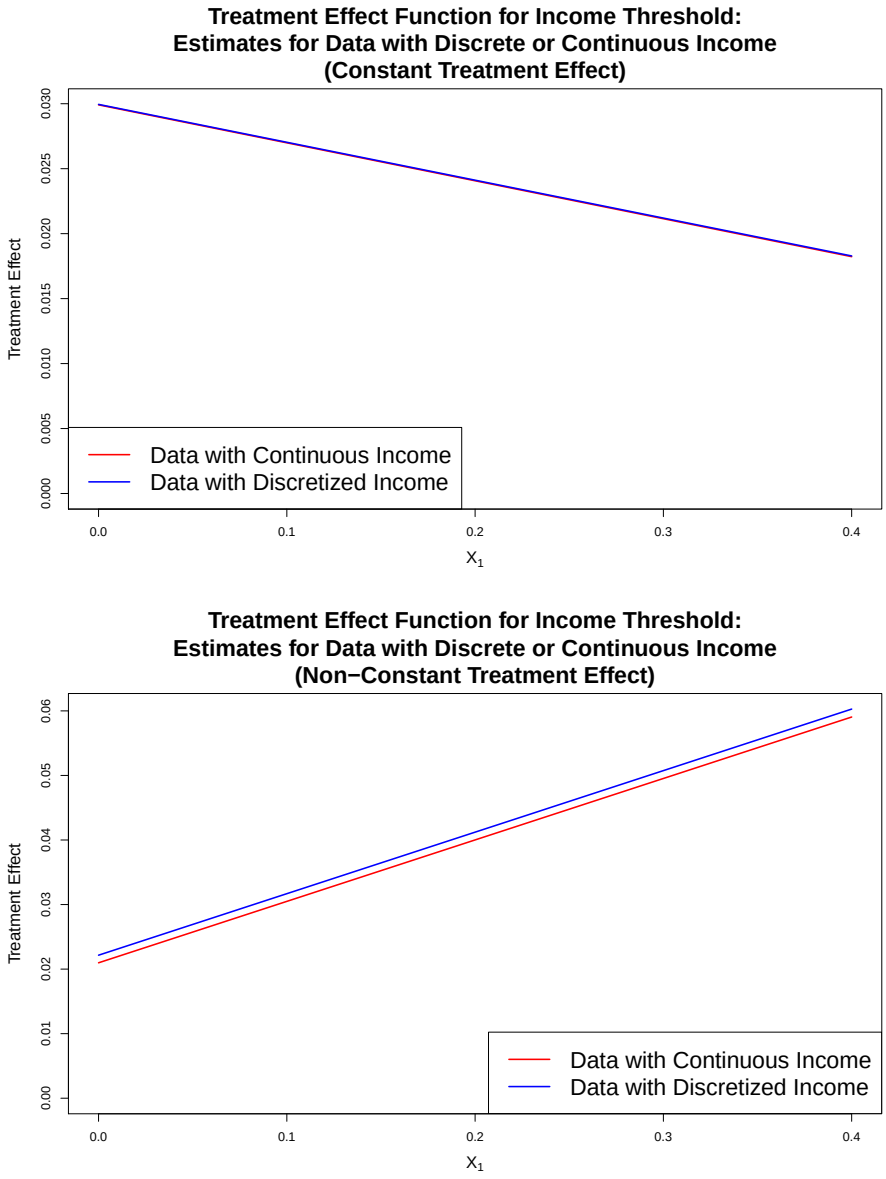

Figure 21: The blue and red lines represent the MDRD treatment effect estimates at the income threshold using TPRS for the dataset with discretized income, and with continuous income respectively.

Table 8: Probit Regressions using Simulated Data for Income Threshold (with and without discretized income)

	Dependent variable:			
	College Going			
$\mathbb{I}[X_2 \geq 0]$	−0.004	−0.005	0.010	0.009
	(0.027)	(0.027)	(0.024)	(0.024)
Observations	4,760	4,760	4,760	4,760
Discretized Income	No	Yes	No	Yes
Constant Treatment	Yes	Yes	No	No
Note:	*p<0.1; **p<0.05; ***p<0.01			

treatment functions using various relatively large (non-symmetric) bandwidths to check how sensitive the treatment effect estimates are to this choice.

The results presented in the rest of this paper uses all the data, which corresponds to "Bandwidth 1" in this subsection. As the histograms of the assignment variables in an earlier subsection show, the distribution of X_2 values has a long left tail that extends far below zero (corresponding to applicants with parental incomes too high for CalGrant eligibility), and the distribution of X_1 values also has a left tail (albeit a shorter one). Hence, I select the other bandwidths in this subsection by truncating the left tail of X_1 and/or X_2, so that the ranges of X_1 and/or X_2 values are centered at zero. Specifically, the chosen bandwidths are:

Bandwidth 1: 20,180 observations with $X_1 \in [-0.65, 0.45]$ and $X_2 \in [-100, 35]$;

Bandwidth 2: 16,805 observations with $X_1 \in [-0.65, 0.45]$ and $X_2 \in [-100, 35]$;

Bandwidth 3: 18,030 observations with $X_1 \in [-0.45, 0.45]$ and $X_2 \in [-35, 35]$;

Bandwidth 4: 15,027 observations with $X_1 \in [-0.45, 0.45]$ and $X_2 \in [-35, 35]$.

Encouragingly, the plots below show that the estimated treatment effects for these bandwidths are relatively similar, implying that bandwidth choice is of second order concern when using TPRS for estimation (as long as it is chosen to be large enough). This is in contrast to local linear regression, where bandwidth selection is a critical yet unresolved issue.

5.8 Results in Repeated Simulations

The simulation exercise so far used two simulated datasets generated with appropriate distributions for the covariates and suitable DGPs, to approximate the true dataset for Kane's paper. While the TPRS approach outperforms the

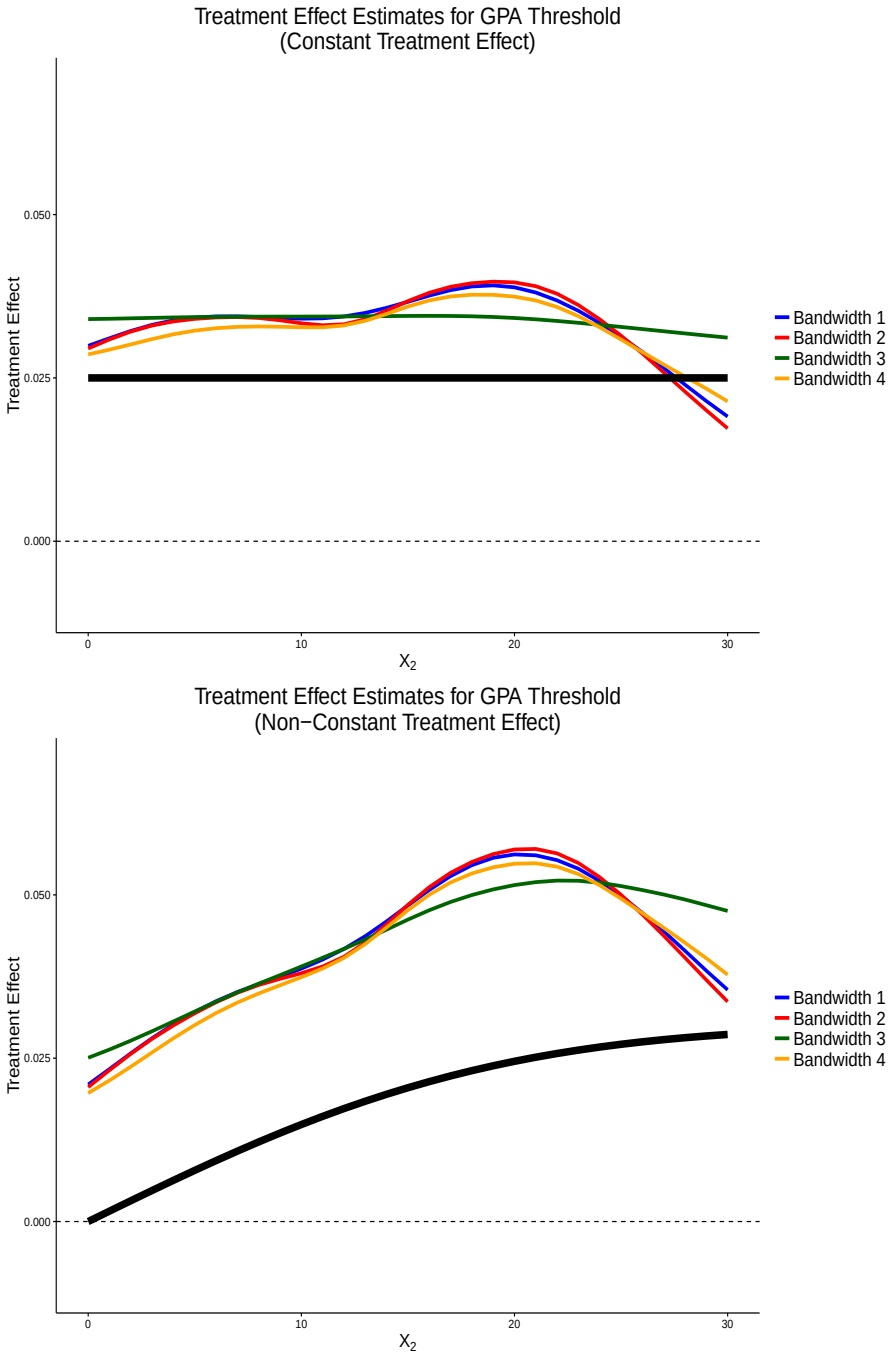

Figure 22: TPRS estimates of the treatment effect function at the GPA threshold are plotted for various choices of bandwidth. The true effect is shown in black.

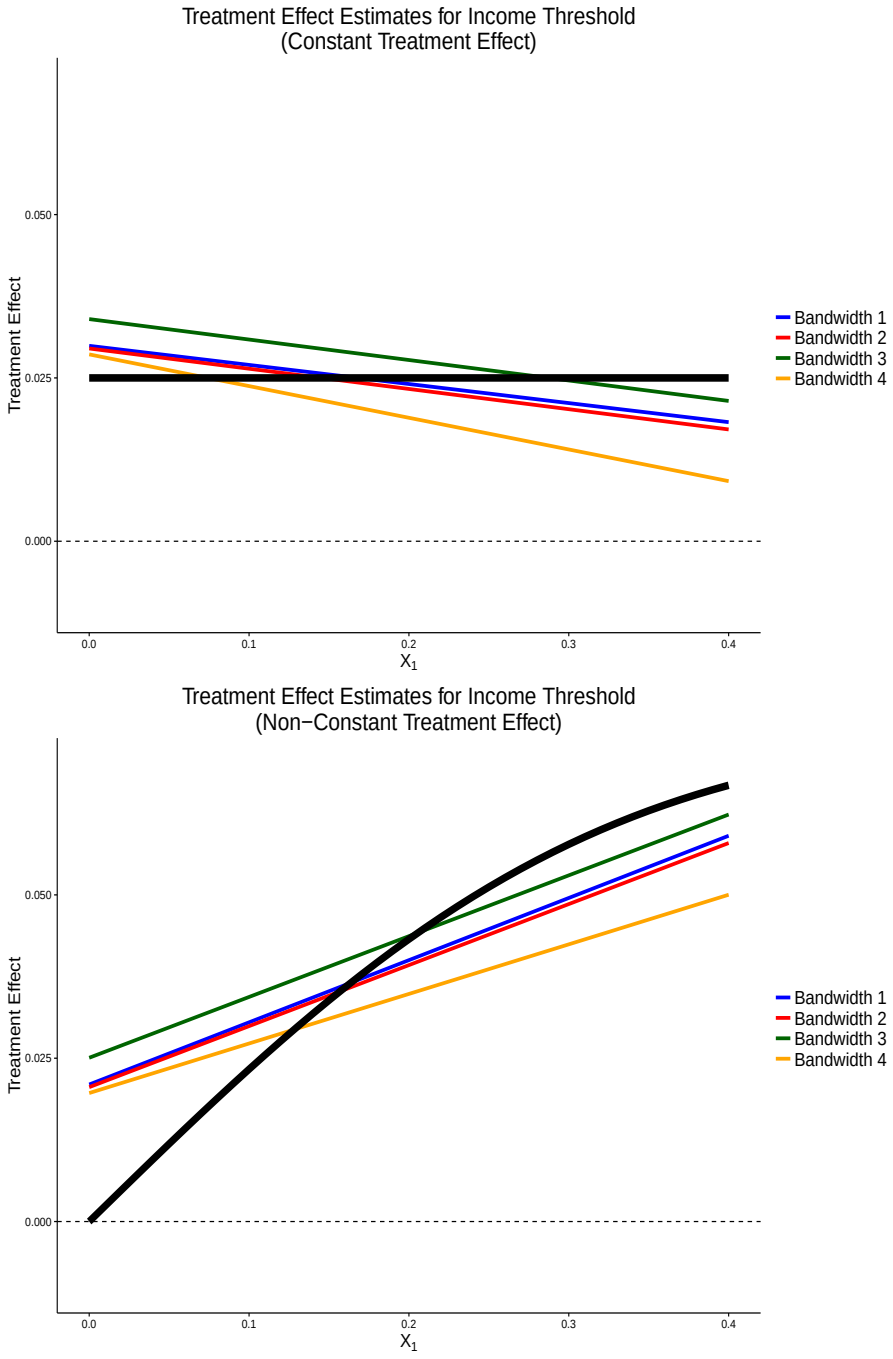

Figure 23: TPRS estimates of the treatment effect function at the income threshold are plotted for various choices of bandwidth. The true effect is shown in black.

probit regression on these simulated datasets, it is possible that the TPRS estimate just happened to perform well on these particular realizations of the data. In this subsection, I assess the performance of the TPRS estimation approach by repeatedly generating datasets (a total of 1,000 times) using the same covariate distributions and DGPs as before, and obtaining TPRS and probit estimates for each realization of the data.[38]

5.8.1 Mean Squared Integrated Error of the TPRS and Probit Estimates

A metric is needed to compare the TPRS and probit estimates of the treatment function. MSE is a common measure for comparing the accuracies of scalar estimators when the true parameter value is known. However, the estimate in this case is a function rather than a scalar. Hence, rather than reporting the MSEs for a grid of points, I report the Mean Integrated Squared Error (MISE), which provides a more convenient summary of the global accuracy of the function estimate. In this example, given the true treatment function τ_1 (or τ_2) and its estimate $\hat{\tau}_1$ (or $\hat{\tau}_2$), the MISE is defined by:[39]

$$(21) \quad MISE(\tau_k, \hat{\tau_k}) \equiv \int_{F_k} (\hat{\tau_k} - \tau_k)^2 dx_{k'}, \text{ where } k \cup k^{'} = \{1, 2\}.$$

The following table shows that on average the TPRS estimate far outperforms the probit estimate in terms of MISE. The four cases considered correspond to the estimates for the two different DGPs (constant/non-constant treatment effect) along the two frontiers (GPA and income). In all four cases, the MISE for the TPRS estimate is smaller than that for the probit estimate (in fact, the TPRS MISE is 2.4 to 3.3 times smaller along the income frontier).

5.8.2 Performance of the TPRS Bayesian Confidence Intervals

In addition to comparing the accuracies of the TPRS and probit estimates, the repeated simulations also allow me to assess the performance of the conservative Bayesian confidence intervals. One may recall from the previous section that provided the error term has a normal distribution, the coefficient estimates for the basis functions of the TPRS (and hence, the point estimates of the treatment effect) are normally distributed. However, this simulation exercise is based on

[38] I did not perform this exercise for the local linear estimate due to the computational expense of the bandwidth selection procedure. For instance, selecting a bandwidth using $\delta = 1$ takes more than three hours, which makes repeating the procedure 1,000 times on a single machine intractable.

[39] For computational reasons, I approximate this integral using the value of the integrand at the left endpoint for a grid of values along the frontier (31 values from 0 to 30 for F_1 and 41 values from 0 to 0.40 for F_2). While this integral approximation is rather crude, it would not be badly biased unless the integrand is systematically increasing or decreasing over the range of the frontier (which does not appear to be the case). Moreover, the focus here is on the difference in MISE between the TPRS and probit estimators, rather than the exact value of the MISE. It turns out that the difference between the MISE of these two estimators is large enough to alleviate such concerns.

Table 9: Comparison of MISE's for TPRS and Probit Estimates

Treatment	Frontier	TPRS MISE	Probit MISE	Ratio of MISE's (Probit/TPRS)
Constant	GPA	0.006	0.009	1.421
Non-Constant	GPA	0.009	0.010	1.106
Constant	Income	0.0001	0.0003	2.431
Non-Constant	Income	0.0001	0.0005	3.262

a probability model where the dichotomous outcome variable can only take the values of zero or one, which implies that the assumption of normally distributed error terms is violated. Hence, normality of the treatment point estimates is only an approximation based on large sample theory. The large number of TPRS estimates and their Bayesian CIs will shed light on whether this approximation is a valid one.

There are two questions to answer concerning the distributions of the point estimates. The first and arguably more important concern is whether the Bayesian 95 percent CIs cover the true function value at least 95 percent of the time. The second is whether the distributions of the point estimates are approximately normal.

The following plots show the coverage rates for the TPRS and probit CIs in the 1,000 simulations under the two DGPs (as before, the conservative Bayesian CIs are used for TPRS). For the simulated data with constant treatment effect, the TPRS CIs have very high coverage rates (about 98 to 99 percent) along the entire ranges of both frontiers. The TPRS CIs tend to have lower coverage rates near the origin for the simulated data with non-constant treatment effect, although moving away from the origin, the coverage rates are typically at least 95 percent. Regardless of the frontier or the dataset in question, the TPRS CIs tend to have higher coverage rates on average than the probit CIs.

To address the second concern, histograms of several point estimates are shown below. These empirical distributions seem approximately normal, in that they are unimodal and are not systematically skewed in either direction.

5.8.3 Shape of the TPRS Treatment Effect Estimate

The final issue explored in this subsection concerns whether the shape of the average TPRS function estimate is similar to that of the true treatment function. To this end, I plot the means of the TPRS point estimates along a grid of points on each frontier (for both the DGPs with constant and non-constant treatments). I also overlay the true treatment functions in these graphs for comparison.

The graphs below show that the TPRS estimates at the income threshold tend to approximate the shape of the true function quite well, while TPRS estimates at the GPA frontier show some sinusoidal behavior. This is an artifact

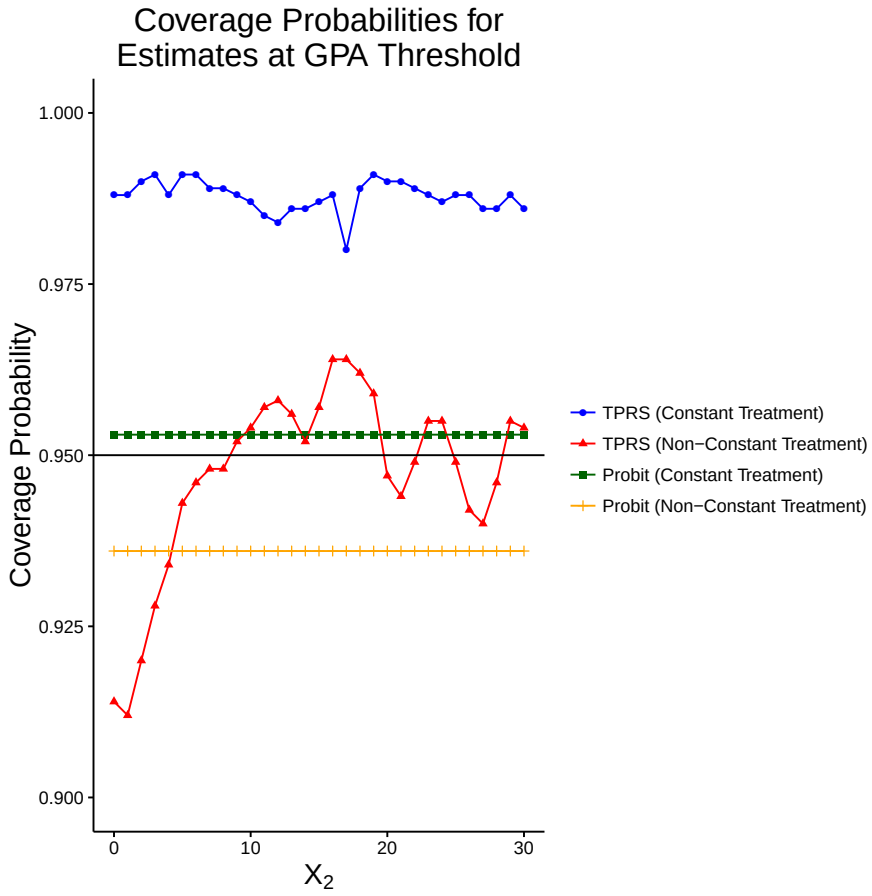

Figure 24: The various lines show the empirical coverage over 1,000 simulations of the two estimators using the two different simulated datasets for different values of X_2 along the GPA frontier. The black line shows the 95 percent coverage level.

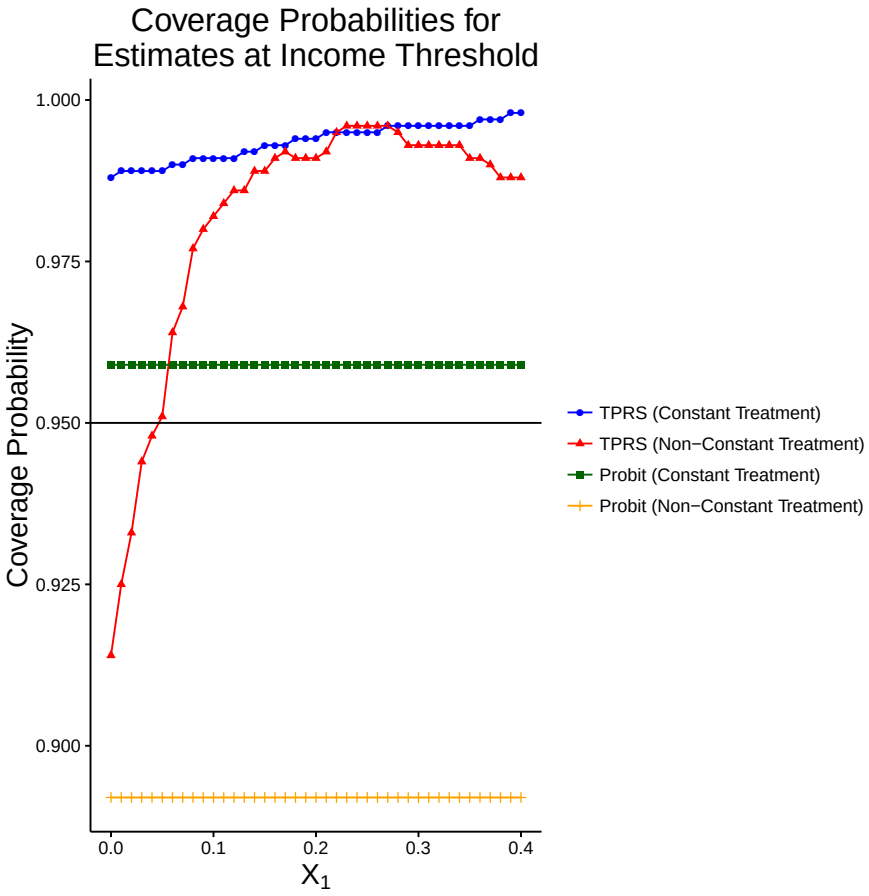

Figure 25: The various lines show the empirical coverage over 1,000 simulations of the two estimators using the two different simulated datasets for different values of X_1 along the income frontier. The black line shows the 95 percent coverage level.

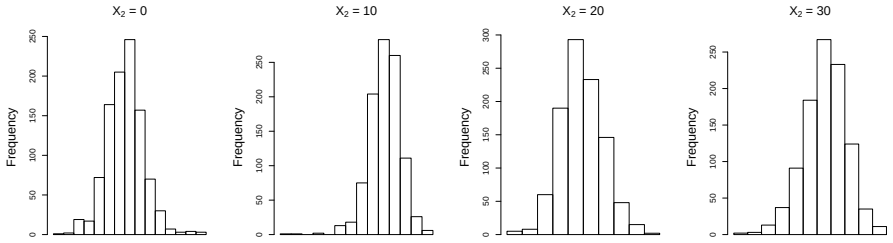

Figure 26: Distribution of TPRS point estimates along the GPA frontier for the simulated dataset with constant treatment effect.

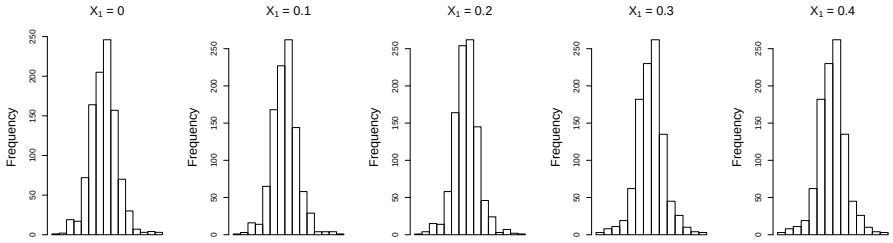

Figure 27: Distribution of TPRS point estimates along the income frontier for the simulated dataset with constant treatment effect.

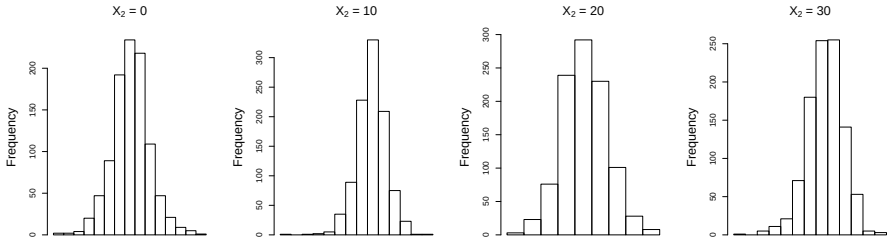

Figure 28: Distribution of TPRS point estimates along the GPA frontier for the simulated dataset with non-constant treatment effect.

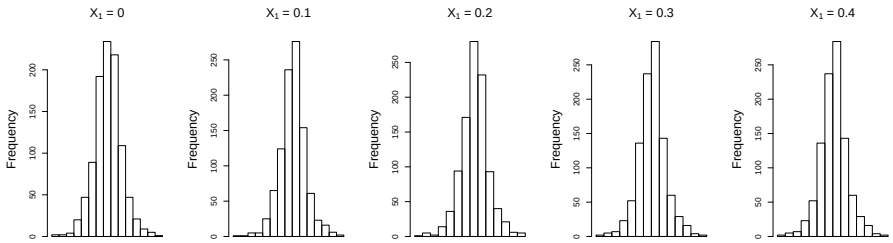

Figure 29: Distribution of TPRS point estimates along the income frontier for the simulated dataset with non-constant treatment effect.

of the periodic behavior in the true DGPs along the GPA frontier that is imperfectly captured by the TPRS estimates (as mentioned several times earlier). Nonetheless, if one abstracts from this oscillation about the trend, the general shape of the TPRS estimate at the GPA threshold is also similar to that of the true function.

6 Additional Considerations

6.1 Fuzzy MRD

Thus far, this paper has only considered sharp MRD, where treatment is completely determined by values of the assignment variables. Here, I discuss fuzzy MRD and describe how the treatment effect function may be estimated, starting with the case for dichotomous treatment. I abbreviate fuzzy MRD in general as FMRD, with the dichotomous and multiple treatment cases respectively abbreviated as FMDRD and FMMRD.

6.1.1 FMDRD

Similar to conventional fuzzy RD, FMRD refers to situations where the jump in probability of receiving treatment when crossing a threshold is less than one. However, instead of a scalar threshold as in conventional RD, there are several one-dimensional frontiers in FMRD. Complicating matters further, it is possible for the FMRD on one frontier to be fuzzy even when the FMRD on another frontier is sharp, or for the jump in treatment probability to be one on part of a frontier and less than one on the rest of the same frontier. The following two examples show how such situations may arise in reality.

To illustrate how the jump in probability may vary along a single frontier, consider a state financial aid package where eligibility depends, without exceptions, on high school students having GPA and family income respectively above and below their thresholds. However, not all eligible applicants apply. Assuming the sample is the entire population of high school students (rather

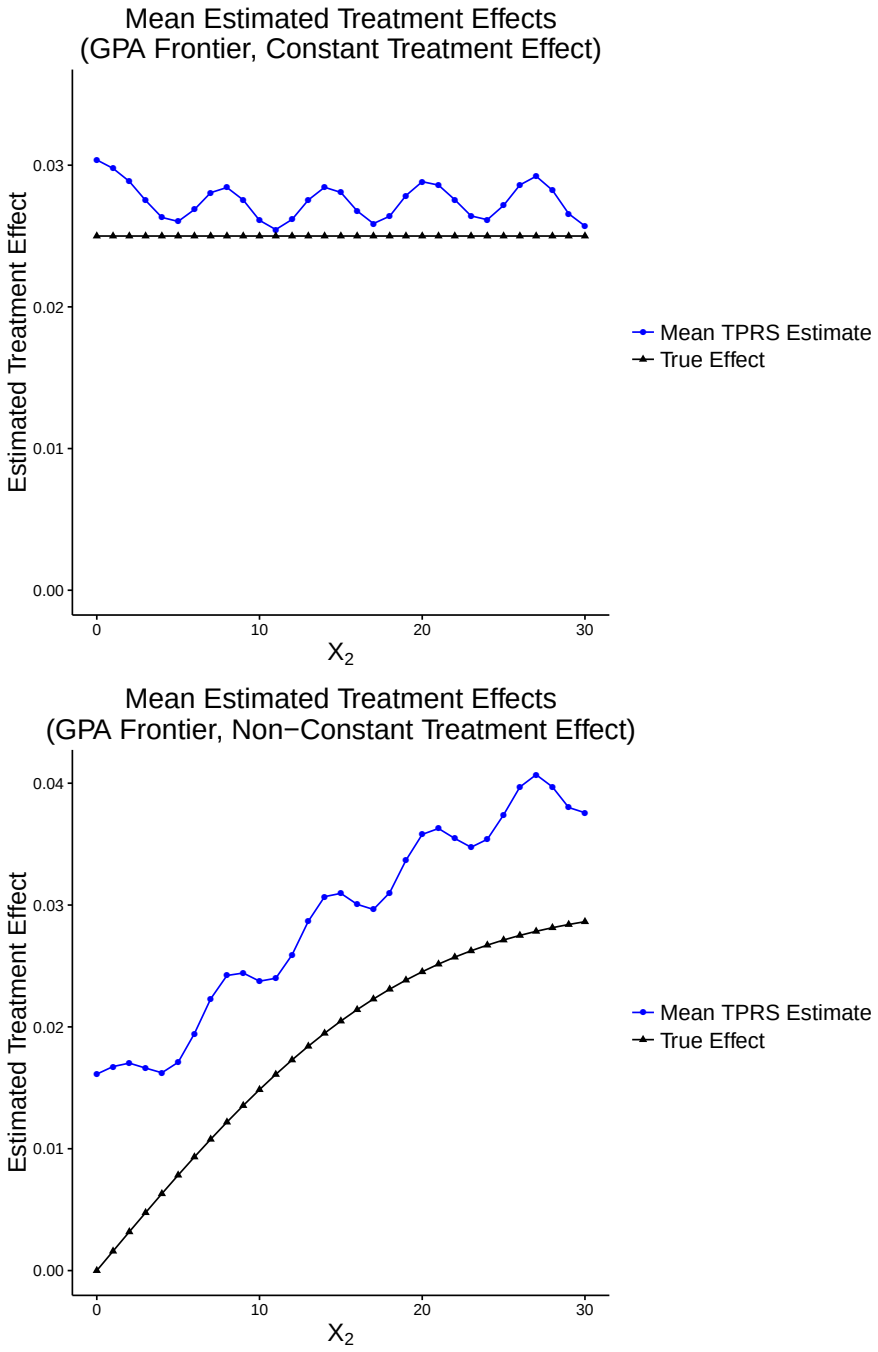

Figure 30: Mean TPRS estimates at the GPA threshold are shown in blue. The black line represents the true treatment effect.

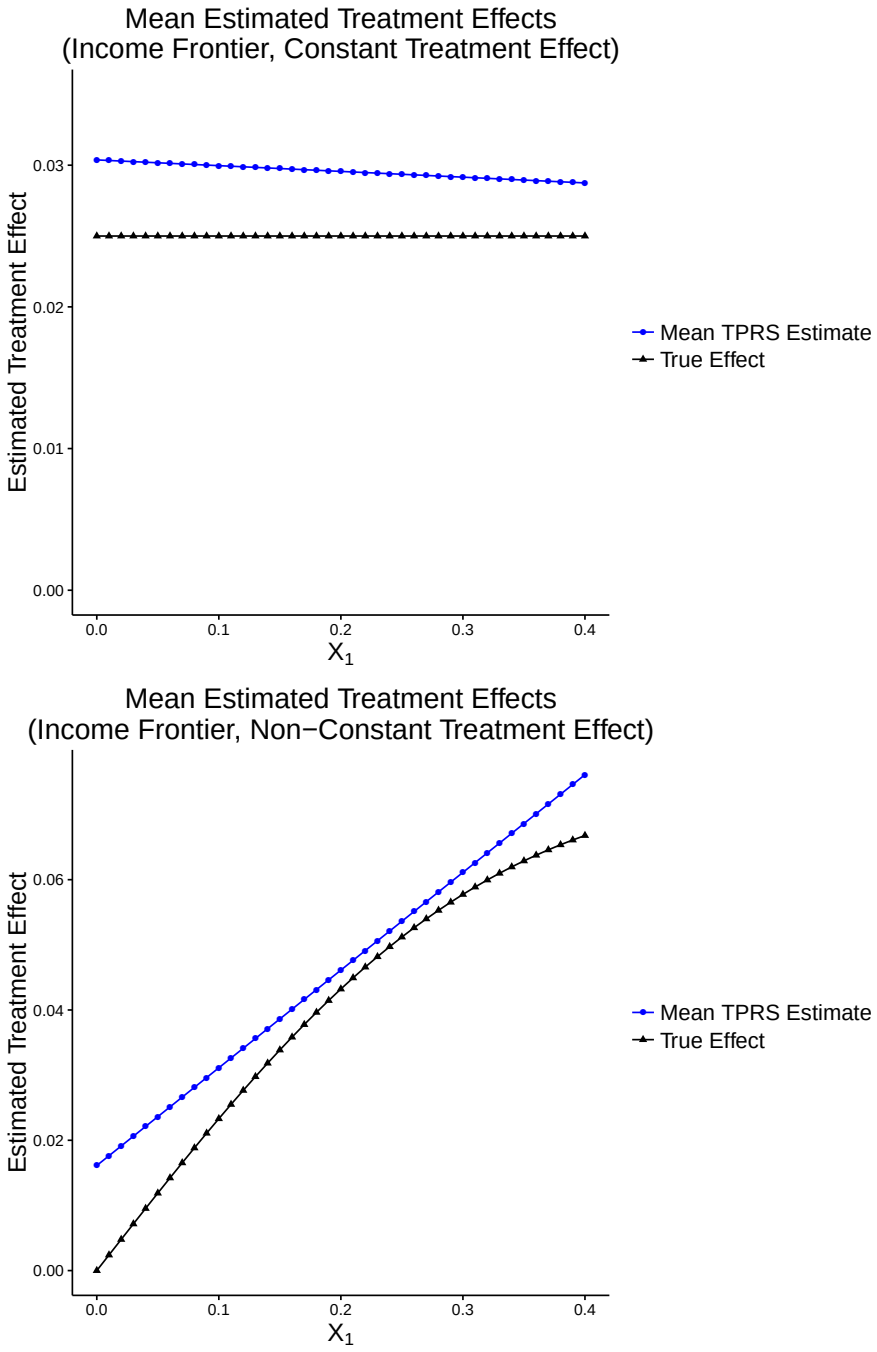

Figure 31: Mean TPRS estimates at the income threshold are shown in blue. The black line represents the true treatment effect.

than just the applicants), and using whether one receives the aid package as the treatment variable, this becomes a FMDRD problem. Now, the decision of eligible applicants to apply (or not) may vary depending on his/her family income. It is quite plausible that for students with family incomes far below the income threshold, the aid package is extremely attractive, so that the jump in probability at the GPA threshold is one for low values of income. On the other hand, students with family income just below the income threshold may be less influenced by the aid package, so the jump in probability at the GPA threshold is less than one for higher values of income (that are still below the income threshold).

This next example has the same basic setup, with eligibility for a financial aid package depending on GPA and family income. Now, suppose that the family income threshold is very low and that the aid package can only be used for public colleges. Assume further that all eligible students who barely pass the GPA threshold apply, since the aid package is extremely attractive for these students from very poor families (by definition of being below the income threshold). Hence, the jump in treatment probability at the GPA threshold is one. On the other hand, suppose that among the eligible students just below the income threshold, some of the students with high GPAs (far above the GPA threshold) do not apply because of scholarship offers they receive from private colleges. This constitutes a case where the MDRD on one frontier (i.e. the GPA threshold) is sharp, whereas the MDRD on another frontier (specifically, the income threshold) is fuzzy.

In practice, one can get a sense of whether the FMRD is fuzzy along a frontier by examining discontinuity plots created using the "slicing" approach (described in section 3.1) with probability of receiving treatment on the vertical axis.

Having discussed the basic concepts of FMDRD, I now define it more formally. The following definition pertains to a FMDRD for the frontier F_1 (the definition of FMDRD for F_2 is completely analogous). Assume that

$$\lim_{x_1 \to 0^+} \mathbb{E}[W_i | X_{1i} = x_1, X_{2i} = x_2] - \lim_{x_1 \to 0^-} \mathbb{E}[W_i | X_{1i} = x_1, X_{2i} = x_2] < 1$$

for some $x_2 \geq 0$, and

$$\lim_{x_1 \to 0^+} \mathbb{E}[W_i | X_{1i} = x_1, X_{2i} = x_2] \neq \lim_{x_1 \to 0^-} \mathbb{E}[W_i | X_{1i} = x_1, X_{2i} = x_2]$$

for all $x_2 \geq 0$.[40] Then, denoting

$$f(x_1, x_2) \equiv \mathbb{E}[Y | X_{1i} = x_1, X_{2i} = x_2],$$
$$p(x_1, x_2) \equiv \mathbb{E}[W | X_{1i} = x_1, X_{2i} = x_2],$$

[40] One may relax this assumption by requiring it only for a subset of F_1. However, should one choose to do so, care must be taken in choosing the domain of definition for the treatment function estimate to avoid division by zero.

the FMDRD treatment effect function along F_1 is defined as:

$$(22)\quad \tau_1^{Fuzzy}(x_2) \equiv \frac{\lim_{x_1 \to 0^+} f(x_1, x_2) - \lim_{x_1 \to 0^-} f(x_1, x_2)}{\lim_{x_1 \to 0^+} p(x_1, x_2) - \lim_{x_1 \to 0^-} p(x_1, x_2)} \quad \text{for } x_2 \geq 0.$$

TPRS estimates for the numerator and denominator can be obtained using the approach outlined in section 4.3 (with ultimate outcome Y and treatment status W as the outcome variables respectively).

To elaborate, I denote the TPRS fitted using observations in the non-negative quadrant R_1 with Y and W as the outcome variables respectively by $\hat{f}_1(x_1, x_2)$ and $\hat{p}_1(x_1, x_2)$, and those fitted on $\mathbb{R}^2 \setminus R_1$ by $\hat{f}_0(x_1, x_2)$ and $\hat{p}_0(x_1, x_2)$. This allows the point estimate of the treatment effect at any point along the frontier F_1 to be written as:

$$(23)\quad \hat{\tau}_1^{Fuzzy}(x_2) \equiv \frac{\hat{f}_1(0, x_2) - \hat{f}_0(0, x_2)}{\hat{p}_1(0, x_2) - \hat{p}_0(0, x_2)} \quad \text{for } x_2 \geq 0.$$

The point estimates for the numerator and denominator of the FMDRD treatment functions have posterior Bayesian distributions that are (approximately) normal, as explained at the end of section 4. However, it is unlikely that a useful theoretical approximation for the distribution of the FMDRD point estimate can be found without knowing the exact distributions of the numerator and denominator as well as their correlation. In particular, ratios of normal random variables need not be well behaved.[41]

Hence, rather than deriving theoretical CIs for the FMDRD estimate, a more practical approach may be to obtain CIs by bootstrapping. Unless the dataset is huge, the computational burden is unlikely to be excessive – computing a MDRD estimate for the simulated dataset in the previous section (which contains almost 20,000 observations) takes less than five seconds on a personal laptop. This flexibility is the main reason TPRS is preferred over thin plate splines, since bootstrapping with the latter would not be feasible even with moderately large datasets due to its computational expense.

If the FMDRD is fuzzy on both frontiers, one can simply use the formula above to obtain $\hat{\tau}_1^{Fuzzy}(x_2)$, and compute $\hat{\tau}_2^{Fuzzy}(x_1)$ similarly using the formula:

$$(24)\quad \hat{\tau}_2^{Fuzzy}(x_1) \equiv \frac{\hat{f}_1(x_1, 0) - \hat{f}_0(x_1, 0)}{\hat{p}_1(x_1, 0) - \hat{p}_0(x_1, 0)} \quad \text{for } x_1 \geq 0.$$

Point-wise CIs can then be obtained via bootstrap for both functions.

However, there is a choice to be made when the FMDRD is fuzzy on one frontier (say F_1) but sharp on the other (F_2 in this case). $\hat{\tau}_1^{Fuzzy}(x_2)$ should of course be estimated using the FMDRD formula as before, with point-wise CIs obtained by bootstrapping. On the other hand, one can either use $\hat{\tau}_2^{Fuzzy}(x_1)$

[41] For instance, the ratio of two independent standard normal random variables follows a Cauchy distribution, which has undefined mean and variance.

for estimation of the treatment effect function along F_2, or $\hat{\tau}_2^{Sharp}(x_1)$, which is defined by:

$$(25) \quad \hat{\tau}_2^{Sharp}(x_1) \equiv \hat{f}_1(x_1, 0) - \hat{f}_0(x_1, 0) \quad \text{for } x_1 \geq 0.$$

The advantages of using $\hat{\tau}_2^{Sharp}(x_1)$ are that this estimator has lower bias if the FMDRD is truly sharp on the frontier F_2, and Bayesian CIs can be used. However, it will be the case that in general, $\hat{\tau}_2^{Sharp}(0) \neq \hat{\tau}_1^{Fuzzy}(0)$ as long as $\hat{f}_1(0,0) - \hat{f}_0(0,0) \neq 0$ and $\hat{p}_1(0,0) - \hat{p}_0(0,0) \neq 1$. By contrast, using the fuzzy treatment effect estimates for both frontiers ensures that $\hat{\tau}_2^{Fuzzy}(0) = \hat{\tau}_1^{Fuzzy}(0)$. Moreover, obtaining bootstrapped CIs for $\hat{\tau}_2^{Fuzzy}(x_1)$ imposes almost no additional computational burden, since bootstrapping is already done for $\hat{\tau}_1^{Fuzzy}(x_2)$.[42]

6.1.2 FMMRD

Next, I show that with the exception of a slight complication, the procedure for estimating FMMRD is essentially the same. Suppose that there are four treatments conditions, roughly corresponding to each of the four quadrants of $\mathbb{R}^2$, and consider for this example, estimation of the treatment function $\tau_{12}(x_2)$ along F_{12}, the non-negative x_2-axis.

Most of the observations in R_k $(k = 1, 2, 3, 4)$ consists of observations with treatment k (i.e. satisfy $W_{ki} = 1$). However, there may be observations in $R_1 \cup R_2$ that satisfy neither $W_{1i} = 1$ nor $W_{2i} = 1$. For this example, suppose for concreteness that there are observations in R_2 that receive treatment 3.[43]

Estimation of $\tau_{12}(x_2)$, which compares treatment 1 to treatment 2, may be invalidated by observations that receive neither treatment. Therefore, only observations receiving treatments 1 or 2 should be used in estimating the treatment effect $\tau_{12}(x_2)$. After discarding observations in $R_1 \cup R_2$ with neither $W_{1i} = 1$ nor $W_{2i} = 1$, one may then use graphical analysis to decide whether to estimate $\tau_{12}^{Sharp}(x_2)$ or $\tau_{12}^{Fuzzy}(x_2)$. Also, notice that in this example, the presence of observations receiving treatment 3 in R_2 suggests that the MMRD along the frontier F_{23} may be a fuzzy one.

6.2 Inclusion of Baseline Covariates

In this subsection, the vector of assignment variables is denoted by $\boldsymbol{X}$ and the vector of baseline covariates by $\boldsymbol{Z}$. If the MRD assumptions are valid, one would expect the density of $\boldsymbol{Z}$ conditional on $\boldsymbol{X}$, $f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x})$, to be continuous at the

[42] To see this, note that during each bootstrap replication b to obtain the bootstrap estimate $\hat{\tau}_{1b}^{Fuzzy}(x_2)$ of $\tau_1(x_2)$, the functions $\hat{f}_{1b}$, $\hat{f}_{0b}$, $\hat{p}_{1b}$ and $\hat{p}_{0b}$ are fitted. Therefore, the only additional calculation required to obtain a bootstrap estimate $\hat{\tau}_{2b}^{Fuzzy}(x_1)$ of $\tau_2(x_1)$ is simply to compute the formula for $\hat{\tau}_{2b}^{Fuzzy}(x_1)$ using the functions $\hat{f}_{1b}$, $\hat{f}_{0b}$, $\hat{p}_{1b}$ and $\hat{p}_{0b}$, which have already been fit.

[43] This situation can only arise in FMMRD, since observations in FMDRD can only be in the treatment or control group by definition.

frontiers[44]. Assuming this holds, it follows that for $\boldsymbol{X}$ arbitrarily close to a frontier, the baseline covariates $\boldsymbol{Z}$ are independent of the treatment W. Taking the frontier F_1 in a MDRD as an example, this implies that for $x_2 \geq 0$,

$$\begin{aligned}
&\lim_{x_1 \to 0^+} \mathbb{E}[Y|X_{1i} = x_1, X_{2i} = x_2, \boldsymbol{Z} = \boldsymbol{z}] \\
&- \lim_{x_1 \to 0^-} \mathbb{E}[Y|X_{1i} = x_1, X_{2i} = x_2, \boldsymbol{Z} = \boldsymbol{z}] \\
&\quad = \lim_{x_1 \to 0^+} \mathbb{E}[Y|X_{1i} = x_1, X_{2i} = x_2] - \lim_{x_1 \to 0^-} \mathbb{E}[Y|X_{1i} = x_1, X_{2i} = x_2].
\end{aligned}$$

So, provided the MRD assumptions hold, the inclusion of additional covariates in the estimation equation for MRD should not change the estimated function by much. Testing the sensitivity of MRD estimates to the inclusion of baseline covariates can thus be used as an additional check on the validity of the MRD design.

A useful feature of TPRS is that it can easily be incorporated into generalized additive models. This provides a great deal of flexibility for how a user may wish to include the baseline covariates in the model. For example, one can estimate

$$Y_i = g_1(\boldsymbol{X_i}) + g_2(\boldsymbol{Z_i}) + \epsilon_i,$$

where g_1 is the TPRS described in section 4, and g_2 is a linear function of the baseline covariates, or

$$Y_i = g(\boldsymbol{X_i}, \boldsymbol{Z_i}) + \epsilon_i,$$

where g is a TPRS in $\boldsymbol{X}$ and $\boldsymbol{Z}$, just to list two possibilities.

Another alternative, motivated by the discussion in Lee and Lemieux (2010), is to "residualize" the outcome variable Y. Specifically, one first fits a model with Y as a function of $\boldsymbol{Z}$, then conducts MRD estimation using the residuals from the first model as the outcome variable.

6.3 More than Two Assignment Variables

Although most of the discussion in this paper has focused on MRD with two assignment variables, the estimation methods described still work in theory for more than two assignment variables, provided appropriate adjustments are made. First, with d assignment variables, the treatment frontiers are $(d-1)$-dimensional surfaces. Hence, the estimated treatment effect for a given frontier is typically a function of $d-1$ assignment variables. Second, the smoothness penalty term J_{md} that is used for fitting TPRS surfaces should satisfy $2m > d+1$.

However, estimation in higher-dimensional assignment variable space is not without difficulties. Perhaps the most important of these is sparsity of data

[44] In fact, the diagnostic graphs that plot predetermined outcomes as functions of the assignment variables (discussed in section 3.2) are designed to check whether this condition holds.

resulting from the curse of dimensionality. The following toy example illustrates this point.

Consider a MDRD with d assignment variables, and assume that values of the assignment variables in the data is uniformly distributed in a d-dimensional unit hypercube, with the cutoffs for each assignment variable at 0.5. Typically, only data that is "close" to the boundary is truly relevant for MRD estimation[45], which I assume to be data that is at most a distance of b from the boundary. Using the distance function induced by the L_∞-norm (i.e. the max-norm) to be generous[46], the total fraction of data that is relevant in this toy example is given by $(0.5+b)^d - (0.5-b)^d$. The fraction of data that is relevant for estimating the treatment function along each frontier is approximately this quantity divided by d, since there are d treatment frontiers in the d-dimensional assignment variable space. The following table shows that the proportion of data that is relevant for estimation decreases rapidly as the dimension d grows (using $b = 0.1$)[47].

Table 10: Toy Example of Curse of Dimensionality

d	Proportion of Data Relevant	Proportion of Data Relevant for Each Frontier
2	0.200	0.100
3	0.152	0.051
4	0.104	0.026
5	0.068	0.014
6	0.043	0.007
7	0.026	0.004
8	0.016	0.002
9	0.010	0.001
10	0.006	0.001

This suggests that when possible, one should try to reduce the dimension of the assignment variable space, especially when several assignment variables represent different measures of the "same" underlying characteristic (in a qualitative sense).

To elaborate, consider an example where eligibility for a college financial aid package depends on family income being below a threshold, and the test score

[45] One should not confuse this point with the fact that TPRS estimation can be done using nearly all the data. In particular, TPRS is a local estimator, so that data far from the boundary has limited effect on estimates at the boundary. Hence, the decision to use most of the data for TPRS estimation is simply to avoid the tedious and tricky task of optimal bandwidth selection, rather than in the hope that data far from the boundary can improve estimates at the boundary.

[46] Using any other distance function induced by L_p-norms for $p < \infty$ will lead to less data being considered relevant for treatment effect estimation.

[47] For MMRD, a greater proportion of the data is closer to at least one frontier. However, the number of frontiers separating the multiple treatment conditions is also greater, so the problem is no less serious.

on each of the three SAT components (reading, writing and language, and math) being at least 600. It would be easy to set this up as a four-dimensional MDRD problem, but this is undesirable due to the curse of dimensionality (which will lead to imprecise estimates due to the sparsity of data close to each frontier in the four-dimensional space). Instead, one can collapse the three SAT scores (X_2, X_3, X_4) into a single assignment variable $\tilde{X}_2$ defined by

$$\tilde{X}_2 = \min\{X_2 - 600, X_3 - 600, X_4 - 600\},$$

since they all represent academic performance. In particular, one is unlikely to gain useful insights by distinguishing between say, the effect of crossing the reading score threshold versus the effect of crossing the writing and language score threshold. So in this context, essentially no information is lost using this dimension reduction method, and the resulting treatment effect estimate is also likely to be more precise, since a greater proportion of data will be close to the treatment boundaries in the new two-dimensional assignment variable space.

However, one would need to make sure that the variables are scaled similarly when using this approach. While this is not a problem for the SAT scores, the inclusion of a GPA requirement will necessitate rescaling of variables before using this dimension-reduction technique, since GPA will almost always be closer to its cutoff than the SAT scores are from theirs.

6.4 Uncertainty over the Number of Treatments

Sometimes, it is not clear exactly what constitutes a treatment condition for a particular MRD. For example, suppose the two assignment variables are reading and math scores which determine whether a student is required to take summer remediation classes for reading and/or math (depending on the respective test scores). Defining the treatment as whether or not a student attends *any* summer remediation results in a MDRD design, whereas distinguishing between *types* of summer remediation leads to a MMRD design.

A practical way to deal with uncertainty over treatment definition is to first estimate a MMRD (which defines a greater number of treatment conditions) and check whether the point-wise CIs for estimated treatment effects along each frontier typically contains zero. In the context of the example above, one may examine the estimated treatment functions for the frontiers separating students who failed one test, and students who failed both tests. If these estimated treatment functions tend not to be significantly different from zero, then one may consider reformulating the problem as a MDRD.

7 Conclusion

While MRD is a useful tool for a wide range of policy evaluations, the current literature on this research design is limited. As such, there is little by way of consensus on how to conduct analysis for MRD. This paper discusses the existing research on MRD, and propose novel methods that address their shortcomings.

In particular, this paper contains the first extensive discussion of graphical methods for MRD. Graphical analysis is a critical component of conventional RD, but has been largely neglected in the MRD literature thus far. I discuss the advantages of estimating flexible treatment effect functions over scalar treatment effects, and provide numerous examples where the former may yield valuable insights. This paper also generalizes an existing method for estimating treatment effect functions using local linear regression, and proposes a separate new method using thin plate regression splines (TPRS) that is easy to implement. Finally, through a simulation study based closely on real data, I show that the estimation approach using TPRS consistently outperforms existing estimation methods on the simulated data.

The simulation results suggest that MRD treatment function estimation using TPRS holds promise. Here, I suggest two related issues that merit further investigation. First, the theoretical properties of the TPRS estimation method have not been developed in depth. For instance, this paper does not provide formal results on its asymptotic convergence properties. Second, a comparison between MRD treatment effect functions estimated on a real dataset using TPRS and local linear regression (with bandwidth chosen either via my generalization of LM CV or a multidimensional analogue of the IK algorithm) may yield additional insights.

Appendix A Simulated Data with Correlated Income and GPA

The simulated dataset in the main text was generated under the assumption that the assignment variables – high school GPA and parental income – were uncorrelated. However, as Patterson and Mattern (2013) document, there is a small but positive correlation between GPA and household income. In this section, I generate simulated datasets similar to the ones in the main paper in most respects, except that a small degree of positive correlation is introduced between high school GPA and parental income. I then compute treatment effect estimates using local linear regression (with $\delta = 0.05$) and TPRS.

Since the values of the assignment variables in this new simulated dataset are different, I modified the DGPs accordingly to ensure that Kane's probit regression for the GPA threshold still produces similar results to the ones presented in his paper. It turns out that with positively correlated assignment variables, Kane's probit regressions consistently overestimate the true treatment effect. Hence, both the constant and non-constant treatment effect functions had to be negative (at least over part of their domains of definition) to guarantee probit regression results similar to those in Kane's paper. Regression tables for the probit model and local linear regression, as well as plots of the estimated MDRD treatment functions for the three estimation methods are shown below.

Table 11: Probit Regression for GPA Threshold for the Simulated Dataset with Correlated Assignment Variables

	Dependent variable:		
	College Going		
	Kane's Data	Simulated Data (Constant Treat)	Simulated Data (Non-Constant Treat)
$\mathbb{I}[X_1 \geq 0]$	0.029*	0.032*	0.030*
	(0.015)	(0.018)	(0.016)
Observations	11,750	11,806	11,806
Order of Polynomial in GPA	3	3	3
Note:	*p<0.1; **p<0.05; ***p<0.01		

As mentioned above, the probit estimates for these datasets are seriously biased upwards. On the other hand, the local linear approach performs far better than with the datasets used in the main text. The local linear estimates are less biased compared to the probit estimates, and have a positive slope for the dataset with an increasing treatment effect. Moreover, the local linear estimate for β_5 is statistically significant at the 10 percent level for the dataset

Table 12: Local Linear Regressions for Simulated Data with Correlated Assignment Variables)

	Dependent variable:	
	College Going	
	Constant Treat	Non-Constant Treat
β_0	0.818^{***}	0.818^{***}
	(0.007)	(0.006)
β_1	0.017	-0.006
	(0.023)	(0.021)
β_2	0.088^{**}	0.088^{**}
	(0.040)	(0.038)
β_3	-0.001^{***}	-0.001^{***}
	(0.0003)	(0.0003)
β_4	0.001	0.001
	(0.002)	(0.002)
β_5	-0.027	0.141^{*}
	(0.090)	(0.084)
β_6	-0.0002	0.001
	(0.001)	(0.001)
β_7	0.002	0.002
	(0.004)	(0.004)
Observations	11,245	11,245
R^2	0.004	0.015
Adjusted R^2	0.003	0.015
Residual Std. Error (df = 11237)	0.366	0.341
F Statistic (df = 7; 11237)	6.053^{***}	24.837^{***}

Note: *p<0.1; **p<0.05; ***p<0.01

The treatment effect estimates along the GPA and income frontiers are respectively given by $\hat{\beta}_1 + \hat{\beta}_6 X_2$ and $\hat{\beta}_1 + \hat{\beta}_5 X_1$. The treatment coefficients implied by the DGP with constant treatment effect are $\beta_1 = -0.005$, $\beta_5 = 0$ and $\beta_6 = 0$. For the DGP with non-constant treatment effect, $\beta_1 = -0.04$, although the coefficients β_5 and β_6 are unable to capture the non-linear perfectly. Nonetheless, specification error aside, one would expect β_5 and β_6 to be positive, given that the treatment effect functions on both treatment frontiers are increasing.

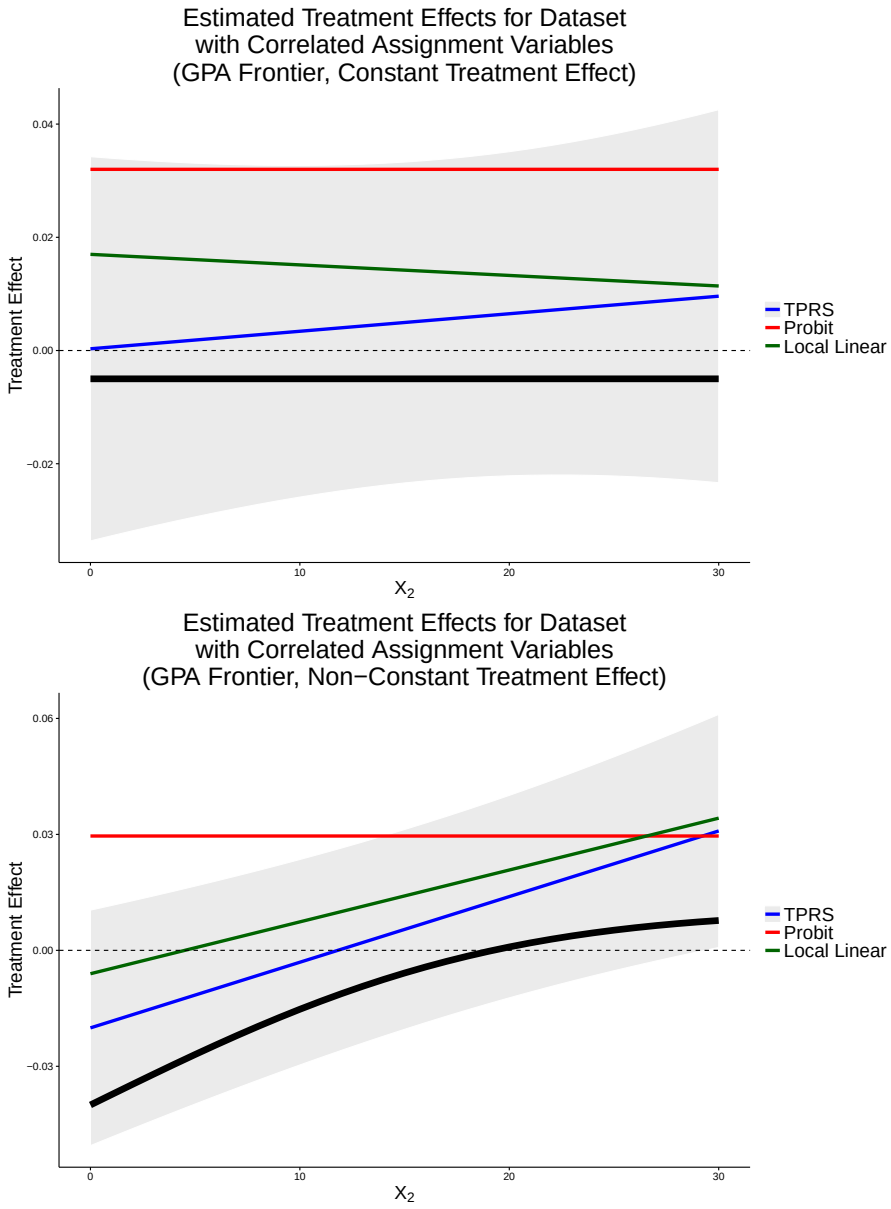

Figure 32: The assignment variables used in the dataset for this graph assume a small positive correlation between high school GPA and parental income. The thick black line in these plots represents the true treatment effect at the GPA threshold for the simulated data. The blue line is the MDRD estimate using TPRS, the red line is the probit estimate using Kane's approach and the dark green line is the local linear estimate with bandwidth selection via LM CV using $\delta = 0.05$. The shaded region represents the 95 percent conservative Bayesian confidence intervals for the TPRS point estimates of the treatment function.

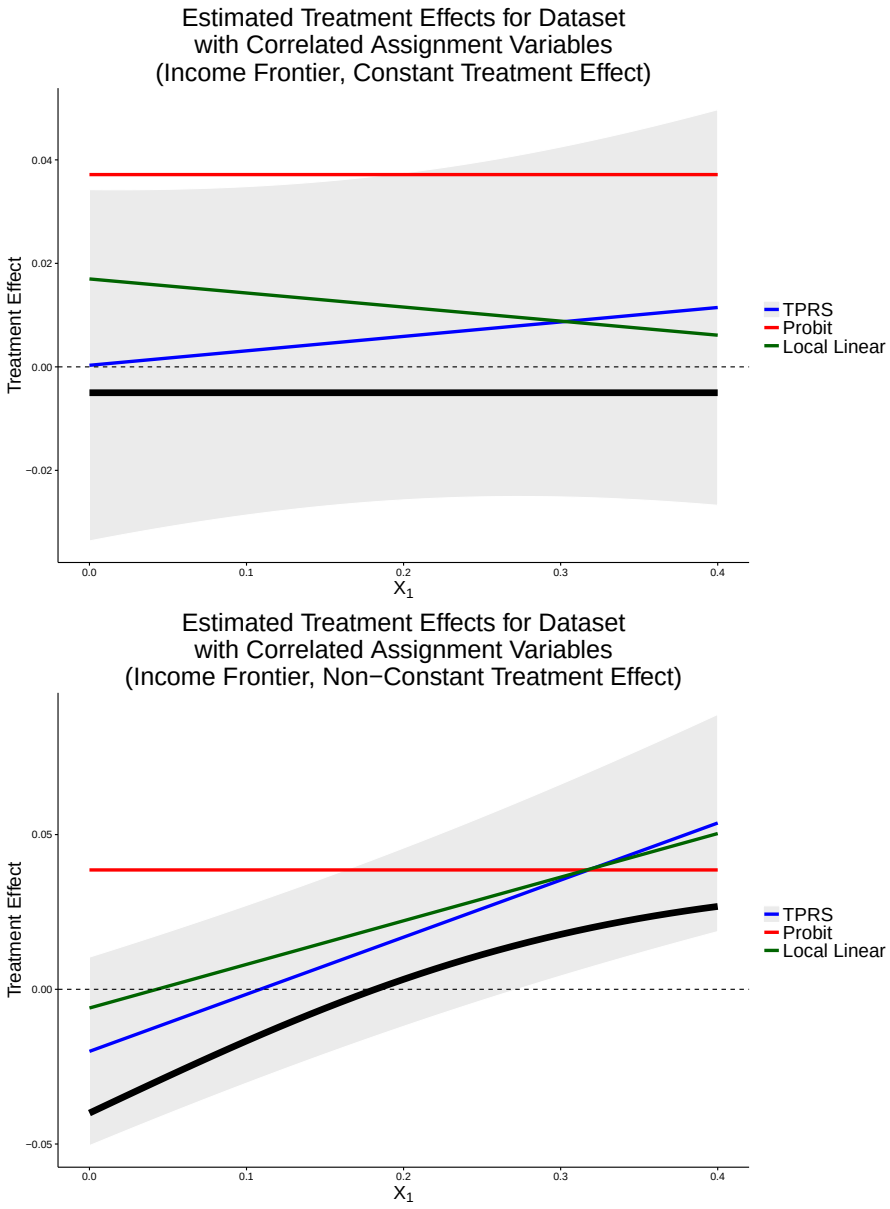

Figure 33: The assignment variables used in the dataset for this graph assume a small positive correlation between high school GPA and parental income. The thick black line in these plots represents the true treatment effect at the income threshold for the simulated data. The blue line is the MDRD estimate using TPRS, the red line is the probit estimate using Kane's approach and the dark green line is the local linear estimate with bandwidth selection via LM CV using $\delta = 0.05$. The shaded region represents the 95 percent conservative Bayesian confidence intervals for the TPRS point estimates of the treatment function.

with non-constant treatment, which indicates (correctly) that there is statistical evidence of a non-constant treatment effect at the income threshold.

Nonetheless, the TPRS approach still outperforms the local linear and probit methods. Its estimates have lower bias than the other two methods, and are increasing for the dataset where the true treatment function is increasing. Again, the true treatment functions fall well within the 95 percent point-wise CIs of the TPRS estimates.

Appendix B Probit Marginal Effects

It was briefly mentioned in the main text that strictly speaking, the marginal probit effect at the income threshold varies for different values of X_1, since the probit regression includes linear and quadratic terms in X_1. Hence, the constant probit effect shown in earlier plots is a simplification.

Yet, as the following graphs show, the actual marginal effects vary minimally as a function of X_1. Moreover, the varying marginal effects shown in these graphs reveal that the probit regressions actually perform worse than implied by the simplified plots. The probit marginal effect should ideally, not vary at all for the simulated data with constant treatment effect. For the simulated data with a non-constant treatment effect that is increasing in X_1, the marginal probit effect varies in the opposite direction, decreasing slightly with X_1.

Therefore, the simplification that assumes a constant probit marginal effect understates how well the TPRS estimates perform relative to the probit estimates.

Appendix C Sensitivity of the TPRS Estimate to Choice of Basis Dimension

In section 4 of the paper, I note that the approximation of thin plate splines by TPRS results in an additional tuning parameter k, which represents the basis dimension used for fitting the TPRS. I also mentioned that as long as k is not chosen to be too small, the exact value chosen is not critical, with Kim and Gu (2004) suggesting $10n^{2/9}$ as a rule-of-thumb.

The estimated treatment effects on the simulated datasets for different choices of k are shown below. Specifically, I considered the default choice provided by the "gam" function in the R package "mgcv" (that is used throughout the paper), $k = 50$, $k = 100$ and $k = 200$. The rule-of-thumb value of k for the simulated dataset lies somewhere between 50 and 100. Reassuringly, the TPRS estimates are almost identical for these different values of k, supporting the assertion that the choice of basis dimension for TPRS is of second-order concern.

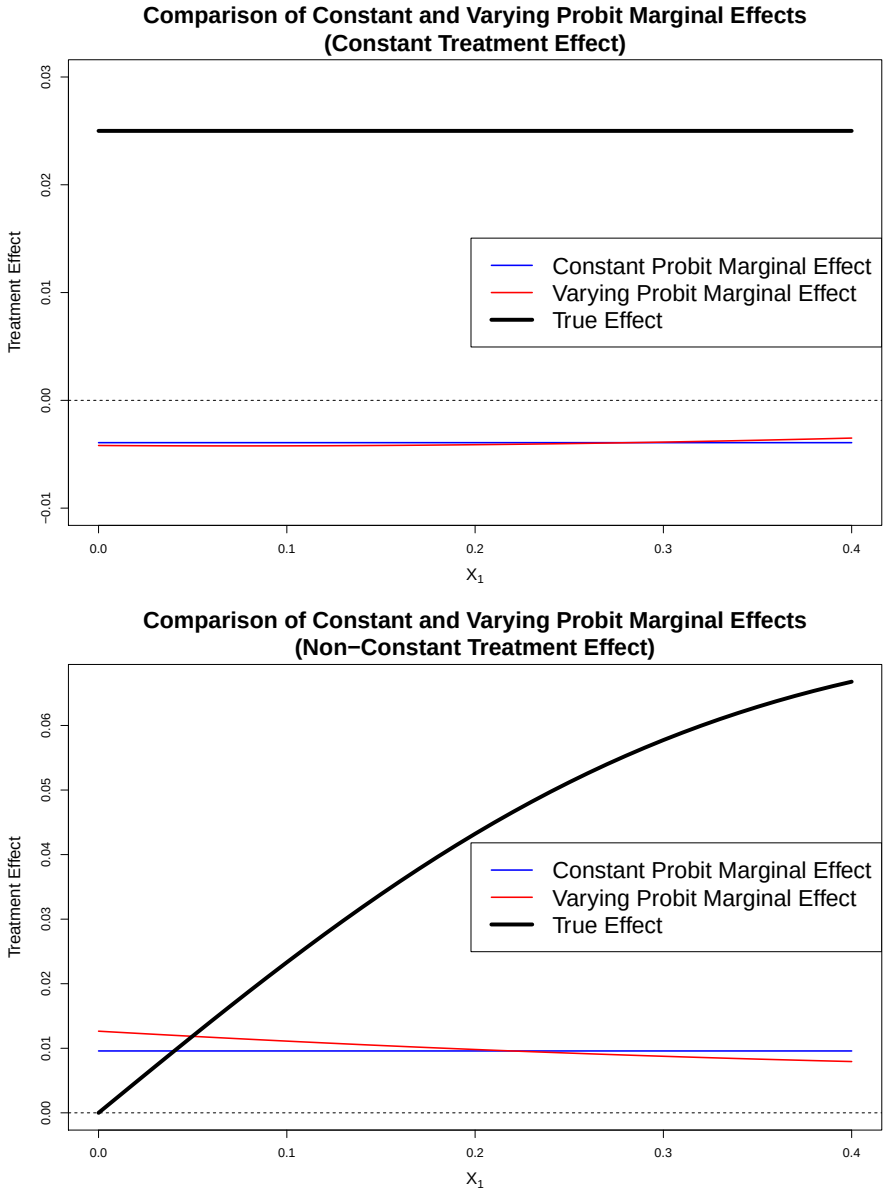

Figure 34: The thick black line in these plots represents the true treatment effect at the income threshold for the simulated data. The blue line represents the constant marginal probit effect that was shown as a simplification throughout the paper. The red line more accurately reflects the marginal probit effect, which varies slightly with X_1 because Kane's probit regression equation at the income threshold includes linear and quadratic terms in X_1.

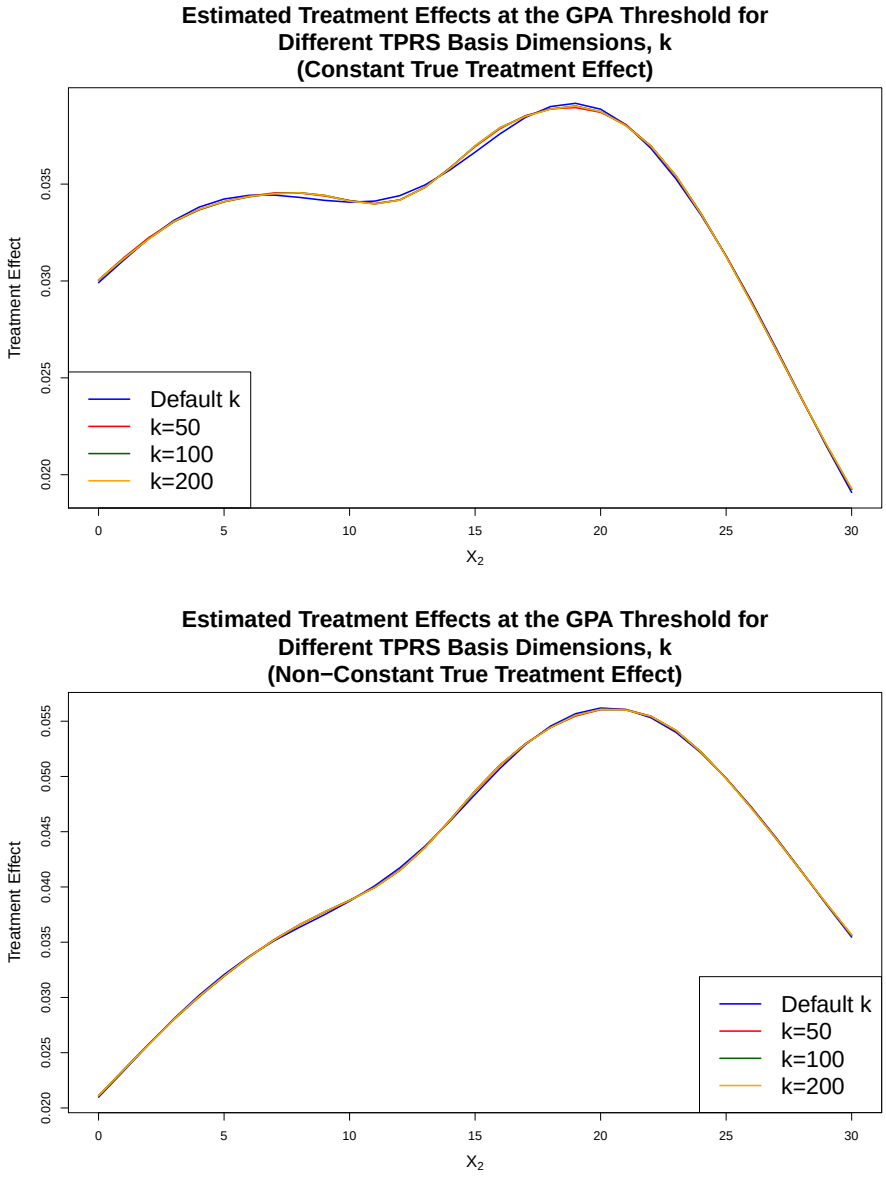

Figure 35: TPRS estimates at the GPA threshold using different choices of basis dimension k.

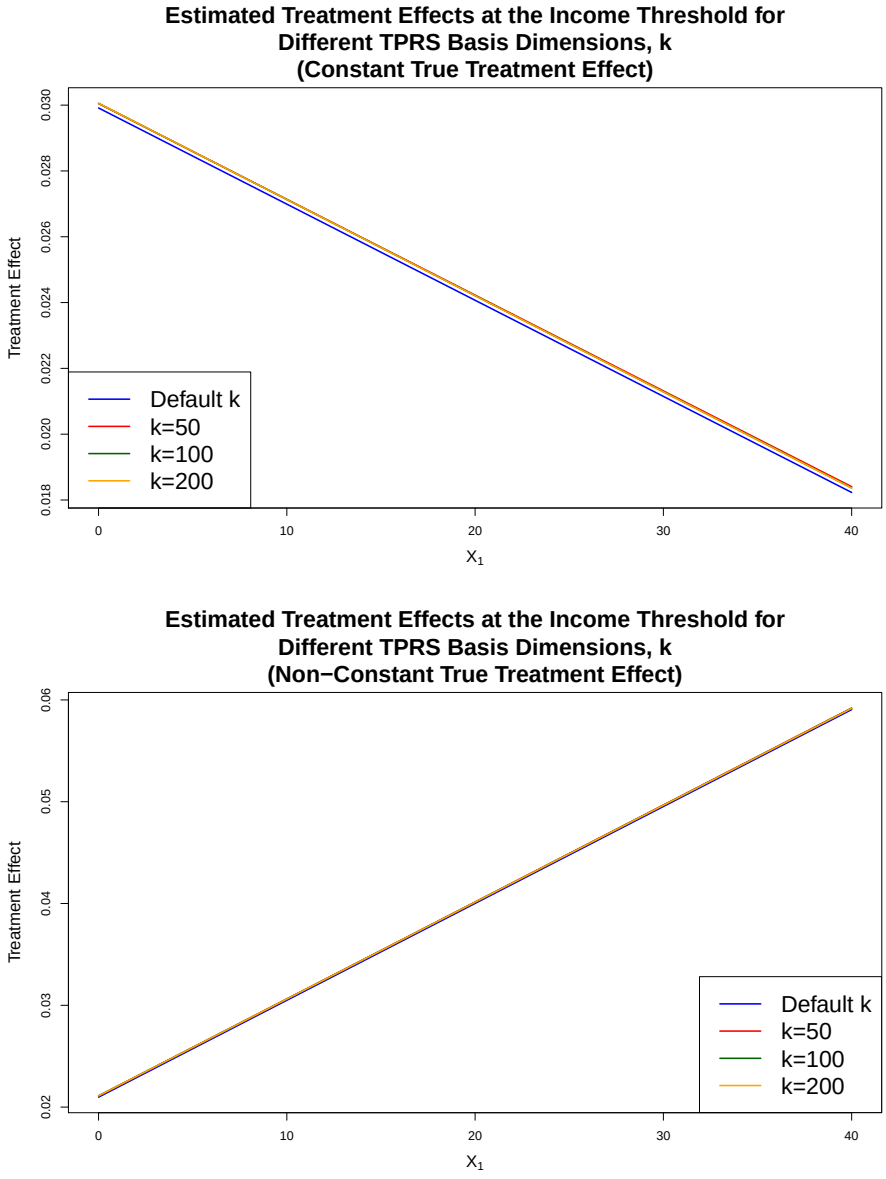

Figure 36: TPRS estimates at the income threshold using different choices of basis dimension k.

Appendix D Additional Details on TPRS

This section of the appendix provides additional details on the approximation of thin plate splines by TPRS. Purely for simplicity of exposition, I assume for this discussion that no two observations have an identical combination of covariate values.

Recall from section 4.3 that a thin plate spline f^* is a solution to the following minimization problem:

$$\underset{f}{\text{minimize}} \quad \sum_{i=1}^{n}(y_i - f(\boldsymbol{x_i}))^2 + \lambda J_{md}(f),$$

where the definition J_{md} can be found in the main text. As long as the restriction $2m > d$ is satisfied, the solution can be written as:

$$f^*(\boldsymbol{x}) = \sum_{i=1}^{n} \delta_i^* \eta_{md}(||\boldsymbol{x} - \boldsymbol{x_i}||) + \sum_{j=1}^{M} \alpha_j^* \phi_j(\boldsymbol{x}),$$

where the orthogonality constraint $\boldsymbol{T'\delta^*} = \boldsymbol{0}$ is satisfied, with $\boldsymbol{T}$ defined by $T_{ij} \equiv \phi_j(\boldsymbol{x_i})$. The functions $M \equiv \binom{m+d-1}{d}$ functions ϕ_j are linearly independent polynomials that span the space of polynomials of degree less than m, and are thus not penalized at all by the penalty term J_{md}. These correspond to the components of "zero wiggliness" mentioned in the main text. The function η_{md} is defined by:

$$\eta_{md}(r) = \begin{cases} \dfrac{(-1)^{m+1+d/2}}{2^{2m-1}\pi^{d/2}(m-1)!(m-d/2)!} r^{2m-d}\log(r) & \text{if } d \text{ is even} \\ \dfrac{\Gamma(d/2-m)}{2^{2m}\pi^{d/2}(m-1)!} r^{2m-d} & \text{if } d \text{ is odd.} \end{cases}$$

Now, defining $\boldsymbol{E}$ by $E_{ij} \equiv \eta_{md}(||\boldsymbol{x_i} - \boldsymbol{x_j}||)$, the minimization problem that defines the thin plate spline can alternatively be written as:

$$(*) \quad \underset{\boldsymbol{\delta},\boldsymbol{\alpha}}{\text{minimize}} \quad ||\boldsymbol{y} - \boldsymbol{E\delta} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta' E \delta} \quad \text{s.t.} \quad \boldsymbol{T'\delta} = \boldsymbol{0}.$$

Leaving the basis for the unpenalized functions untouched, the TPRS focuses on truncating the basis for the $\boldsymbol{\delta}$ parameter space in a way that perturbs the minimization problem as little as possible.

To elaborate, let k be the basis dimension for the TPRS chosen by the user. Instead of searching for the value of $\boldsymbol{\delta}$ over the entire space $\mathbb{R}^n$ that (along with $\boldsymbol{\alpha}$) minimizes the objective function and satisfies the orthogonality constraint, the minimization problem that defines the TPRS only considers possible values of $\boldsymbol{\delta}$ within a k-dimensional subspace, W of $\mathbb{R}^n$.

To make precise how the subspace W is chosen for TPRS (for a given value of k), I introduce the following notation. Given a k-dimensional subspace W of

$\mathbb{R}^n$, let $\boldsymbol{\Gamma_k}$ be a $n \times k$ matrix of rank k with columns that form an orthonormal basis for W. The TPRS minimization problem is thus

$$\underset{\boldsymbol{\delta_k}, \boldsymbol{\alpha}}{\text{minimize}} \quad ||\boldsymbol{y} - \boldsymbol{E\Gamma_k\delta_k} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta_k'\Gamma_k'E\Gamma_k\delta_k} \quad \text{s.t.} \quad \boldsymbol{T'\Gamma_k\delta_k} = \boldsymbol{0},$$

where $\boldsymbol{\delta_k'} \in \mathbb{R}^k$.

In order to express this in a form closer to that of the minimization problem for the thin plate spline, I define the $n \times n$ matrices $\boldsymbol{\tilde{E}_k} \equiv \boldsymbol{E\Gamma_k\Gamma_k'}$ and $\boldsymbol{\hat{E}_k} \equiv \boldsymbol{\Gamma_k\Gamma_k'E\Gamma_k\Gamma_k'}$. This allows me to write the TPRS minimization problem as,

$$(**) \quad \underset{\boldsymbol{\delta}, \boldsymbol{\alpha}}{\text{minimize}} \quad ||\boldsymbol{y} - \boldsymbol{\tilde{E}_k\delta} - \boldsymbol{T\alpha}||^2 + \lambda \boldsymbol{\delta'\hat{E}_k\delta} \quad \text{s.t.} \quad \boldsymbol{T'\delta} = \boldsymbol{0},$$

since $\boldsymbol{\delta} \in W$ if and only if $\boldsymbol{\Gamma_k\delta_k} = \boldsymbol{\delta}$ for some $\boldsymbol{\delta_k} \in \mathbb{R}^k$, by definition of W.

Now, the goal of TPRS is to choose W, or equivalently $\boldsymbol{\Gamma_k}$, so that replacing the matrix $\boldsymbol{E}$ in problem $(*)$ by $\boldsymbol{\tilde{E}_k}$ and $\boldsymbol{\hat{E}_k}$ in problem $(**)$ perturbs problem $(*)$ as little as possible. Unfortunately, there is no k-dimensional subspace that minimizes the change in objective value for *all* possible values of $\boldsymbol{\delta}$. Hence, TPRS instead chooses $\boldsymbol{\Gamma_k}$ to minimize the *worst* possible change in objective value. In other words, $\boldsymbol{\Gamma_k}$ is the orthonormal basis matrix of rank k in $\mathbb{R}^{n \times k}$ that simultaneously minimizes

$$\epsilon_k \equiv \max_{\boldsymbol{\delta} \neq \boldsymbol{0}} \left\{ \frac{||(\boldsymbol{E} - \boldsymbol{\tilde{E}_k})\boldsymbol{\delta}||}{||\boldsymbol{\delta}||} \right\} \quad \text{and} \quad e_k \equiv \max_{\boldsymbol{\delta} \neq \boldsymbol{0}} \left\{ \frac{\boldsymbol{\delta'}(\boldsymbol{E} - \boldsymbol{\hat{E}_k})\boldsymbol{\delta}}{||\boldsymbol{\delta}||^2} \right\},$$

where ϵ_k and e_k correspond to the worst possible change in the least squares and penalty terms respectively.

It turns out that the solution that simultaneously minimizes ϵ_k and e_k is a truncated eigenbasis of $\boldsymbol{E}$. To elaborate, write the spectral decomposition of $\boldsymbol{E}$ as $\boldsymbol{E} = \boldsymbol{UDU'}$ where $\boldsymbol{D}$ is the diagonal matrix containing the eigenvalues of $\boldsymbol{E}$, arranged in decreasing order of magnitude, i.e. $|\boldsymbol{D_{i,i}}| \geq |\boldsymbol{D_{i+1,i+1}}|$ for $i = 1, ..., n-1$.[48] Then, the solution $\boldsymbol{\Gamma_k}$ is the first k columns of $\boldsymbol{U}$, appropriately scaled so that the columns are orthonormal. One may also easily verify that this solution results in $\boldsymbol{\tilde{E}_k} = \boldsymbol{\hat{E}_k}$.

For further details on TPRS, interested readers are encouraged to refer to Wood (2003) and Wood (2006).

[48] This decomposition is possible because $\boldsymbol{E}$ is a real symmetric matrix by definition.

References

[1] **Almond, Douglas, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams.** 2010. "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns." *The Quarterly Journal of Economics* 125 (2): 591–634.

[2] **Clementi, Fabio, and Mauro Gallegati.** 2007. "Pareto's Law of Income Distribution: Evidence for Germany, the United Kingdom, and the United States." In *Econophysics of Wealth Distributions*, edited by Arnab Chatterjee, Sudhakar Yarlagadda, and Bikas K. Chakrabarti. Milan: Springer.

[3] **Jean Duchon.** 1977. "Splines minimizing rotation-invariant semi-norms in Sobolev spaces." In *Construction Theory of Functions of Several Variables*, edited by Walter Schempp, and Karl Zeller. Berlin: Springer.

[4] **Fan, Jianqing, and Irene Gijbels.** 1992. "Variable Bandwidth and Local Linear Regression Smoothers." *The Annals of Statistics* 20 (4): 2008–2036.

[5] **Gelman, Andrew, and Guido W. Imbens.** 2014. "Why High-order Polynomials Should Not be Used in Regression Discontinuity Designs." National Bureau of Economic Research Working Paper 20405.

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

[7] **Imbens, Guido W., and Thomas Lemieux.** 2008. "Regression discontinuity designs: A guide to practice." *Journal of Econometrics* 142 (2): 615–635.

[8] **Jacob, Brian A., and Lars Lefgren.** 2004. "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis." *Review of Economics and Statistics* 86 (1): 226–244.

[9] **Kane, Thomas J.** 2003. "A Quasi-Experimental Estimate of the Impact of Financial Aid on College-Going." National Bureau of Economic Research Working Paper 9703.

[10] **Kim, Young-Ju, and Chong Gu.** 2004. "Smoothing spline Gaussian regression: more scalable computation via efficient approximation." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66 (2): 337–356.

[11] **Lee, David S. and David Card.** 2008. "Regression discontinuity inference with specification error." *Journal of Econometrics* 142 (2): 655–674.

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

[13] **Ludwig, Jens, and Douglas L. Miller.** 2005. "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design." National Bureau of Economic Research Working Paper 11702.

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

[15] **National Center for Health Statistics.** 2008. "2008 Cohort Linked Birth/Infant Death Data Set." United States Department of Health and Human Services. `http://www.cdc.gov/nchs/data_access/vitalstatsonline.htm` (accessed February 21, 2016).

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

[17] **Papay, John P., John B. Willett, and Richard J. Murnane.** 2011b. "High-School Exit Examinations and the Schooling Decisions of Teenagers: A Multi-Dimensional Regression-Discontinuity Analysis." National Bureau of Economic Research Working Paper 17112.

[18] **Patterson, Brian F., and Krista D. Mattern.** 2013. "Validity of the SAT for Predicting First-Year Grades: 2010 SAT Validity Sample." *College Board Statistical Report 2013-2*. New York: The College Board.

[19] **Porter, Jack.** 2003. "Estimation in the Regression Discontinuity Model". Mimeo. Department of Economics, Wisconsin. `http://www.ssc.wisc.edu/~jrporter/reg_discont_2003.pdf` (accessed April 2, 2016).

[20] **Rau, Tomas.** 2011. "Bayesian Inference in the Regression Discontinuity Model." *Vigesimosextas Jornadas Anuales de Economa.*

[21] **Reardon, Sean F., and Joseph P. Robinson.** 2012. "Regression Discontinuity Designs With Multiple Rating-Score Variables." *Journal of Research on Educational Effectiveness* 5 (1): 83–104.

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

[23] **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.

[24] **Wong, Vivian C., Peter M. Steiner, and Thomas D. Cook.** 2013. "Analyzing Regression-Discontinuity Designs With Multiple Assignment Variables: A Comparative Study of Four Estimation Methods." *Journal of Educational and Behavioral Statistics* 38 (2): 107–141.

[25] **Wood, Simon N.** 2006. *Generalized Additive Models: An Introduction with R.* Boca Raton: CRC Press.

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