NBER WORKING PAPER SERIES

REGRESSION DISCONTINUITY DESIGNS: A GUIDE TO PRACTICE

Guido Imbens
Thomas Lemieux

Working Paper 13039
http://www.nber.org/papers/w13039

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 2007

This paper was prepared as an introduction to a special issue of the Journal of Econometrics on regression discontinuity designs. We are grateful for discussions with David Card and Wilbert van der Klaauw. Financial support for this research was generously provided through NSF grant SES 0452590 and the SSHRC of Canada. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

Regression Discontinuity Designs: A Guide to Practice
Guido Imbens and Thomas Lemieux
NBER Working Paper No. 13039
April 2007
JEL No. C14,C21

ABSTRACT

In Regression Discontinuity (RD) designs for evaluating causal effects of interventions, assignment to a treatment is determined at least partly by the value of an observed covariate lying on either side of a fixed threshold. These designs were first introduced in the evaluation literature by Thistlewaite and Campbell (1960). With the exception of a few unpublished theoretical papers, these methods did not attract much attention in the economics literature until recently. Starting in the late 1990s, there has been a large number of studies in economics applying and extending RD methods. In this paper we review some of the practical and theoretical issues involved in the implementation of RD methods.

Guido Imbens
Department of Economics
Littauer Center
Harvard University
1805 Cambridge Street
Cambridge, MA 02138
and NBER
imbens@fas.harvard.edu

Thomas Lemieux
Department of Economics
University of British Columbia
#997-1873 East Mall
Vancouver, BC V6T 1Z1
Canada
and NBER
tlemieux@interchange.ubc.ca

1 Introduction

Since the late 1990s there has been a large number of studies in economics applying and extending RD methods, including Van der Klaauw (2002), Black (1999), Angrist and Lavy (1999), Lee (this volume), Chay and Greenstone (2005), DiNardo and Lee (2004), Chay, McEwan, and Urquiola (2005), McEwan and Shapiro (2007), and Card, Mas and Rothstein (2006). Key theoretical and conceptual contributions include the interpretation of estimates for fuzzy regression discontinuity designs allowing for general heterogeneity of treatment effects (Hahn, Todd and Van der Klaauw, 2001, HTV from hereon), adaptive estimation methods (Sun, 2005), specific methods for choosing bandwidths (Ludwig and Miller, 2005), and various tests for discontinuities in means and distributions of non-affected variables (Lee, this volume; McCrary, this volume).

In this paper, we review some of the practical issues in implementation of RD methods. There is relatively little novel in this discussion. Our general goal is instead to address practical issues in implementing RD designs and review some of the new theoretical developments.

After reviewing some basic concepts in Section 2, the paper focuses on five specific issues in the implementation of RD designs. In Section 3 we stress graphical analyses as powerful methods for illustrating the design. In Section 4 we discuss estimation and suggest using local linear regression methods using only the observations close to the discontinuity point. In Section 5 we propose choosing the bandwidth using cross validation. In Section 6 we provide a simple plug-in estimator for the asymptotic variance and a second estimator that exploits the link with instrumental variables methods derived by HTV. In Section 7 we discuss a number of specification tests and sensitivity analyses based on tests for (a) discontinuities in the average values for covariates, (b) discontinuities in the conditional density of the forcing variable, as suggested by McCrary, and (c) discontinuities in the average outcome at other values of the forcing variable.

2 Sharp and Fuzzy Regression Discontinuity Designs

2.1 Basics

Our discussion will frame the RD design in the context of the modern literature on causal effects and treatment effects, using the Rubin Causal Model (RCM) set up with potential outcomes (Rubin, 1974; Holland, 1986; Imbens and Rubin, 2007), rather than the regression framework that was originally used in this literature. For a general discussion of the RCM and its use in the economic literature, see the survey by Imbens and Wooldridge (2007).

In the basic setting for the RCM (and for the RD design), researchers are interested in the causal effect of a binary intervention or treatment. Units, which may be individuals, firms, countries, or other entities, are either exposed or not exposed to a treatment. The effect of the treatment is potentially heterogenous across units. Let $Y_i(0)$ and $Y_i(1)$ denote the pair of potential outcomes for unit i: $Y_i(0)$ is the outcome without exposure to the treatment, and $Y_i(1)$ is the outcome given exposure to the treatment. Interest is in some comparison of $Y_i(0)$ and $Y_i(1)$. Typically, including in this discussion, we focus on differences $Y_i(1) - Y_i(0)$. The fundamental problem of causal inference is that we never observe the pair $Y_i(0)$ and $Y_i(1)$ together. We therefore typically focus on average effects of the treatment, that is, averages of $Y_i(1) - Y_i(0)$ over (sub-)populations, rather than on unit-level effects. For unit i we observe the outcome corresponding to the treatment received. Let $W_i \in \{0, 1\}$ denote the treatment received, with $W_i = 0$ if unit i was not exposed to the treatment, and $W_i = 1$ otherwise. The outcome observed can then be written as

$$Y_i = (1 - W_i) \cdot Y_i(0) + W_i \cdot Y_i(1) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

In addition to the assignment W_i and the outcome Y_i, we may observe a vector of covariates or pretreatment variables denoted by (X_i, Z_i), where X_i is a scalar and Z_i is an M-vector. A key characteristic of X_i and Z_i is that they are known not to have been affected by the treatment. Both X_i and Z_i are covariates, with a special role played by X_i in the RD design. For each unit we observe the quadruple (Y_i, W_i, X_i, Z_i). We assume that we observe this quadruple for a random sample from some well-defined population.

The basic idea behind the RD design is that assignment to the treatment is deter-

mined, either completely or partly, by the value of a predictor (the covariate X_i) being on either side of a fixed threshold. This predictor may itself be associated with the potential outcomes, but this association is assumed to be smooth, and so any discontinuity of the conditional distribution (or of a feature of this conditional distribution such as the conditional expectation) of the outcome as a function of this covariate at the cutoff value is interpreted as evidence of a causal effect of the treatment.

The design often arises from administrative decisions, where the incentives for units to participate in a program are partly limited for reasons of resource constraints, and clear transparent rules rather than discretion by administrators are used for the allocation of these incentives. Examples of such settings abound. For example, Hahn, Todd and Van der Klaauw (1999) study the effect of an anti-discrimination law that only applies to firms with at least 15 employees. In another example, Matsudaira (this volume) studies the effect of a remedial summer school program that is mandatory for students who score less than some cutoff level on a test (see also Jacob and Lefgren, 2004). Access to public goods such as libraries or museums is often eased by lower prices for individuals depending on an age cutoff value (senior citizen discounts, and discounts for children under some age limit). Similarly, eligibility for medical services through medicare is restricted by age (Card, Dobkin and Maestas, 2006).

2.2 The Sharp Regression Discontinuity Design

It is useful to distinguish between two general settings, the Sharp and the Fuzzy Regression Discontinuity (SRD and FRD from hereon) designs (e.g., Trochim, 1984, 2001; HTV). In the SRD design the assignment W_i is a deterministic function of one of the covariates, the forcing (or treatment-determining) variable X:[1]

$$W_i = 1\{X_i \geq c\}.$$

All units with a covariate value of at least c are assigned to the treatment group (and participation is mandatory for these individuals), and all units with a covariate value

[1]Here we take X_i to be a scalar. More generally, the assignment can be a function of a vector of covariates. Formally, we can write this as the treatment indicator being an indicator for the vector X_i being an element of a subset of the covariate space, or

$$W_i = 1\{X_i \in \mathbb{X}_1\},$$

where $\mathbb{X}_1 \subset \mathbb{X}$, and $\mathbb{X}$ is the covariate space.

less than c are assigned to the control group (members of this group are not eligible for the treatment). In the SRD design we look at the discontinuity in the conditional expectation of the outcome given the covariate to uncover an average causal effect of the treatment:

$$\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x],$$

which is interpreted as the average causal effect of the treatment at the discontinuity point:

$$\tau_{\text{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i = c]. \tag{2.1}$$

Figures 1 and 2 illustrate the identification strategy in the SRD set up. Based on artificial population values, we present in Figure 1 the conditional probability of receiving the treatment, $\Pr(W = 1 | X = x)$ against the covariate x. At $x = 6$ the probability jumps from zero to one. In Figure 2, three conditional expectations are plotted. The two dashed lines in the figure are the conditional expectations of the two potential outcomes given the covariate, $\mu_w(x) = \mathbb{E}[Y(w) | X = x]$, for $w = 0, 1$. These two conditional expectations are continuous functions of the covariate. Note that we can only estimate $\mu_0(x)$ for $x < c$, and $\mu_1(x)$ for $x \geq c$. In addition we plot the conditional expectation of the observed outcome,

$$\mathbb{E}[Y | X = x] =$$

$$\mathbb{E}[Y | W = 0, X = x] \cdot \Pr(W = 0 | X = x) + \mathbb{E}[Y | W = 1, X = x] \cdot \Pr(W = 1 | X = x)$$

in Figure 2, indicated by a solid line. Although the two conditional expectations of the potential outcomes $\mu_w(x)$ are continuous, the conditional expectation of the observed outcome jumps at $x = c = 6$.

Now let us discuss the interpretation of $\lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]$ as an average causal effect in more detail. In the SRD design, the widely used unconfoundedness assumption (e.g., Rosenbaum and Rubin, 1983, Imbens, 2004) underlying most matching-type estimators still holds:

$$Y_i(0), Y_i(1) \;\perp\!\!\!\perp\; W_i \;\Big|\; X_i.$$

This assumption holds in a trivial manner, because conditional on the covariates there is no variation in the treatment. However, this assumption cannot be exploited directly. The problem is that the second assumption that is typically used for matching-type approaches, the overlap assumption which requires that for all values of the covariates there are both treated and control units, or

$$0 < \Pr(W_i = 1|X_i = x) < 1,$$

is fundamentally violated. In fact, for all values of x the probability of assignment is either zero or one, rather than always between zero and one as required by the overlap assumption. As a result, there are no values of x with overlap.

This implies there is a unavoidable need for extrapolation. However, in large samples the amount of extrapolation required to make inferences is arbitrarily small, as we only need to infer the conditional expectation of $Y(w)$ given the covariates ε away from where it can be estimated. To avoid non-trivial extrapolation we focus on the average treatment effect at $X = c$,

$$\tau_{\text{SRD}} = \mathbb{E}\left[Y(1) - Y(0)|X = c\right] = \mathbb{E}\left[Y(1)|X = c\right] - \mathbb{E}\left[Y(0)|X = c\right]. \tag{2.2}$$

By design, there are no units with $X_i = c$ for whom we observe $Y_i(0)$. We therefore will exploit the fact that we observe units with covariate values arbitrarily close to c.[2] In order to justify this averaging we make a smoothness assumption. Typically this assumption is formulated in terms of conditional expectations:

Assumption 2.1 (Continuity of Conditional Regression Functions)

$$\mathbb{E}[Y(0)|X = x] \qquad \text{and} \quad \mathbb{E}[Y(1)|X = x],$$

are continuous in x.

More generally, one might want to assume that the conditional distribution function is smooth in the covariate. Let $F_{Y(w)|X}(y|x) = \Pr(Y(w) \leq y|X = x)$ denote the conditional distribution function of $Y(w)$ given X. Then the general version of the assumption is:

[2] Although in principle the first term in the difference in (2.2) would be straightforward to estimate if we actually observed individuals with $X_i = x$, with continuous covariates we also need to estimate this term by averaging over units with covariate values close to c.

Assumption 2.2 (Continuity of Conditional Distribution Functions)

$$F_{Y(0)|X}(y|x) \qquad \text{and} \quad F_{Y(1)|X}(y|x),$$

are continuous in x for all y.

Both these assumptions are stronger than required, as we will only use continuity at $x = c$, but it is rare that it is reasonable to assume continuity for one value of the covariate, but not at other values of the covariate. We therefore make the stronger assumption.

Under either assumption,

$$\mathbb{E}[Y(0)|X = c] = \lim_{x \uparrow c} \mathbb{E}[Y(0)|X = x] = \lim_{x \uparrow c} \mathbb{E}[Y(0)|W = 0, X = x] = \lim_{x \uparrow c} \mathbb{E}[Y|X = x],$$

and similarly

$$\mathbb{E}[Y(1)|X = c] = \lim_{x \downarrow c} \mathbb{E}[Y|X = x].$$

Thus, the average treatment effect at c, τ_{SRD}, satisfies

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mathbb{E}[Y|X = x] - \lim_{x \uparrow c} \mathbb{E}[Y|X = x].$$

The estimand is the difference of two regression functions at a point. Hence, if we try to estimate this object without parametric assumptions on the two regression functions, we do not obtain root$-N$ consistent estimators. Instead we get consistent estimators that converge to their limits at a slower, nonparametric rates.

As an example of a SRD design, consider the study of the effect of party affiliation of a congressman on congressional voting outcomes by Lee (this volume). See also Lee, Moretti and Butler (2004). The key idea is that electoral districts where the share of the vote for a Democrat in a particular election was just under 50% are on average similar in many relevant respects to districts where the share of the Democratic vote was just over 50%, but the small difference in votes leads to an immediate and big difference in the party affiliation of the elected representative. In this case, the party affiliation always jumps at 50%, making this is a SRD design. Lee looks at the incumbency effect. He is interested in the probability of Democrats winning the subsequent election, comparing districts where the Democrats won the previous election with just over 50% of the popular vote with districts where the Democrats lost the previous election with just under 50% of the vote.

2.3 The Fuzzy Regression Discontinuity Design

In the Fuzzy Regression Discontinuity (FRD) design, the probability of receiving the treatment needs not change from zero to one at the threshold. Instead, the design allows for a smaller jump in the probability of assignment to the treatment at the threshold:

$$\lim_{x \downarrow c} \Pr(W_i = 1 | X_i = x) \neq \lim_{x \uparrow c} \Pr(W_i = 1 | X_i = x),$$

without requiring the jump to equal 1. Such a situation can arise if incentives to participate in a program change discontinuously at a threshold, without these incentives being powerful enough to move all units from nonparticipation to participation. In this design we interpret the ratio of the jump in the regression of the outcome on the covariate to the jump in the regression of the treatment indicator on the covariate as an average causal effect of the treatment. Formally, the estimand is

$$\tau_{\text{FRD}} = \frac{\lim_{x \downarrow c} \mathbb{E}[Y|X=x] - \lim_{x \uparrow c} \mathbb{E}[Y|X=x]}{\lim_{x \downarrow c} \mathbb{E}[W|X=x] - \lim_{x \uparrow c} \mathbb{E}[W|X=x]}.$$

Let us first consider the interpretation of this ratio. HTV, in arguably the most important theoretical paper in the recent RD literature, exploit the instrumental variables connection to interpret the fuzzy regression discontinuity design when the effect of the treatment varies by unit, as in Imbens and Angrist (1994).[3] Let $W_i(x)$ be potential treatment status given cutoff point x, for x in some small neighborhood around c. $W_i(x)$ is equal to one if unit i would take or receive the treatment if the cutoff point was equal to x. This requires that the cutoff point is at least in principle manipulable. For example, if X is age, one could imagine changing the age that makes an individual eligible for the treatment from c to $c + \epsilon$. Then it is useful to assume monotonicity (see HTV):

Assumption 2.3 *$W_i(x)$ is non-increasing in x at $x = c$.*

Next, define compliance status. This concept is similar to the one used in instrumental variables settings (e.g., Angrist, Imbens and Rubin, 1996). A complier is a unit such that

$$\lim_{x \downarrow X_i} W_i(x) = 0, \qquad \text{and} \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

[3] The close connection between FRD and instrumental variables models led researchers in a number of cases to interpret RD designs as instrumental variables settings. See, for example, Angrist and Krueger (1991) and Imbens and Van der Klaauw (1995). The main advantage of thinking of these designs as RD designs is that it suggests the specification analyses from Section 7.

Compliers are units that would get the treatment if the cutoff were at X_i or below, but that would not get the treatment if the cutoff were higher than X_i. To be specific, consider an example where individuals with a test score less than c are encouraged for a remedial teaching program (Matsudaira, this issue). Interest is in the effect of the program on subsequent test scores. Compliers are individuals who would participate if encouraged (if the test score is below the cutoff for encouragement), but not if not encouraged (if test score is above the cutoff for encouragement). Nevertakers are units with

$$\lim_{x \downarrow X_i} W_i(x) = 0, \qquad \text{and} \quad \lim_{x \uparrow X_i} W_i(x) = 0,$$

and alwaystakers are units with

$$\lim_{x \downarrow X_i} W_i(x) = 1, \qquad \text{and} \quad \lim_{x \uparrow X_i} W_i(x) = 1.$$

Then

$$\begin{aligned} \tau_{\text{FRD}} &= \frac{\lim_{x \downarrow c} \mathbb{E}[Y|X=x] - \lim_{x \uparrow c} \mathbb{E}[Y|X=x]}{\lim_{x \downarrow c} \mathbb{E}[W|X=x] - \lim_{x \uparrow c} \mathbb{E}[W|X=x]} \\ &= \mathbb{E}[Y_i(1) - Y_i(0) | \text{unit } i \text{ is a complier and } X_i = c]. \end{aligned}$$

The estimand is an average effect of the treatment, but only averaged for units with $X_i = c$ (by regression discontinuity), and only for compliers (people who are affected by the threshold).

In Figure 3 we plot the conditional probability of receiving the treatment for an FRD design. As in the SRD design, this probability still jumps at $x = 6$, but now by an amount less than one. Figure 4 presents the expectation of the potential outcomes given the covariate and the treatment, $\mathbb{E}[Y(w)|W = w, X = x]$, represented by the dashed lines, as well as the conditional expectation of the observed outcome given the covariate (solid line):

$$\begin{aligned} &\mathbb{E}[Y|X=x] \\ &= \mathbb{E}[Y(0)|W=0, X=x] \cdot \Pr(W=0|X=x) + \mathbb{E}[Y(1)|W=1, X=x] \cdot \Pr(W=1|X=x). \end{aligned}$$

Note that it is no longer necessarily the case here that $\mathbb{E}[Y(w)|W = w, X = x] = \mathbb{E}[Y(w)|X = x]$. Under some assumptions (unconfoundedness) this will be true, but this is not necessary for inference regarding causal effects in the FRD setting.

As an example of a FRD design, consider the study of the effect of financial aid on college attendance by Van der Klaauw (2002). Van der Klaauw looks at the effect of financial aid on acceptance on college admissions. Here X_i is a numerical score assigned to college applicants based on the objective part of the application information (SAT scores, grades) used to streamline the process of assigning financial aid offers. During the initial stages of the admission process, the applicants are divided into L groups based on discretized values of these scores. Let

$$G_i = \begin{cases} 1 & \text{if } 0 \leq X_i < c_1 \\ 2 & \text{if } c_1 \leq X_i < c_2 \\ \vdots & \\ L & \text{if } c_{L-1} \leq X_i \end{cases}$$

denote the financial aid group. For simplicity, let us focus on the case with $L = 2$, and a single cutoff point c. Having a score of just over c will put an applicant in a higher category and increase the chances of financial aid discontinuously compared to having a score of just below c. The outcome of interest in the Van der Klaauw study is college attendance. In this case, the simple association between attendance and the financial aid offer is ambiguous. On the one hand, an aid offer makes the college more attractive to the potential student. This is the causal effect of interest. On the other hand, a student who gets a generous financial aid offer is likely to have better outside opportunities in the form of financial aid offers from other colleges. College aid is emphatically not a deterministic function of the financial aid categories, making this a fuzzy RD design. Other components of the application that are not incorporated in the numerical score (such as the essay and recommendation letters) undoubtedly play an important role. Nevertheless, there is a clear discontinuity in the probability of receiving an offer of a larger financial aid package.

2.4 The FRD Design and Unconfoundedness

In the FRD setting, it is useful to contrast the RD approach with estimation of average causal effects under unconfoundedness. The unconfoundedness assumption (e.g., Rosenbaum and Rubin, 1983; Imbens, 2004) requires that

$$Y(0), Y(1) \perp\!\!\!\perp W \,\Big|\, X.$$

If this assumption holds, then we can estimate the average effect of the treatment at $X = c$ as

$$\mathbb{E}[Y(1) - Y(0)|X = x] = \mathbb{E}[Y|W = 1, X = c] - \mathbb{E}[Y|W = 0, X = c].$$

This approach does not exploit the jump in the probability of assignment at the discontinuity point. Instead it assumes that differences between treated and control units with $X_i = c$ are interpretable as average causal effects.

In contrast, the assumptions underlying a FRD analysis implies that comparing treated and control units with $X_i = c$ is likely to be the wrong approach. Treated units with $X_i = c$ include compliers and alwaystakers, and control units at $X_i = c$ consist of nevertakers. Comparing these different types of units has no causal interpretation under the FRD assumptions. Although, in principle, one cannot test the unconfoundedness assumption, one aspect of the problem makes this assumption fairly implausible. Unconfoundedness is fundamentally based on units being comparable if their covariates are similar. This is not an attractive assumption in the current setting where the probability of receiving the treatment is discontinuous in the covariate. Thus, units with similar values of the forcing variable (but on different sides of the threshold) must be different in some important way related to the receipt of treatment. Unless there is a substantive argument that this difference is immaterial for the comparison of the outcomes of interest, an analysis based on unconfoundedness is not attractive.

2.5 External Validity

One important aspect of both the SRD and FRD designs is that they, at best, provide estimates of the average effect for a subpopulation, namely the subpopulation with covariate value equal to $X_i = c$. The FRD design restricts the relevant subpopulation even further to that of compliers at this value of the covariate. Without strong assumptions justifying extrapolation to other subpopulations (e.g., homogeneity of the treatment effect), the designs never allow the researcher to estimate the overall average effect of the treatment. In that sense the design has fundamentally only a limited degree of external validity, although the specific average effect that is identified may well be of special interest, for example in cases where the policy question concerns changing the location of the threshold. The advantage of RD designs compared to other non-experimental analyses

that may have more external validity, such as those based on unconfoundedness, is that RD designs may have a relatively high degree of internal validity (in settings where they are applicable).

3 Graphical Analyses

3.1 Introduction

Graphical analyses should be an integral part of any RD analysis. The nature of RD designs suggests that the effect of the treatment of interest can be measured by the value of the discontinuity in the expected value of the outcome at a particular point. Inspecting the estimated version of this conditional expectation is a simple yet powerful way to visualize the identification strategy. Moreover, to assess the credibility of the RD strategy, it is useful to inspect two additional graphs for covariates and the density of the forcing variable. The estimators we discuss later use more sophisticated methods for smoothing but these basic plots will convey much of the intuition. For strikingly clear examples of such plots, see Lee, Moretti, and Butler (2004), Lalive (this volume), and Lee (this volume). Note that, in practice, the visual clarity of the plots is often improved by adding smoothed regression lines based on polynomial regressions (or other flexible methods) estimated separately on the two sides of the cutoff point.

3.2 Outcomes by Forcing Variable

The first plot is a histogram-type estimate of the average value of the outcome for different values of the forcing variable, the estimated counterpart to the solid line in Figures 2 and 4. For some binwidth h, and for some number of bins K_0 and K_1 to the left and right of the cutoff value, respectively, construct bins $(b_k, b_{k+1}]$, for $k = 1, \ldots, K = K_0 + K_1$, where

$$b_k = c - (K_0 - k + 1) \cdot h.$$

Then calculate the number of observations in each bin:

$$N_k = \sum_{i=1}^{N} 1\{b_k < X_i \leq b_{k+1}\},$$

and the average outcome in the bin:

$$\overline{Y}_k = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Y_i \cdot 1\{b_k < X_i \leq b_{k+1}\}.$$

The first plot of interest is that of the $\overline{Y}_k$, for $k = 1, \ldots, K$ against the mid point of the bins, $\tilde{b}_k = (b_k + b_{k+1})/2$. The question is whether around the threshold c there is any evidence of a jump in the conditional mean of the outcome. The formal statistical analyses discussed below are essentially just sophisticated versions of this, and if the basic plot does not show any evidence of a discontinuity, there is relatively little chance that the more sophisticated analyses will lead to robust and credible estimates with statistically and substantially significant magnitudes. In addition to inspecting whether there is a jump at this value of the covariate, one should inspect the graph to see whether there are any other jumps in the conditional expectation of Y given X that are comparable to, or larger than, the discontinuity at the cutoff value. If so, and if one cannot explain such jumps on substantive grounds, it would call into question the interpretation of the jump at the threshold as the causal effect of the treatment. In order to optimize the visual clarity it is important to calculate averages that are not smoothed over the cutoff point.

3.3 Covariates by Forcing Variable

The second set of plots compares average values of other covariates in the K bins. Specifically, let Z_i be the M-vector of additional covariates, with m-th element Z_{im}. Then calculate

$$\overline{Z}_{km} = \frac{1}{N_k} \cdot \sum_{i=1}^{N} Z_{im} \cdot 1\{b_k < X_i \leq b_{k+1}\}.$$

The second plot of interest is that of the $\overline{Z}_{km}$, for $k = 1, \ldots, K$ against the mid point of the bins, $\tilde{b}_k$, for all $m = 1, \ldots, M$. In the case of FRD designs, it is also particularly useful to plot the mean values of the treatment variable W_i to make sure there is indeed a jump in the probability of treatment at the cutoff point (as in Figure 3). Plotting other covariates is also useful for detecting possible specification problems (see Section 7.1) in the case of either SRD or FRD designs.

3.4 The Density of the Forcing Variable

In the third graph, one should plot the number of observations in each bin, N_k, against the mid points $\tilde{b}_k$. This plot can be used to inspect whether there is a discontinuity in the distribution of the forcing variable X at the threshold. Such discontinuity would raise the question of whether the value of this covariate was manipulated by the individual agent, invalidating the design. For example, suppose that the forcing variable is a test score. If individuals know the threshold and have the option of re-taking the test, individuals with test scores just below the threshold may do so, and invalidate the design. Such a situation would lead to a discontinuity of the conditional density of the test score at the threshold, and thus be detectable in the kind of plots described here. See Section 7.2 for more discussion of tests based on this idea.

4 Estimation: Local Linear Regression

4.1 Nonparametric Regression at the Boundary

The practical estimation of the treatment effect τ in both the SRD and FRD designs is largely a standard nonparametric regression problem (e.g., Pagan and Ullah, 1999; Härdle, 1990; Li and Racine, 2007). However, there are two unusual features. In this case we are interested in the regression function at a single point, and in addition that single point is a boundary point. As a result, standard nonparametric kernel regression does not work very well. At boundary points, such estimators have a slower rate of convergence than they do at interior points. Here we discuss a more attractive implementation suggested by HTV, among others. First define the conditional means

$$\mu_l(x) = \lim_{z \uparrow x} \mathbb{E}[Y(0)|X = z], \qquad \text{and} \quad \mu_r(x) = \lim_{z \downarrow x} \mathbb{E}[Y(1)|X = z].$$

The estimand in the SRD design is, in terms of these regression functions,

$$\tau_{\text{SRD}} = \mu_r(c) - \mu_l(c).$$

A natural approach is to use standard nonparametric regression methods for estimation of $\mu_l(x)$ and $\mu_r(x)$. Suppose we use a kernel $K(u)$, with $\int K(u)du = 1$. Then the regression functions at x can be estimated as

$$\hat{\mu}_l(x) = \frac{\sum_{i:X_i<c} Y_i \cdot K\left(\frac{X_i - x}{h}\right)}{\sum_{i:X_i<c} K\left(\frac{X_i - x}{h}\right)}, \qquad \text{and} \quad \hat{\mu}_r(x) = \frac{\sum_{i:X_i \geq c} Y_i \cdot K\left(\frac{X_i - x}{h}\right)}{\sum_{i:X_i \geq c} K\left(\frac{X_i - x}{h}\right)},$$

where h is the bandwidth.

The estimator for the object of interest is then

$$\hat{\tau}_{\text{SRD}} = \hat{\mu}_r(x) - \hat{\mu}_l(x) = \frac{\sum_{i:X_i>c} Y_i \cdot K\left(\frac{X_i-x}{h}\right)}{\sum_{i:X_i>c} K\left(\frac{X_i-x}{h}\right)} - \frac{\sum_{i:X_i\leq c} Y_i \cdot K\left(\frac{X_i-x}{h}\right)}{\sum_{i:X_i\leq c} K\left(\frac{X_i-x}{h}\right)}.$$

In order to see the nature of this estimator for the SRD case, it is useful to focus on a special case. Suppose we use a rectangular kernel, e.g., $K(u) = 1/2$ for $-1 < u < 1$, and zero elsewhere. Then the estimator can be written as

$$\begin{aligned}\hat{\tau}_{\text{SRD}} &= \frac{\sum_{i=1}^N Y_i \cdot 1\{c \leq X_i \leq c+h\}}{\sum_{i=1}^N 1\{c \leq X_i \leq c+h\}} - \frac{\sum_{i=1}^N Y_i \cdot 1\{c-h \leq X_i < c\}}{\sum_{i=1}^N 1\{c-h \leq X_i < c\}} \\ &= \overline{Y}_{hr} - \overline{Y}_{hl},\end{aligned}$$

the difference between the average outcomes for observations within a distance h of the cutoff point on the right and left of the cutoff, respectively. N_{hr} and N_{hl} denote the number of observations with $X_i \in [c, c+h]$ and $X_i \in [c-h, c)$, respectively. This estimator can be interpreted as first discarding all observations with a value of X_i more than h away from the discontinuity point c, and then simply differencing the average outcomes by treatment status in the remaining sample.

This simple nonparametric estimator is in general not very attractive, as pointed out by HTV and Porter (2003). Let us look at the approximate bias of this estimator through the probability limit of the estimator for fixed bandwidth. The probability limit of $\hat{\mu}_r(c)$, using the rectangular kernel, is

$$\text{plim}\left[\hat{\mu}_r(c)\right] = \frac{\int_c^{c+h} \mu(x) f(x) dx}{\int_c^{c+h} f(x) dx} = \mu_r(c) + \lim_{x \downarrow c} \frac{\partial}{\partial x}\mu(x) \cdot \frac{h}{2} + O\left(h^2\right).$$

Combined with the corresponding calculation for the control group, we obtain the bias

$$\text{plim}\left[\hat{\mu}_r(c) - \hat{\mu}_l(c)\right] - \mu_r(c) - \mu_l(c) = \frac{h}{2} \cdot \left(\lim_{x \downarrow c} \frac{\partial}{\partial x}\mu(x) + \lim_{x \uparrow c} \frac{\partial}{\partial x}\mu(x)\right) + O\left(h^2\right).$$

Hence the bias is linear in the bandwidth h, whereas when we nonparametrically estimate a regression function in the interior of the support we typically get a bias of order h^2.

Note that we typically do expect the regression function to have a non-zero derivative, even in cases where the treatment has no effect. In many applications the eligibility criterion is based on a covariate that does have some correlation with the outcome, so

that, for example, those with poorest prospects in the absence of the program are in the eligible group. Hence it is likely that the bias for the simple kernel estimator is relatively high.

One practical solution to the high order of the bias is to use a local linear regression (e.g., Fan and Gijbels, 1996). An alternative is to use series regression or sieve methods. Such methods could be implemented in the current setting by adding higher order terms to the regression function. For example, Lee, Moretti and Butler (2004) include fourth order polynomials in the covariate to the regression function. The formal properties of such methods are equally attractive to those of kernel type methods. The main concern is that they are more sensitive to outcome values for observations far away from the cutoff point. Kernel methods using kernels with compact support rule out any sensitivity to such observations, and given the nature of RD designs this can be an attractive feature. Certainly, it would be a concern if results depended in an important way on using observations far away from the cutoff value. In addition, global methods put effort into estimating the regression functions in areas (far away from the discontinuity point) that are of no interest in the current setting.

4.2 Local Linear Regression

Here we discuss local linear regression. See for a general discussion Fan and Gijbels (1996). Instead of locally fitting a constant function, we can fit linear regression functions to the observations within a distance h on either side of the discontinuity point:

$$\min_{\alpha_l,\beta_l} \sum_{i|c-h<X_i<c}^{N} \left(Y_i - \alpha_l - \beta_l \cdot (X_i - c)\right)^2,$$

and

$$\min_{\alpha_r,\beta_r} \sum_{i|c\leq X_i<c+h}^{N} \left(Y_i - \alpha_r - \beta_r \cdot (X_i - c)\right)^2.$$

The value of $\mu_l(c)$ is then estimated as

$$\widehat{\mu_l(c)} = \hat{\alpha}_l + \hat{\beta}_l \cdot (c - c) = \hat{\alpha}_l,$$

and the value of $\mu_r(c)$ is then estimated as

$$\widehat{\mu_r(c)} = \hat{\alpha}_r + \hat{\beta}_r \cdot (c - c) = \hat{\alpha}_r,$$

Given these estimates, the average treatment effect is estimated as

$$\hat{\tau}_{\text{SRD}} = \hat{\alpha}_r - \hat{\alpha}_l.$$

Alternatively one can estimate the average effect directly in a single regression, by solving

$$\min_{\alpha,\beta,\tau,\gamma} \sum_{i=1}^{N} 1\{c-h \leq X_i \leq c+h\} \cdot (Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot W_i - \gamma \cdot (X_i - c) \cdot W_i)^2,$$

which will numerically yield the same estimate of τ_{SRD}.

An alternative is to impose the restriction that the slope coefficients are the same on both sides of the discontinuity point, or $\lim_{x \downarrow c} \frac{\partial}{\partial x}\mu(x) = \lim_{x \uparrow c} \frac{\partial}{\partial x}\mu(x)$. This can be imposed by requiring that $\beta_l = \beta_r$. Although it may be reasonable to expect the slope coefficients for the covariate to be similar on both sides of the discontinuity point, this procedure also has some disadvantages. Specifically, by imposing this restriction one allows for observations on $Y(1)$ from the right of the discontinuity point to affect estimates of $\mathbb{E}[Y(0)|X = c]$ and, similarly, for observations on $Y(0)$ from the left of discontinuity point to affect estimates of $\mathbb{E}[Y(1)|X = c]$. In practice, one might wish to have the estimates of $\mathbb{E}[Y(0)|X = c]$ based solely on observations on $Y(0)$, and not depend on observations on $Y(1)$, and vice versa.

We can make the nonparametric regression more sophisticated by using weights that decrease smoothly as the distance to the cutoff point increases, instead of the zero/one weights based on the rectangular kernel. However, even in this simple case the asymptotic bias can be shown to be of order h^2, and the more sophisticated kernels rarely make much difference. Furthermore, if using different weights from a more sophisticated kernel does make a difference, it likely suggests that the results are highly sensitive to the choice of bandwidth. So the only case where more sophisticated kernels may make a difference is when the estimates are not very credible anyway because of too much sensitivity to the choice of bandwidth. From a practical point of view, one may just want to focus on the simple rectangular kernel, but verify the robustness of the results to different choices of bandwidth.

For inference we can use standard least squares methods. Under appropriate conditions on the rate at which the bandwidth goes to zero as the sample size increases, the resulting estimates will be asymptotically normally distributed, and the (robust) standard errors from least squares theory will be justified. Using the results from HTV, the

optimal bandwidth is $h \propto N^{-1/5}$. Under this sequence of bandwidths the asymptotic distribution of the estimator $\hat{\tau}$ will have a non-zero bias. If one does some undersmoothing, by requiring that $h \propto N^{-\delta}$ with $1/5 < \delta < 2/5$, then the asymptotic bias disappears and standard least squares variance estimators will lead to valid confidence intervals. See Section 6 for more details.

4.3 Covariates

Often there are additional covariates available in addition to the forcing covariate that is the basis of the assignment mechanism. These covariates can be used to eliminate small sample biases present in the basic specification, and improve the precision. In addition, they can be useful for evaluating the plausibility of the identification strategy, as discussed in Section 7.1. Let the additional vector of covariates be denoted by Z_i. We make three observations on the role of these additional covariates.

The first and most important point is that the presence of these covariates rarely changes the identification strategy. Typically, the conditional distribution of the covariates Z given X is continuous at $x = c$. In fact, as we discuss in Section 7, one may wish to test for discontinuities at that value of x in order to assess the plausibility of the identification strategy. If such discontinuities in other covariates are found, the justification of the identification strategy may be questionable. If the conditional distribution of Z given X is continuous at $x = c$, then including Z in the regression

$$\min_{\alpha,\beta,\tau,\delta} \sum_{i=1}^{N} 1\{c-h \leq X_i \leq c+h\} \cdot \left(Y_i - \alpha - \beta \cdot (X_i - c) - \tau \cdot W_i - \gamma \cdot (X_i - c) \cdot W_i - \delta' Z_i\right)^2,$$

will have little effect on the expected value of the estimator for τ, since conditional on X being close to c, the additional covariates Z are independent of W.

The second point is that even though the presence of Z in the regression does not affect any bias when X is very close to c, in practice we often include observations with values of X not too close to c. In that case, including additional covariates may eliminate some bias that is the result of the inclusion of these additional observations.

Third, the presence of the covariates can improve precision if Z is correlated with the potential outcomes. This is the standard argument, which also supports the inclusion of covariates in analyses of randomized experiments. In practice the variance reduction

will be relatively small unless the contribution to the R^2 from the additional regressors is substantial.

4.4 Estimation for the Fuzzy Regression Discontinuity Design

In the FRD design, we need to estimate the ratio of two differences. The estimation issues we discussed earlier in the case of the SRD arise now for both differences. In particular, there are substantial biases if we do simple kernel regressions. Instead, it is again likely to be better to use local linear regression. We use a uniform kernel, with the same bandwidth for estimation of the discontinuity in the outcome and treatment regressions.

First, consider local linear regression for the outcome, on both sides of the discontinuity point. Let

$$\left(\hat{\alpha}_{yl}, \hat{\beta}_{yl}\right) = \arg\min_{\alpha_{yl},\beta_{yl}} \sum_{i:c-h\leq X_i<c} \left(Y_i - \alpha_{yl} - \beta_{yl}\cdot(X_i-c)\right)^2, \tag{4.3}$$

$$\left(\hat{\alpha}_{yr}, \hat{\beta}_{yr}\right) = \arg\min_{\alpha_{yr},\beta_{yr}} \sum_{i:c\leq X_i\leq c+h} \left(Y_i - \alpha_{yr} - \beta_{yr}\cdot(X_i-c)\right)^2. \tag{4.4}$$

The magnitude of the discontinuity in the outcome regression is then estimated as

$$\hat{\tau}_y = \hat{\alpha}_{yr} - \hat{\alpha}_{yl}.$$

Second, consider the two local linear regression for the treatment indicator:

$$\left(\hat{\alpha}_{wl}, \hat{\beta}_{wl}\right) = \arg\min_{\alpha_{wl},\beta_{wl}} \sum_{i:c-h\leq X_i<c} \left(W_i - \alpha_{wl} - \beta_{wl}\cdot(X_i-c)\right)^2, \tag{4.5}$$

$$\left(\hat{\alpha}_{wr}, \hat{\beta}_{wr}\right) = \arg\min_{\alpha_{wr},\beta_{wr}} \sum_{i:c\leq X_i\leq c+h} \left(Y_i - \alpha_{wr} - \beta_{wr}\cdot(X_i-c)\right)^2. \tag{4.6}$$

The magnitude of the discontinuity in the treatment regression is then estimated as

$$\hat{\tau}_w = \hat{\alpha}_{wr} - \hat{\alpha}_{wl}.$$

Finally, we estimate the effect of interest as the ratio of the two discontinuities:

$$\hat{\tau}_{\text{FRD}} = \frac{\hat{\tau}_y}{\hat{\tau}_w} = \frac{\hat{\alpha}_{yr} - \hat{\alpha}_{yl}}{\hat{\alpha}_{wr} - \hat{\alpha}_{wl}}. \tag{4.7}$$

Because of the specific implementation we use here, with a uniform kernel, and the same bandwidth for estimation of the denominator and the numerator, we can characterize the estimator for τ as a Two-Stage-Least-Squares (TSLS) estimator. HTV were the first to note this equality, in the setting with standard kernel regression and no additional covariates. It is a simple extension to show that the equality still holds when we use local linear regression and include additional regressors. Define

$$V_i = \begin{pmatrix} 1 \\ 1\{X_i < c\} \cdot (X_i - c) \\ 1\{X_i \geq c\} \cdot (X_i - c) \end{pmatrix}, \qquad \text{and} \quad \delta = \begin{pmatrix} \alpha_{yl} \\ \beta_{yl} \\ \beta_{yr} \end{pmatrix}. \tag{4.8}$$

Then we can write

$$Y_i = \delta' V_i + \tau \cdot W_i + \varepsilon_i. \tag{4.9}$$

Estimating τ based on the regression function (4.9) by TSLS methods, with the indicator $1\{X_i \geq c\}$ as the excluded instrument and V_i as the set of exogenous variables is numerically identical to $\hat{\tau}_{\text{FRD}}$ as given in (4.7).

5 Bandwidth Selection

An important issue in practice is the selection of the smoothing parameter, the binwidth h. In general there are two approaches to choosing bandwidths. A first approach consists of characterizing the optimal bandwidth in terms of the unknown joint distribution of all variables. The relevant components of this distribution can then be estimated, and plugged into the optimal bandwidth function. The second approach, on which we focus here, is based on a cross-validation procedure. The specific methods discussed here are similar to those developed by Ludwig and Miller (2005, 2007). In particular, their proposals, like ours, are aimed specifically at estimating the regression function at the boundary. Initially we focus on the SRD case, and in Section 5.2 we extend the recommendations to the FRD setting.

To set up the bandwidth choice problem we generalize the notation slightly. In the SRD setting we are interested in

$$\tau_{\text{SRD}} = \lim_{x \downarrow c} \mu(x) - \lim_{x \uparrow c} \mu(x).$$

We estimate the two terms as

$$\widehat{\lim_{x\downarrow c}\mu(x)} = \hat{\alpha}_r(c),$$

and

$$\widehat{\lim_{x\uparrow c}\mu(x)} = \hat{\alpha}_l(c),$$

where $\hat{\alpha}_l(x)$ and $\hat{\beta}_l(x)$ solve

$$\left(\hat{\alpha}_l(x), \hat{\beta}_l(x)\right) = \arg\min_{\alpha,\beta} \sum_{j|x-h<X_j<x} \left(Y_j - \alpha - \beta\cdot(X_j - x)\right)^2. \tag{5.10}$$

and $\hat{\alpha}_r(x)$ and $\hat{\beta}_r(x)$ solve

$$\left(\hat{\alpha}_r(x), \hat{\beta}_r(x)\right) = \arg\min_{\alpha,\beta} \sum_{j|x<X_j<x+h} \left(Y_j - \alpha - \beta\cdot(X_j - x)\right)^2. \tag{5.11}$$

Let us focus first on estimating $\lim_{x\downarrow c}\mu(x)$. For estimation of this limit we are interested in the bandwidth h that minimizes

$$Q_r(x,h) = \mathbb{E}\left[\left(\lim_{z\downarrow x}\mu(z) - \hat{\alpha}_r(x)\right)^2\right],$$

at $x = c$. In principle this could be different from the bandwidth that minimizes the corresponding criterion on the lefthand side,

$$Q_l(x,h) = \mathbb{E}\left[\left(\lim_{x\uparrow c}\mu(x) - \hat{\alpha}_l(c)\right)^2\right],$$

at $x = c$. However, we will focus on a single bandwidth for both sides of the threshold, and therefore focus on minimizing

$$Q(c,h) = \frac{1}{2}\cdot(Q_l(c,h) + Q_r(c,h)) = \frac{1}{2}\cdot\left(\mathbb{E}\left[\left(\lim_{x\uparrow c}\mu(x) - \hat{\alpha}_l(c)\right)^2\right] + \mathbb{E}\left[\left(\lim_{x\downarrow c}\mu(x) - \hat{\alpha}_r(c)\right)^2\right]\right).$$

We now discuss two methods for choosing the bandwidth.

5.1 Bandwidth Selection for the SRD Design

For a given binwidth h, let the estimated regression function at x be

$$\hat{\mu}(x) = \begin{cases} \hat{\alpha}_l(x) & \text{if } x < c, \\ \hat{\alpha}_r(x) & \text{if } x \geq c, \end{cases}$$

where $\hat{\alpha}_l(x)$, $\hat{\beta}_l(x)$, $\hat{\alpha}_r(x)$ and $\hat{\beta}_r(x)$ solve (5.10) and (5.11). Note that in order to mimic the fact that we are interested in estimation at the boundary, we only use the observations on one side of x in order to estimate the regression function at x, rather than the observations on both sides of x, that is, observations with $x - h < X_j < x + h$. In addition, the strict inequality in the definition implies that $\hat{\mu}(x)$ evaluated at $x = X_i$ does not depend on Y_i.

Now define the cross-validation criterion as

$$\text{CV}_Y(h) = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \hat{\mu}(X_i)\right)^2, \tag{5.12}$$

with the corresponding cross-validation choice for the binwidth

$$h_{\text{CV}}^{\text{opt}} = \arg\min_h \text{CV}_Y(h).$$

The expected value of this cross-validation function is, ignoring the term that does not involve h, equal to $\mathbb{E}[\text{CV}_Y(h)] = C + \mathbb{E}[Q(X,h)] = C + \int Q(x,h) f_X(dx)$. Although the modification to estimate the regression using one-sided kernels mimics more closely the estimand of interest, this is still not quite what we are interested in. Ultimately, we are solely interested in estimating the regression function in the neighborhood of a single point, the threshold c, and thus in minimizing $Q(c,h)$, rather than $\int_x Q(x,h) f_X(x) dx$. If there are quite a few observations in the tails of the distribution, minimizing the criterion in (5.12) may lead to larger bins than is optimal for estimating the regression function around $x = c$, if c is in the center of the distribution. We may therefore wish to minimize the cross-validation criterion after first discarding observations from the tails. Let $q_{X,\delta,l}$ be the δ quantile of the empirical distribution of X for the subsample with $X_i < c$, and let $q_{X,\delta,r}$ be the δ quantile of the empirical distribution of X for the subsample with $X_i \geq c$. Then, we may wish to use the criterion

$$\text{CV}_Y^{\delta}(h) = \frac{1}{N} \sum_{i: q_{X,\delta,l} \leq X_i \leq q_{X,1-\delta,r}} \left(Y_i - \hat{\mu}(X_i)\right)^2. \tag{5.13}$$

The modified cross-validation choice for the bandwidth is

$$h_{\text{CV}}^{\delta,\text{opt}} = \arg\min_h \text{CV}_Y^{\delta}(h). \tag{5.14}$$

The modified cross-validation function has expectation, again ignoring terms that do not involve h, proportional to $\mathbb{E}[Q(X,h) | q_{X,\delta,l} < X < q_{X,\delta,r}]$. Choosing a smaller value of δ

makes the expected value of the criterion closer to what we are ultimately interested in, that is, $Q(c,h)$, but has the disadvantage of leading to a noisier estimate of $\mathbb{E}[\text{CV}_Y^\delta(h)]$. In practice, one may wish to choose $\delta = 1/2$, and discard 50% of the observations on either side of the threshold, and afterwards assess the sensitivity of the bandwidth choice to the choice of δ. Ludwig and Miller (2005) implement this by using only data within 5 percentage points of the threshold on either side.

Note that, in principle, we can use a different binwidth on either side of the cutoff value. However, it is likely that the density of the forcing variable x is similar on both sides of the cutoff point. If, in addition, the curvature is similar on both sides close to the cutoff point, then in large samples the optimal binwidth will be similar on both sides. Hence, the benefits of having different binwidths on the two sides may not be sufficient to balance the disadvantage of the additional noise in estimating the optimal value from a smaller sample.

5.2 Bandwidth Selection for the FRD Design

In the FRD design, there are four regression functions that need to be estimated: the expected outcome given the forcing variable, both on the left and right of the cutoff point, and the expected value of the treatment variable, again on the left and right of the cutoff point. In principle, we can use different binwidths for each of the four nonparametric regressions.

In the section on the SRD design, we argued in favor of using identical bandwidths for the regressions on both sides of the cutoff point. The argument is not so clear for the pairs of regression functions by outcome we have here. In principle, we have two optimal bandwidths, one based on minimizing $\text{CV}_Y^\delta(h)$, and one based on minimizing $\text{CV}_W^\delta(h)$, defined correspondingly. It is likely that the conditional expectation of the treatment variable is relatively flat compared to the conditional expectation of the outcome variable, suggesting one should use a larger binwidth for estimating the former.[4] Nevertheless, in practice it is appealing to use the same binwidth for numerator and denominator. To avoid asymptotic biases, one may wish to use the smallest bandwidth selected by the

[4] In the extreme case of the SRD design where the conditional expectation of W given X is flat on both sides of the threshold, the optimal bandwidth would be infinity. Therefore, in practice it is likely that the optimal bandwidth for estimating the jump in the conditional expectation of the treatment would be larger than the bandwidth for estimating the conditional expectation of the outcome.

cross validation criterion applied separately to the outcome and treatment regression:

$$h_{\text{CV}}^{\text{opt}} = \min\left(\arg\min_h \text{CV}_Y^\delta(h), \arg\min_h \text{CV}_W^\delta(h)\right),$$

where $\text{CV}_Y^\delta(h)$ is as defined in (5.12), and $\text{CV}_W^\delta(h)$ is defined similarly. Again, a value of $\delta = 1/2$ is likely to lead to reasonable estimates in many settings.

6 Inference

We now discuss some asymptotic properties for the estimator for the FRD case given in (4.7) or its alternative representation in (4.9).[5] More general results are given in HTV. We continue to make some simplifying assumptions. First, as in the previous sections, we use a uniform kernel. Second, we use the same bandwidth for the estimator for the jump in the conditional expectation of the outcome and treatment. Third, we undersmooth, so that the square of the bias vanishes faster than the variance, and we can ignore the bias in the construction of confidence intervals. Fourth, we continue to use the local linear estimator.

Under these assumptions we do two things. First, we give an explicit expression for the asymptotic variance. Second, we present two estimators for the asymptotic variance. The first estimator follows explicitly the analytic form for the asymptotic variance, and substitutes estimates for the unknown quantities. The second estimator is the standard robust variance for the Two-Stage-Least-Squares (TSLS) estimator, based on the sample obtained by discarding observations when the forcing covariate is more than h away from the cutoff point. The asymptotic variance and the corresponding estimators reported here are robust to heteroskedasticity.

6.1 The Asymptotic Variance

To characterize the asymptotic variance we need a couple of additional pieces of notation. Define the four variances

$$\sigma_{Yl}^2 = \lim_{x\uparrow c} \text{Var}(Y|X=x), \qquad \sigma_{Yr}^2 = \lim_{x\downarrow c} \text{Var}(Y|X=x),$$

[5] The results for the SRD design are a special case of those for the FRD design. In the SRD design, only the first term of the asympotic variance in equation (6.18) is left since $V_{\tau_w} = C_{\tau_y,\tau_w} = 0$, and the variance can also be estimated using the standard robust variance for OLS instead of TSLS.

$$\sigma^2_{Wl} = \lim_{x \uparrow c} \text{Var}(W|X=x), \qquad \sigma^2_{Wr} = \lim_{x \downarrow c} \text{Var}(W|X=x),$$

and the two covariances

$$C_{YWl} = \lim_{x \uparrow c} \text{Cov}(Y,W|X=x), \qquad C_{YWr} = \lim_{x \downarrow c} \text{Cov}(Y,W|X=x).$$

Note that, because of the binary nature of W, it follows that $\sigma^2_{Wl} = \mu_{Wl} \cdot (1 - \mu_{Wl})$, where $\mu_{Wl} = \lim_{x \uparrow c} \Pr(W=1|X=x)$, and similarly for σ^2_{Wr}. To discuss the asymptotic variance of $\hat{\tau}$, it is useful to break it up in three pieces. The asymptotic variance of $\sqrt{Nh}(\hat{\tau}_y - \tau_y)$ is

$$V_{\tau_y} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Yr} + \sigma^2_{Yl}\right). \tag{6.15}$$

The asymptotic variance of $\sqrt{Nh}(\hat{\tau}_w - \tau_w)$ is

$$V_{\tau_w} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Wr} + \sigma^2_{Wl}\right) \tag{6.16}$$

The asymptotic covariance of $\sqrt{Nh}(\hat{\tau}_y - \tau_y)$ and $\sqrt{Nh}(\hat{\tau}_w - \tau_w)$ is

$$C_{\tau_y,\tau_w} = \frac{4}{f_X(c)} \cdot \left(C_{YWr} + C_{YWl}\right). \tag{6.17}$$

Finally, the asymptotic distribution has the form

$$\sqrt{Nh} \cdot (\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\tau^2_w} \cdot V_{\tau_y} + \frac{\tau^2_y}{\tau^4_w} \cdot V_{\tau_w} - 2 \cdot \frac{\tau_y}{\tau^3_w} \cdot C_{\tau_y,\tau_w}\right). \tag{6.18}$$

This asymptotic distribution is a special case of that in HTV (page 208), using the rectangular kernel, and with $h \propto N^{-\delta}$, for $1/5 < \delta < 2/5$ (so that the asymptotic bias can be ignored).

6.2 A Plug-in Estimator for the Asymptotic Variance

We now discuss two estimators for the asymptotic variance of $\hat{\tau}$. First, we can estimate the asymptotic variance of $\hat{\tau}$ by estimating each of the components, τ_w, τ_y, V_{τ_w}, V_{τ_y}, and C_{τ_y,τ_w} and substituting them into the expression for the variance in (6.18). In order to do this we first estimate the residuals

$$\hat{\varepsilon}_i = Y_i - \hat{\mu}_y(X_i) = Y_i - 1\{X_i < c\} \cdot \hat{\alpha}_{yl} - 1\{X_i \geq c\} \cdot \hat{\alpha}_{yr},$$

$$\hat{\eta}_i = W_i - \hat{\mu}_w(X_i) = W_i - 1\{X_i < c\} \cdot \hat{\alpha}_{wl} - 1\{X_i \geq c\} \cdot \hat{\alpha}_{wr}.$$

Then we estimate the variances and covariances consistently as

$$\hat{\sigma}^2_{Yl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\varepsilon}^2_i, \qquad \hat{\sigma}^2_{Yr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\varepsilon}^2_i,$$

$$\hat{\sigma}^2_{Wl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\eta}^2_i, \qquad \hat{\sigma}^2_{Wr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\eta}^2_i,$$

$$\hat{C}_{YWl} = \frac{1}{N_{hl}} \sum_{i|c-h \leq X_i < c} \hat{\varepsilon}_i \cdot \hat{\eta}_i, \qquad \hat{C}_{YWr} = \frac{1}{N_{hr}} \sum_{i|c \leq X_i \leq c+h} \hat{\varepsilon}_i \cdot \hat{\eta}_i.$$

Finally, we estimate the density consistently as

$$\hat{f}_X(x) = \frac{N_{hl} + N_{hr}}{2 \cdot N \cdot h}.$$

Then we can plug in the estimated components of V_{τ_y}, V_{τ_W}, and C_{τ_Y,τ_W} from (6.15)-(6.17), and finally substitute these into the variance expression in (6.18).

6.3 The TSLS Variance Estimator

The second estimator for the asymptotic variance of $\hat{\tau}$ exploits the interpretation of the $\hat{\tau}$ as a TSLS estimator, given in (4.9). The variance estimator is equal to the robust variance for TSLS based on the subsample of observations with $c-h \leq X_i \leq c+h$, using the indicator $1\{X_i \geq c\}$ as the excluded instrument, the treatment W_i as the endogenous regressor and the V_i defined in (4.8) as the exogenous covariates.

7 Specification Testing

There are generally two main conceptual concerns in the application of RD designs, sharp or fuzzy. A first concern about RD designs is the possibility of other changes at the same cutoff value of the covariate. Such changes may affect the outcome, and these effects may be attributed erroneously to the treatment of interest. For example, at age 65 individuals become eligible for discounts at many cultural institutions. However, if one finds that there is a discontinuity in the number of hours worked by age at 65, this is unlikely to be the result of these discounts. The more plausible explanation is that there are other institutional changes that affect incentives to work at age 65. The effect of discounts on

attendance at these cultural institutions, which may well be present, may be difficult to detect due to the many other changes at age 65.

The second concern is that of manipulation of the forcing variable. Consider the Van der Klaauw example where the value of an aggregate admission score affected the likelihood of receiving financial aid. If a single admissions officer scores the entire application packet of any one individual, and if this person is aware of the importance of this cutoff point, they may be more or less likely to score an individual just below the cutoff value. Alternatively, if applicants know the scoring rule, they may attempt to change particular parts of their application in order to end up on the right side of the threshold, for example by retaking tests. If it is costly to do so, the individuals retaking the test may be a selected sample, invalidating the basic RD design.

We also address the issue of sensitivity to the bandwidth choice, and more generally small sample concerns. We end the section by discussing how, in the FRD setting, one can compare the RD estimates to those based on unconfoundedness.

7.1 Tests Involving Covariates

One category of tests involves testing the null hypothesis of a zero average effect on pseudo outcomes known not to be affected by the treatment. Such variables includes covariates that are, by definition, not affected by the treatment. Such tests are familiar from settings with identification based on unconfoundedness assumptions (e.g., Heckman and Hotz, 1989; Rosenbaum, 1987; Imbens, 2004). In the RD setting, they have been applied by Lee, Moretti and Butler (2004) and others. In most cases, the reason for the discontinuity in the probability of the treatment does not suggest a discontinuity in the average value of covariates. If we find such a discontinuity, it typically casts doubt on the assumptions underlying the RD design. In principle, it may be possible to make the assumptions underlying the RD design conditional on covariates, and so a discontinuity in the conditional expectation of the covariates does not necessarily invalidate the approach. In practice, however, it is difficult to rationalize such discontinuities with the rationale underlying the RD approach.

7.2 Tests of Continuity of the Density

The second test is conceptually somewhat different, and unique to the RD setting. McCrary (this volume) suggests testing the null hypothesis of continuity of the density of the covariate that underlies the assignment at the discontinuity point, against the alternative of a jump in the density function at that point. Again, in principle, one does not need continuity of the density of X at c, but a discontinuity is suggestive of violations of the no-manipulation assumption. If in fact individuals partly manage to manipulate the value of X in order to be on one side of the cutoff rather than the other, one might expect to see a discontinuity in this density at the cutoff point. For example, if the variable underlying the assignment is age with a publicly known cutoff value c, and if age is self-reported, one might see relatively few individuals with a reported age just below c, and relatively many individuals with a reported age of just over c. Even if such discontinuities are not conclusive evidence of violations of the RD assumptions, at the very least, inspecting this density would be useful to assess whether it exhibits unusual features that may shed light on the plausibility of the design.

7.3 Testing for Jumps at Non-discontinuity Points

A third set of tests involves estimating jumps at points where there should be no jumps. As in the treatment effect literature (e.g., Imbens, 2004), the approach used here consists of testing for a zero effect in settings where it is known that the effect should be zero.

Here we suggest a specific way of implementing this idea by testing for jumps at the median of the two subsamples on either side of the cutoff value. More generally, one may wish to divide the sample up in different ways, or do more tests. As before, let $q_{X,\delta,l}$ and $q_{X,\delta,r}$ be the δ quantiles of the empirical distribution of X in the subsample with $X_i < c$ and $X_i \geq c$, respectively. Now take the subsample with $X_i < c$, and test for a jump at the median of the forcing variable. Splitting this subsample at its median increases the power of the test to find jumps. Also, by only using observations on the left of the cutoff value, we avoid estimating the regression function at a point where it is known to have a discontinuity. To implement the test, use the same method for selecting the binwidth as before, and estimate the jump in the regression function at $q_{X,1/2,l}$. Also, estimate the standard errors of the jump and use this to test the hypothesis of a zero jump. Repeat

this using the subsample to the right of the cutoff point with $X_i \geq c$. Now estimate the jump in the regression function and at $q_{X,1/2,r}$, and test whether it is equal to zero.

7.4 RD Designs with Misspecification

Lee and Card (this volume) study the case where the forcing variable X is discrete. In practice this is of course always the case. This implies that ultimately one relies for identification on functional form assumptions for the regression function $\mu(x)$. Lee and Card consider a parametric specification for the regression function that does not fully saturate the model, that is, it has fewer free parameters than there are support points. They then interpret the deviation between the true conditional expectation $\mathbb{E}[Y|X = x]$ and the estimated regression function as random specification error that introduces a group structure on the standard errors. Lee and Card then show how to incorporate this group structure into the standard errors for the estimated treatment effect. This approach will tend to widen the confidence intervals for the estimated treatment effect, sometimes considerably, and leads to more conservative and typically more credible inferences. Within the local linear regression framework discussed in the current paper, one can calculate the Lee-Card standard errors (possibly based on slightly coarsened covariate data if X is close to continuous) and compare them to the conventional ones.

7.5 Sensitivity to the Choice of Bandwidth

All these tests are based on estimating jumps in nonparametric regression or density functions. This brings us to the third concern, the sensitivity to the bandwidth choice. Irrespective of the manner in which the bandwidth is chosen, one should always investigate the sensitivity of the inferences to this choice, for example, by including results for bandwidths twice (or four times) and half (or a quarter of) the size of the originally chosen bandwidth. Obviously, such bandwidth choices affect both estimates and standard errors, but if the results are critically dependent on a particular bandwidth choice, they are clearly less credible than if they are robust to such variation in bandwidths. See Lee, Moretti, and Butler (2004) and Lemieux and Milligan (this volume) for examples of papers where the sensitivity of the results to bandwidth choices is explored.

7.6 Comparisons to Estimates Based on Unconfoundedness in the FRD Design

When we have a FRD design, we can also consider estimates based on unconfoundedness (Battistin and Rettore, this volume). In fact, we may be able to estimate the average effect of the treatment conditional on any value of the covariate X under that assumption. Inspecting such estimates and especially their variation over the range of the covariate can be useful. If we find that, for a range of values of X, our estimate of the average effect of the treatment is relatively constant and similar to that based on the FRD approach, one would be more confident in both sets of estimates.

8 Conclusion: A Summary Guide to Practice

In this paper, we reviewed the literature on RD designs and discussed the implications for applied researchers interested in implementing these methods. We end the paper by providing a summary guide of steps to be followed when implementing RD designs. We start with the case of SRD, and then add a number of details specific to the case of FRD.

Case 1: Sharp Regression Discontinuity (SRD) Designs

1. Graph the data (Section 3) by computing the average value of the outcome variable over a set of bins. The binwidth has to be large enough to have a sufficient amount of precision so that the plots looks smooth on either side of the cutoff value, but at the same time small enough to make the jump around the cutoff value clear.

2. Estimate the treatment effect by running linear regressions on both sides of the cutoff point. Since we propose to use a rectangular kernel, these are just standard regression estimated within a bin of width h on both sides of the cutoff point. Note that:

    - Standard errors can be computed using standard least square methods (robust standard errors)
    - The optimal bandwidth can be chosen using cross-validation methods (Section 5)

3. The robustness of the results should be assessed by employing various specification tests.

    - Looking at possible jumps in the value of other covariates at the cutoff point (Section 7.1)
    - Testing for possible discontinuities in the conditional density of the forcing variable (Section 7.2).
    - Looking whether the average outcome is discontinuous at other values of the forcing variable is (Section 7.3).
    - Using various values of the bandwidth (Section 7.5, with and without other covariates that may be available.

Case 2: Fuzzy Regression Discontinuity (FRD) Designs

A number of issues arise in the case of FRD designs in addition to those mentioned above.

1. Graph the average outcomes over a set of bins as in the case of SRD, but also graph the probability of treatment.

2. Estimate the treatment effect using TSLS, which is numerically equivalent to computing the ratio in the estimate of the jump (at the cutoff point) in the outcome variable over the jump in the treatment variable.

    - Standard errors can be computed using the usual (robust) TSLS standard errors (Section 6.3), though a plug-in approach can also be used instead (Section 6.2).
    - The optimal bandwidth can again be chosen using a modified cross-validation procedure (Section 5)

3. The robustness of the results can be assessed using the various specification tests mentioned in the case of SRD designs. In addition, FRD estimates of the treatment effect can be compared to standard estimates based on unconfoundedness.

References

Angrist, J.D., G.W. Imbens, and D.B. Rubin, 1996, Identification of Causal Effects Using Instrumental Variables, *Journal of the American Statistical Association* 91, 444–472.

Angrist, J.D. and A.B. Krueger, 1991, Does Compulsory School Attendance Affect Schooling and Earnings?, *Quarterly Journal of Economics* 106, 979-1014.

Angrist, J.D., and V. Lavy, 1999, Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement", *Quarterly Journal of Economics* 114, 533-575.

Battistin, E. and E. Rettore, 2007, Ineligibles and Eligible Non-Participants as a Double Comparison Group in Regression-Discontinuity Designs, *Journal of Econometrics*, this issue.

Black, S., 1999, Do Better Schools Matter? Parental Valuation of Elementary Education, *Quarterly Journal of Economics* 114, 577-599.

Card, D., C. Dobkin and N. Maestas, 2004, The Impact of Nearly Universal Insurance Coverage on Health Care Utilization and Health: Evidence from Medicare, NBER Working Paper No. 10365.

Card, D., A. Mas, and J. Rothstein, 2006, Tipping and the Dynamics of Segregation in Neighborhoods and Schools, Unpublished Manuscript, Department of Economics, Princeton University.

Chay, K., and M. Greenstone, 2005, Does Air Quality Matter; Evidence from the Housing Market, *Journal of Political Economy* 113, 376-424.

Chay, K., McEwan, P., and M. Urquiola, 2005, The Central Role of Noise in Evaluating Interventions That Use Test Scores to Rank Schools, *American Economic Review* 95, 1237-1258.

DiNardo, J., and D.S. Lee, 2004, Economic Impacts of New Unionization on Private Sector Employers: 1984-2001, *Quarterly Journal of Economics* 119, 1383-1441.

Fan, J. and I. Gijbels, 1996, *Local Polynomial Modelling and Its Applications* (Chapman and Hall, London).

Hahn, J., P. Todd and W. Van der Klaauw, 1999, Evaluating the Effect of an Anti Discrimination Law Using a Regression-Discontinuity Design, NBER Working Paper 7131.

Hahn, J., P. Todd and W. Van der Klaauw, 2001, Identification and Estimation of Treatment Effects with a Regression Discontinuity Design, *Econometrica* 69, 201-209.

Härdle, W., 1990, *Applied Nonparametric Regression* (Cambridge University Press, New York).

Heckman, J.J. and J. Hotz, 1989, Alternative Methods for Evaluating the Impact of Training Programs (with discussion), *Journal of the American Statistical Association* 84, 862-874.

Holland, P., 1986, Statistics and Causal Inference (with discussion), *Journal of the American Statistical Association*, 81, 945-970.

Imbens, G., 2004, Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review, *Review of Economics and Statistics* 86, 4-30.

Imbens, G., and J. Angrist, 1994, Identification and Estimation of Local Average Treatment Effects, *Econometrica* 61, 467-476.

Imbens, G. and D. Rubin, 2007, *Causal Inference: Statistical Methods for Estimating Causal Effects in Biomedical, Social, and Behavioral Sciences*, Cambridge University Press, forthcoming.

Imbens, G. and W. van der Klaauw, 1995, Evaluating the Cost of Conscription in The Netherlands, *Journal of Business and Economic Statistics* 13, 72-80.

Imbens, G. and J. Wooldridge, 2007, Recent Developments in the Econometrics of Program Evaluation, Unpublished Manuscript, Department of Economics, Harvard University.

Jacob, B.A., and L. Lefgren, 2004, Remedial Education and Student Achievement: A Regression-Discontinuity Analysis, *Review of Economics and Statistics* 68, 226-244.

Lalive, R., 2007, How do Extended Benefits affect Unemployment Duration? A Regression Discontinuity Approach, *Journal of Econometrics*, this issue.

Lee, D.S., 2007, Randomized Experiments from Non-random Selection in U.S. House Elections, *Journal of Econometrics*, this issue.

Lee, D.S. and D. Card, 2007, Regression Discontinuity Inference with Specification Error, *Journal of Econometrics*, this issue.

Lee, D.S., Moretti, E., and M. Butler, 2004, Do Voters Affect or Elect Policies? Evidence from the U.S. House, *Quarterly Journal of Economics* 119, 807-859.

Lemieux, T. and K. Milligan, 2007, Incentive Effects of Social Assistance: A Regression Discontinuity Approach, *Journal of Econometrics*, this issue.

Li, Q., and J. Racine, 2007, *Nonparametric Econometrics* (Princeton University Press, Princeton, New Jersey).

Ludwig, J., and D. Miller, 2005, Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, NBER working paper 11702.

Ludwig, J., and D. Miller, 2007, Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design, *Quarterly Journal of Economics* 122(1), 159-208.

Matsudaira, J., 2007, Mandatory Summer School and Student Achievement, *Journal of Econometrics*, this issue.

McCrary, J., 2007, Testing for Manipulation of the Running Variable in the Regression Discontinuity Design, *Journal of Econometrics*, this issue.

McEwan, P. and J. Shapiro, 2007, The benefits of delayed primary school enrollment: Discontinuity estimates using exact birth dates, Wellesley College and LSE working paper.

Pagan, A. and A. Ullah, 1999, *Nonparametric Econometrics*, Cambridge University Press, New York.

Porter, J., 2003, Estimation in the Regression Discontinuity Model," mimeo, Department of Economics, University of Wisconsin, http://www.ssc.wisc.edu/ jporter/reg_discont_2003.pdf.

ROSENBAUM, P., AND D. RUBIN, 1983, The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika* 70, 41-55.

ROSENBAUM, P., 1987, The role of a second control group in an observational study (with discussion), *Statistical Science* 2, 292–316.

RUBIN, D., 1974, Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies, *Journal of Educational Psychology* 66, 688-701.

SUN, Y., 2005, Adaptive Estimation of the Regression Discontinuity Model, Unpublished Manuscript, Department of Economics, University of California at San Diego.

THISTLEWAITE, D., AND D. CAMPBELL, 1960, Regression-Discontinuity Analysis: An Alternative to the Ex-Post Facto Experiment, *Journal of Educational Psychology* 51, 309-317.

TROCHIM, W., 1984, *Research Design for Program Evaluation; The Regression-discontinuity Design* (Sage Publications, Beverly Hills, CA).

TROCHIM, W., 2001, Regression-Discontinuity Design, in N.J. Smelser and P.B Baltes, eds., *International Encyclopedia of the Social and Behavioral Sciences* 19 (Elsevier North-Holland, Amsterdam) 12940-12945.

VAN DER KLAAUW, W., 2002, Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression–discontinuity Approach, *International Economic Review* 43, 1249-1287.

Fig 1: Assignment Probabilities (Sharp RD)

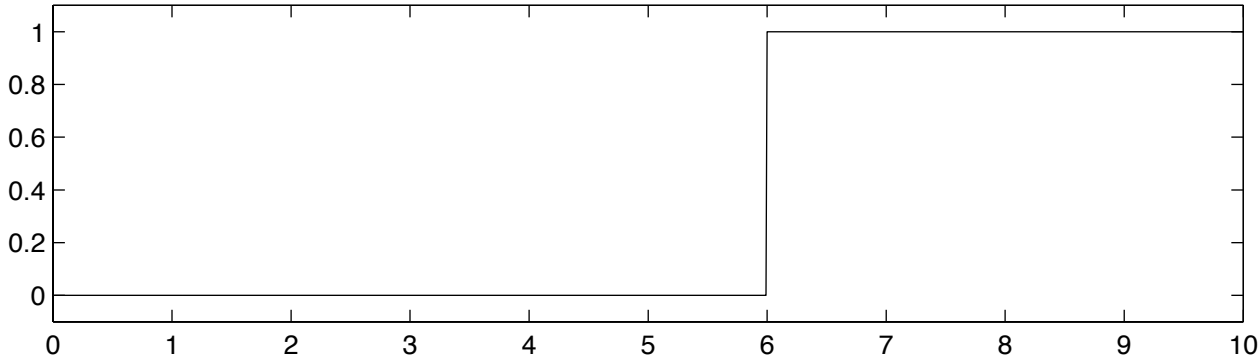

Fig 2: Potential and Observed Outcome Regression Functions

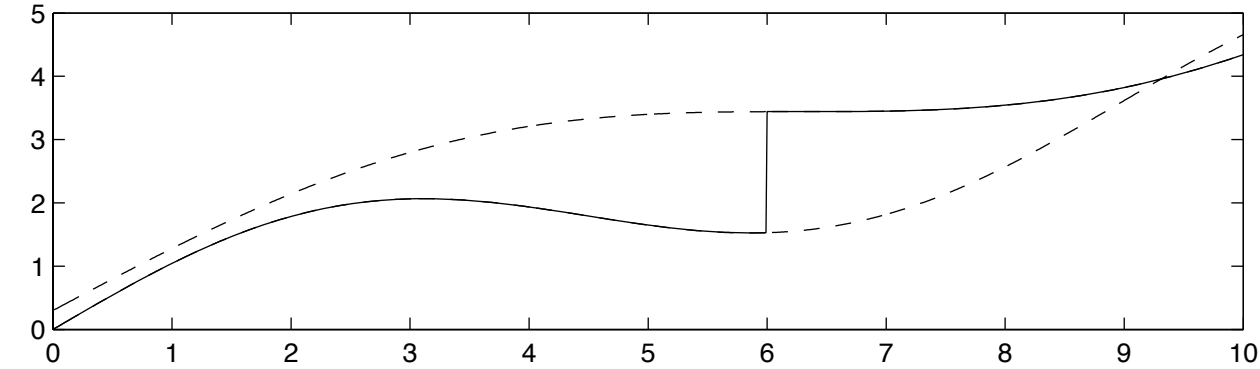

Fig 3: Assignment Probabilities (Fuzzy RD)

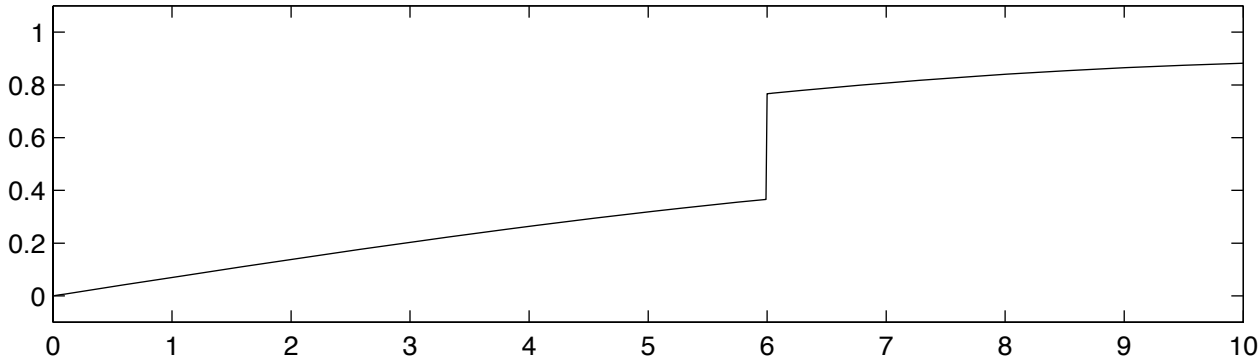

Fig 4: Potential and Observed Outcome Regression (Fuzzy RD)