IRP Lectures **Madison, WI, August 2008**

Lecture 12, Tuesday, Aug 5th, 4.15-5.30pm

Regression Discontinuity Designs[1]

1. Introduction

Since the late 1990s there has been a large number of studies in economics applying and extending Regression Discontinuity (RD) methods from its origins in the statistics literature in the early 60's (Thisthlewaite and Cook, 1960). Here, we review some of the practical issues in implementation of RD methods. The focus is on five specific issues. The first is the importance of graphical analyses as powerful methods for illustrating the design. Second, we suggest using local linear regression methods using only the observations close to the discontinuity point. Third, we discuss choosing the bandwidth using cross validation specifically tailored to the focus on estimation of regression functions on the boundary of the support, following Ludwig and Miller (2005). Fourth, we provide two simple estimators for the asymptotic variance, one of them exploiting the link with instrumental variables methods derived by Hahn, Todd, and VanderKlaauw (2001, HTV). Finally, we discuss a number of specification tests and sensivitity 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 (2007), (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 potential outcomes framework (Rubin, 1974), rather than the regression framework that was originally used in this literature. For unit i there are two potential outcomes, $Y_i(0)$ and $Y_i(1)$, with the causal effect defined as the difference

[1] These notes draw heavily on Imbens and Lemieux (2008).

$Y_i(1) - Y_i(0)$, and the observed outcome equal to

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

where $W_i \in \{0, 1\}$ is the binary indicator for the treatment.

The basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of a predictor (the forcing variable X_i) being on either side of a common threshold. This predictor X_i may itself be associated with the potential outcomes, but this association is assumed to be smooth, and so any discontinuity in the conditional distribution of the outcome, indexed by the value 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.

2.2 The Sharp Regression Discontinuity Design

It is useful to distinguish between two designs, 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:

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

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

$$\lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x] = \lim_{x \downarrow c} \mathbb{E}[Y_i(1)|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i(0)|X_i = x], \quad (1)$$

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

$$\tau_{\text{SRD}} = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = c]. \quad (2)$$

In order to justify this interpretation we make a smoothness assumption. Typically this assumption is formulated in terms of conditional expectations[2]:

Assumption 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.

Under this assumption,

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

The estimand is the difference of two regression functions at a point.

There is a unavoidable need for extrapolation, because 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.[3]

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 (2007). 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

[2]More generally, one might want to assume that the conditional distribution function is smooth in the covariate. Let $F_{Y(w)|X}(y|x) = \Pr(Y_i(w) \leq y|X_i = x)$ denote the conditional distribution function of $Y_i(w)$ given X_i. Then the general version of the assumption assume that $F_{Y(0)|X}(y|x)$ and $F_{Y(1)|X}(y|x)$ are continuous in x for all y. Both 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.

[3]Although in principle the first component in the difference in (1) would be straightforward to estimate if we actually observe 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.

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 need 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

$$\tau_{\text{FRD}}=\frac{\lim_{x\downarrow c}\mathbb{E}[Y_i|X_i=x]-\lim_{x\uparrow c}\mathbb{E}[Y_i|X_i=x]}{\lim_{x\downarrow c}\mathbb{E}[W_i|X_i=x]-\lim_{x\uparrow c}\mathbb{E}[W_i|X_i=x]}.$$

As an example of a FRD design, consider the study of the effect of financial aid on college attendance by VanderKlaauw (2002). VanderKlaauw 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 just over c will put an applicant in a higher category and

increase the chances of financial aid discontinuously compared to having a score just below c. The outcome of interest in the VanderKlaauw study is college attendance. In this case, the statistical association between attendance and the financial aid offer is ambiguous. On the one hand, an aid offer by a college makes that 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 from one college is likely to have better outside opportunities in the form of financial aid offers from other colleges. In the VanderKlaauw application College aid is emphatically not a deterministic function of the financial aid categories, making this a fuzzy RD design. Other components of the college application package 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.

Let us first consider the interpretation of τ_{FRD}. HTV exploit the instrumental variables connection to interpret the fuzzy regression discontinuity design when the effect of the treatment varies by unit. Let $W_i(x)$ be potential treatment status given cutoff point x, for x in some small neigborhood 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.[4] 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 *$W_i(x)$ is non-increasing in x at $x = c$.*

Next, define compliance status. This concept is similar to that in instrumental variables, e.g., Imbens and Angrist (1994), Angrist, Imbens and Rubin (1996). A complier is a unit such that

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

[4] Alternatively, one could think of the individual characteristic X_i as being manipulable, but in many cases this is an immutable characteristic such as age.

Compliers are units who would get the treatment if the cutoff were at X_i or below, but they 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, 2007). Interest is in the effect of the remedial teaching program on subsequent test scores. Compliers are individuals who would participate if encouraged (if the cutoff for encouragement is below or equal to their actual score), but not if not encouraged (if the cutoff for encouragement is higher than their actual score). Then

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

$$=\mathbb{E}[Y_i(1)-Y_i(0)|\text{unit } i \text{ is a complier and } X_i=c].$$

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).

3. The FRD Design, Unconfoundedness and External Validity

3.1 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

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

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

$$\mathbb{E}[Y_i(1)-Y_i(0)|X_i=x]=\mathbb{E}[Y_i|W_i=1,X_i=c]-\mathbb{E}[Y_i|W_i=0,X_i=c].$$

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 an 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 only 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 in this setting.

3.2 The FRD Design and 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 (population) 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 generally have a relatively high degree of internal validity in settings where they are applicable.

4. Graphical Analyses

4.1 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. 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 (2007), and Lee (2007). Two figures from Lee (2007) are attached.

4.2 Outcomes by Forcing Variable

The first plot is a histogram-type estimate of the average value of the outcome by the forcing variable. 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 $k = 1, K$ 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_i given X_i 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. The attached figure is taken from the paper by Lee (2007).

4.2 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 $k = 1, K$ against the mid point of the bins, $\tilde{b}_k$, for all $m = 1, \ldots, M$. Lee (2007) presents such a figure for a lagged value of the outcome, namely the election results from a prior election, against the vote share in the last election. 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. Plotting other covariates is also useful for detecting possible specification problems (see Section 8) in the case of either SRD or FRD designs.

4.3 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. McCrary (2007) suggests that such discontinuity would raise the question whether the value of this covariate was manipulated

Figure IIa: Candidate's Probability of Winning Election t+1, by Margin of Victory in Election t: local averages and parametric fit

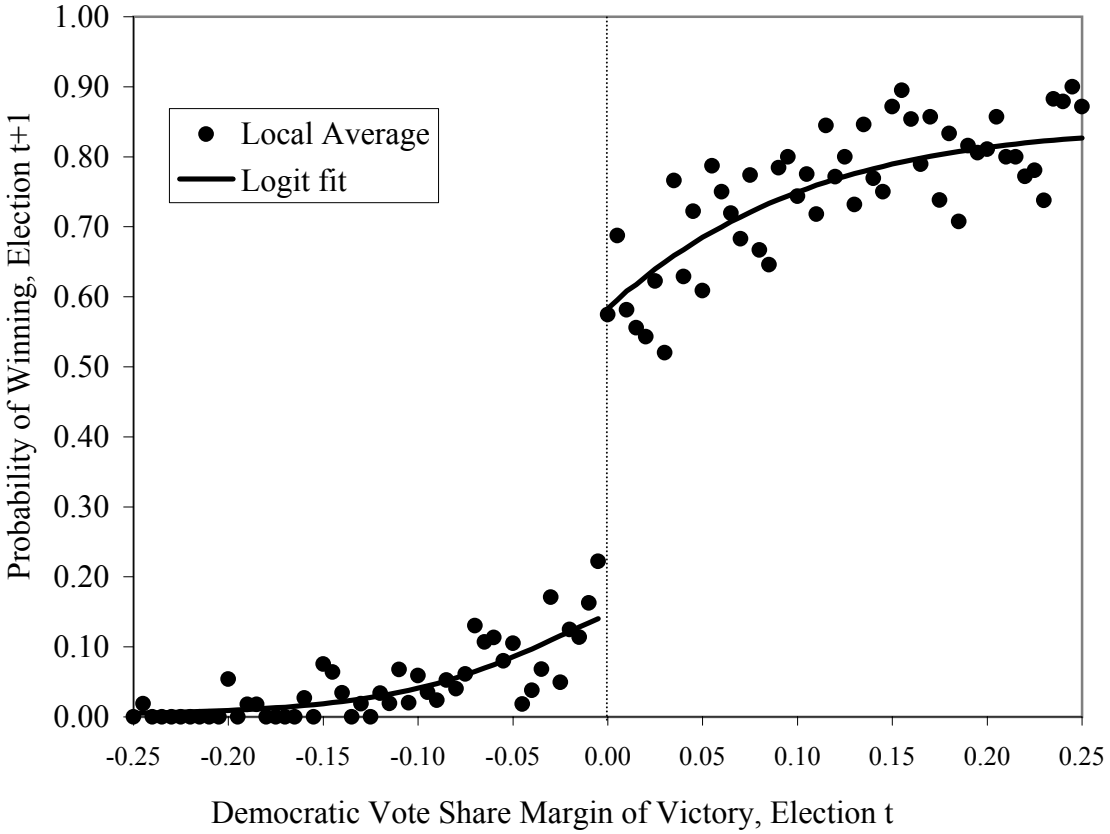
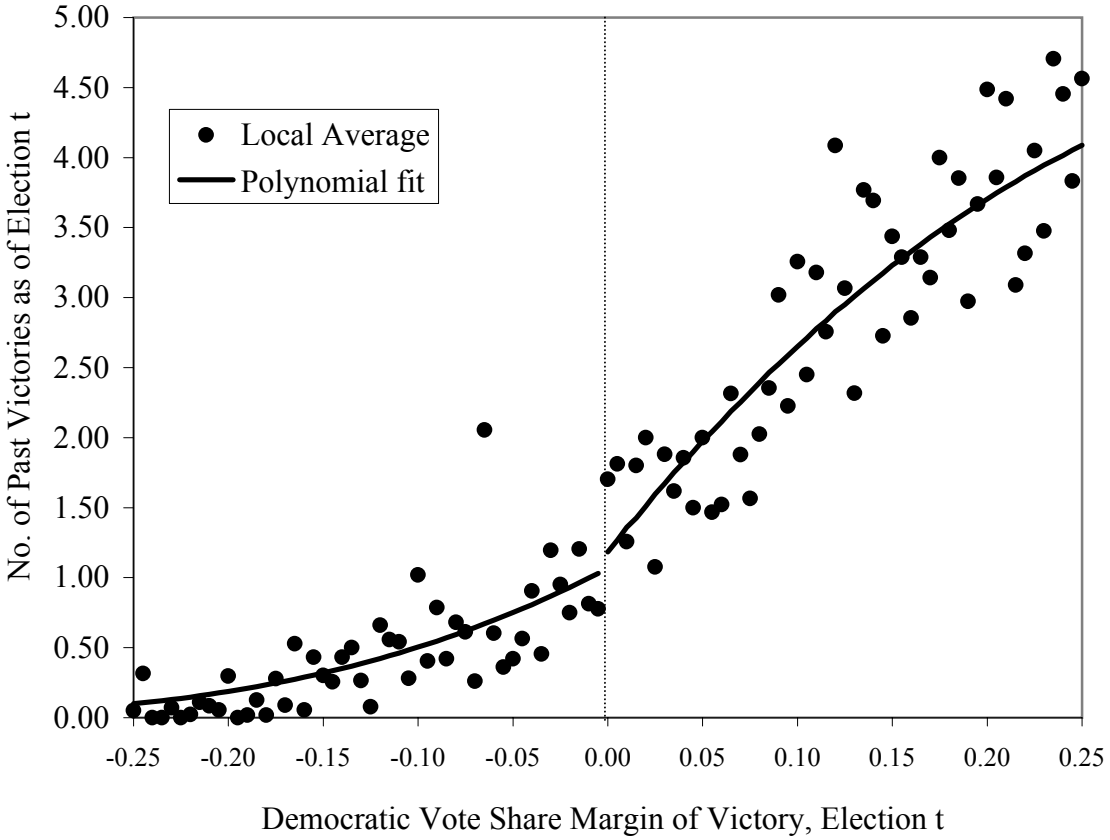


Figure IIb: Candidate's Accumulated Number of Past Election Victories, by Margin of Victory in Election t: local averages and parametric fit



by the individual agents, 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 plots such as described here. See Section 8 for more discussion of the specification tests based on this idea.

5. Estimation: Local Linear Regression

5.1 Nonparametric Regression at the Boundary

The practical estimation of the treatment effect τ in both the SRD and FRD designs is largely standard nonparametric regression (e.g., Pagan and Ullah, 1999; Härdle, 1990; Li and Racine, 2007). However, there are two unusual features to estimation in the RD setting. First, we are interested in the regression function at a single point, and second, 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. Standard methods for choosing the bandwidth are also not designed to provide good choices in this setting.

5.2 Local Linear Regression

Here we discuss local linear regression (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 $\mu_l(c)$ and $\mu_r(c)$ are then estimated as

$$\widehat{\mu_l(c)} = \hat{\alpha}_l + \hat{\beta}_l \cdot (c - c) = \hat{\alpha}_l, \qquad \text{and} \quad \widehat{\mu_r(c)} = \hat{\alpha}_r + \hat{\beta}_r \cdot (c - c) = \hat{\alpha}_r,$$

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

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

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

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 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.

5.3 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 8.1. 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 $x = c$. 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

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

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 with X very close to c, the presence of Z in the regression does not affect any bias, 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 even in analyses of randomized experiments. In practice the variance reduction will be relatively small unless the contribution to the $\mathbb{R}^2$ from the additional regressors is substantial.

5.4 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 (rectangular) 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

$$\left(\hat{\alpha}_{yl},\hat{\beta}_{yl}\right)=\arg\min_{\alpha_{yl},\beta_{yl}}\sum_{i:c-h\leq X_i<c}\left(Y_i-\alpha_{yl}-\beta_{yl}\cdot(X_i-c)\right)^2, \tag{3}$$

$$\left(\hat{\alpha}_{yr},\hat{\beta}_{yr}\right)=\arg\min_{\alpha_{yr},\beta_{yr}}\sum_{i:c\leq X_i\leq c+h}\left(Y_i-\alpha_{yr}-\beta_{yr}\cdot(X_i-c)\right)^2. \tag{4}$$

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:

$$\left(\hat{\alpha}_{wl},\hat{\beta}_{wl}\right)=\arg\min_{\alpha_{wl},\beta_{wl}}\sum_{i:c-h\leq X_i<c}\left(W_i-\alpha_{wl}-\beta_{wl}\cdot(X_i-c)\right)^2, \tag{5}$$

$$\left(\hat{\alpha}_{wr},\hat{\beta}_{wr}\right)=\arg\min_{\alpha_{wr},\beta_{wr}}\sum_{i:c\leq X_i\leq c+h}\left(Y_i-\alpha_{wr}-\beta_{wr}\cdot(X_i-c)\right)^2. \tag{6}$$

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:

$$\hat{\tau}_{\rm FRD}=\frac{\hat{\tau}_y}{\hat{\tau}_w}=\frac{\hat{\alpha}_{yr}-\hat{\alpha}_{yl}}{\hat{\alpha}_{wr}-\hat{\alpha}_{wl}}. \tag{7}$$

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 (See HTV). This equality still holds when we use local linear regression and include additional regressors. Define

$$V_i=\begin{pmatrix}1\\ 1\{X_i<c\}\cdot(X_i-c)\\ 1\{X_i\geq c\}\cdot(X_i-c)\end{pmatrix},\qquad\text{and}\quad\delta=\begin{pmatrix}\alpha_{yl}\\ \beta_{yl}\\ \beta_{yr}\end{pmatrix}. \tag{8}$$

Then we can write

$$Y_i=\delta' V_i+\tau\cdot W_i+\varepsilon_i. \tag{9}$$

Estimating τ based on the regression function (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 $\hat{\tau}_{\rm FRD}$ as given in (7).

6. Bandwidth Selection

An important issue in practice is the selection of the smoothing parameter, the binwidth h. Here we focus on cross-validation procedures rather than plug in methods which would require estimating derivatives nonparametrically. The specific methods discussed here are based on those developed by Ludwig and Miller (2005, 2007). Initially we focus on the SRD case, and in Section 6.2 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 the

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

where $\mu(x) = \mathbb{E}[Y_i|X_i = x]$. We estimate the two terms as

$$\widehat{\lim_{x\downarrow c}\mu(x)} = \hat{\alpha}_r(c), \qquad \text{and} \quad \widehat{\lim_{x\uparrow c}\mu(x)} = \hat{\alpha}_l(c),$$

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

$$\left(\hat{\alpha}_l(x), \hat{\beta}_l(x)\right) = \arg\min_{\alpha,\beta} \sum_{j|x-h<X_j<x} \left(Y_j - \alpha - \beta\cdot(X_j - x)\right)^2. \tag{10}$$

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

$$\left(\hat{\alpha}_r(x), \hat{\beta}_r(x)\right) = \arg\min_{\alpha,\beta} \sum_{j|x<X_j<x+h} \left(Y_j - \alpha - \beta\cdot(X_j - x)\right)^2. \tag{11}$$

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 $x = c$. However, we 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.

6.1 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 (10) and (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

$$\text{CV}_Y(h) = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \hat{\mu}(X_i)\right)^2, \tag{12}$$

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 under some conditions equal to $\mathbb{E}[\text{CV}_Y(h)] = C + \mathbb{E}[Q(X,h)] = C + \int Q(x,h) f_X(dx)$, for some constant that does not depend on h. 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 few observations in the tails of the distributions, minimizing the criterion in (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 δ quantile of the empirical distribution of X for the subsample with $X_i < c$, and let

$q_{X,\delta,r}$ be δ quantile of the empirical distribution of X for the subsample with $X_i \geq c$. Then we may wish to use the criterion

$$\mathrm{CV}_Y^{\delta}(h) = \frac{1}{N} \sum_{i: q_{X,\delta,l} \leq X_i \leq q_{X,1-\delta,r}} \left(Y_i - \hat{\mu}(X_i)\right)^2. \tag{13}$$

The modified cross-validation choice for the bandwidth is

$$h_{\mathrm{CV}}^{\delta,\mathrm{opt}} = \arg\min_h \mathrm{CV}_Y^{\delta}(h). \tag{14}$$

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, that is, $Q(c,h)$, but has the disadvantage of leading to a noisier estimate of $\mathbb{E}[\mathrm{CV}_Y^{\delta}(h)]$. 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.

6.2 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, 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 regressions functions by outcome we have here, and so in principle we have two optimal bandwidths, one based on minimizing $\mathrm{CV}_Y^{\delta}(h)$, and one based on minimizing $\mathrm{CV}_W^{\delta}(h)$, defined correspondingly. It is likely that the conditional expectation of the treatment is relatively flat compared to the conditional expectation of the outcome variable, suggesting one should use a larger binwidth for estimating the former.[5] Nevertheless, in practice it is appealing

[5] In the extreme case of the SRD design the conditional expectation of W given X is flat on both sides

to use the same binwidth for numerator and denominator. Since typically the size of the discontinuity is much more marked in the expected value of the treatment, one option is to use the optimal bandwidth based on the outcome discontinuity. Alternatively, to minimize bias, one may wish to use the smallest bandwidths selected by the cross validation criterion applied separately to the outcome and treatment regression:

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

where ${\rm CV}_Y^\delta(h)$ is as defined in (12), and ${\rm CV}_W^\delta(h)$ is defined similarly. Again a value of $\delta = 1/2$ is likely to lead to reasonable estimates in many settings.

7. Inference

We now discuss some asymptotic properties for the estimator for the FRD case given in (7) or its alternative representation in (9).[6] 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 expectations 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 give an explicit expression for the asymptotic variance, and 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. Both are robust to heteroskedasticity.

7.1 The Asymptotic Variance

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

of the threshold, and so the optimal bandwidth would be infinity. Therefore, in practice it is likely that the optimal bandwidth would be larger for estimating the jump in the conditional expectation of the treatment than in estimating the jump in the conditional expectation of the outcome.

[6] The results for the SRD design are a special case of these for the FRD design.

Define the four variances

$$\sigma^2_{Yl} = \lim_{x \uparrow c} \text{Var}(Y_i | X_i = x), \qquad \sigma^2_{Yr} = \lim_{x \downarrow c} \text{Var}(Y_i | X_i = x),$$

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

and the two covariances

$$C_{YWl} = \lim_{x \uparrow c} \text{Cov}(Y_i, W_i | X_i = x), \qquad C_{YWr} = \lim_{x \downarrow c} \text{Cov}(Y_i, W_i | X_i = x).$$

Note that because of the binary nature of W, it follows that $\sigma^2_{Wl} = \mu_{Wl} \cdot (1 - \mu_{Wl})$, where $\mu_{Wl} = \lim_{x \uparrow c} \Pr(W_i = 1 | X_i = x)$, 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

$$V_{\tau_y} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Yr} + \sigma^2_{Yl} \right). \tag{15}$$

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

$$V_{\tau_w} = \frac{4}{f_X(c)} \cdot \left(\sigma^2_{Wr} + \sigma^2_{Wl} \right) \tag{16}$$

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

$$C_{\tau_y, \tau_w} = \frac{4}{f_X(c)} \cdot \left(C_{YWr} + C_{YWl} \right). \tag{17}$$

Finally, the asymptotic distribution has the form

$$\sqrt{Nh} \cdot (\hat{\tau} - \tau) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\tau^2_w} \cdot V_{\tau_y} + \frac{\tau^2_y}{\tau^4_w} \cdot V_{\tau_w} - 2 \cdot \frac{\tau_y}{\tau^3_w} \cdot C_{\tau_y, \tau_w} \right). \tag{18}$$

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

7.2 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 (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

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

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

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

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 (15)-(17), and finally substitute these into the variance expression in (18).

7.3 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 (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 (8) as the exogenous covariates.

8. 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. The second concern is that of manipulation of the covariate value.

8.1 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. 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. See the second part of the Lee (2007) figure for an example.

8.2 Tests of Continuity of the Density

The second test is conceptually somewhat different, and unique to the RD setting. McCrary (2007) 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 boundary rather than the other, one might expect to see a discontinuity in this density at the discontinuity point.

8.3 Testing for Jumps at Non-discontinuity Points

Taking the subsample with $X_i < c$ we can test for a jump in the conditional mean of the outcome at the median of the forcing variable. To implement the test, use the same method for selecting the binwidth as before. 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.

8.4 RD Designs with Misspecification

Lee and Card (2007) study the case where the forcing variable variable X is discrete. In practice this is of course always true. 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. 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.

8.5 Sensitivity to the Choice of Bandwidth

One should 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.

8.6 Comparisons to Estimates Based on Unconfoundedness in the FRD Design

If we have an FRD design, we can also consider estimates based on unconfoundedness. 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.

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.

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

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.

Cook, T., (2007), "Waiting for Life to Arrive": A History of the Regression-Discontinuity Design in Psychology, Statistics, and Economics, forthcoming, *Journal of Econometrics*.

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, (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).

Imbens, G., and J. Angrist (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, Vol. 61, No. 2, 467-476.

Imbens, G., and T. Lemieux, (2008) "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*.

Lee, D.S. and D. Card, (2007), Regression Discontinuity Inference with Specification Error, forthcoming, *Journal of Econometrics*.

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, forthcoming, *Journal of Econometrics*.

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

McCrary, J., (2007), Testing for Manipulation of the Running Variable in the Regression Discontinuity Design, forthcoming, *Journal of Econometrics*.

McEwan, P., and J. Shapiro, (2007),The Benefits of Delayed Primary School Enrollment: Discontinuity Estimates using exact Birth Dates," Unpublished manuscript.

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.

Shadish, W., T. Campbell and D. Cook, (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference* (Houghton Mifflin, Boston).

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.