

# Optimal bandwidth choice for the regression discontinuity estimator

---

**Guido Imbens**  
**Karthik Kalyanaraman**

The Institute for Fiscal Studies  
Department of Economics, UCL

**cemmap** working paper CWP05/10

# Optimal Bandwidth Choice for the Regression Discontinuity Estimator\*

Guido Imbens<sup>†</sup>  
and Karthik Kalyanaraman<sup>‡</sup>

First Draft: June 2008  
This Draft: January 2010

## Abstract

We investigate the problem of optimal choice of the smoothing parameter (bandwidth) for the regression discontinuity estimator. We focus on estimation by local linear regression, which was shown to be rate optimal (Porter, 2003). We derive the optimal bandwidth. This optimal bandwidth depends on unknown functionals of the distribution of the data and we propose specific, consistent, estimators for these functionals to obtain a fully data-driven bandwidth choice that has the “asymptotic no-regret” property. We illustrate our proposed bandwidth, and the sensitivity to the choices made in this bandwidth proposal, using a data set previously analyzed by Lee (2008), as well as a small simulation study based on the Lee data set. The simulations suggest that the proposed rule performs well.

## JEL Classification:

**Keywords:** *Optimal Bandwidth Selection, Local Linear Regression, Regression Discontinuity Designs*

---

\*Financial support for this research was generously provided through NSF grants 0452590 and 0820361. We are grateful to David Lee for making his data available, and to Tom Cook, Tom Lemieux, Doug Miller, a co-editor and two referees for comments.

<sup>†</sup>Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge, MA 02138, and NBER. Electronic correspondence: [imbens@harvard.edu](mailto:imbens@harvard.edu), <http://www.economics.harvard.edu/faculty/imbens/imbens.html>.

<sup>‡</sup>University College London, Gower Street, London. Electronic correspondence: [kalyana@fas.harvard.edu](mailto:kalyana@fas.harvard.edu).

## 1 Introduction

Regression discontinuity (RD) designs for evaluating causal effects of interventions, where assignment is determined at least partly by the value of an observed covariate lying on either side of a threshold, were introduced by Thistlewaite and Campbell (1960). See Cook (2008) for a historical perspective. A recent surge of applications in economics includes studies of the impact of financial aid offers on college acceptance (Van Der Klaauw, 2002), school quality on housing values (Black, 1999), class size on student achievement (Angrist and Lavy, 1999), air quality on health outcomes (Chay and Greenstone, 2005), incumbency on re-election (Lee, 2008), and many others. Recent important theoretical work has dealt with identification issues (Hahn, Todd, and Van Der Klaauw, 2001, HTV from hereon), optimal estimation (Porter, 2003), tests for validity of the design (McCrary, 2008), quantile effects (Frandsen, 2008; Frölich and Melly, 2008), and the inclusion of covariates (Frölich, 2007). General surveys include Lee and Lemieux (2009), Van Der Klaauw (2008), and Imbens and Lemieux (2008).

In RD settings analyses typically focus on the average effect of the treatment for units with values of the forcing variable close to the threshold, using kernel, local linear, or global polynomial series estimators. Fan and Gijbels (1992) and Porter (2003) show that local linear estimators are rate optimal and have attractive bias properties. A key decision in implementing these methods is the choice of bandwidth. In current practice researchers use a variety of *ad hoc* approaches for bandwidth choice, such as standard plug-in and crossvalidation methods. These are sometimes based on objective functions which take into account the performance of the estimator of the regression function over the entire support, and do not yield optimal bandwidths here. There are no bandwidth choices in the literature (e.g., Härdle 1992, Fan and Gijbels, 1992, Wand and Jones, 1994) for the case where the estimand is the difference in two regression functions at boundary points. In this paper we propose a practical, rule-of-thumb bandwidth choice tailored to the RD setting with some optimality properties. The two contributions of this paper are (i), the derivation of the optimal bandwidth for this setting, taking account of the special features of the RD setting, and (ii), a fully data-dependent method for choosing the bandwidth that is asymptotically optimal in the sense of having the “asymptotic no-regret” property.<sup>1</sup> Although optimal in large samples, the proposed algorithm involves initial bandwidth choices. We analyze the sensitivity of the results to these choices. We illustrate the algorithm using a data set previously analyzed by Lee (2008), and compare our procedure to

---

<sup>1</sup>Matlab and Stata software for implementing this bandwidth rule is available on the website <http://www.economics.harvard.edu/faculty/imbens/imbens.html>.

others proposed in the literature, including the crossvalidation procedure proposed by Ludwig and Miller (2007) and a bandwidth choice proposed by DesJardins and McCall (2008) that uses a different criterion. Simulations indicate that the proposed algorithm works well in realistic settings.

## 2 Basic model

In the basic RD setting, researchers are interested in the causal effect of a binary treatment. In the setting we consider we have a sample of  $N$  units, drawn randomly from a large population. For unit  $i$ ,  $i = 1, \dots, N$ , the variable  $Y_i(1)$  denotes the potential outcome for unit  $i$  given treatment, and  $Y_i(0)$  the potential outcome without treatment. For unit  $i$  we observe the treatment received,  $W_i$ , equal to one if unit  $i$  was exposed to the treatment and 0 otherwise, and the outcome corresponding to the treatment received:

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

We also observe for each unit a scalar covariate, called the forcing variable, denoted by  $X_i$ . Here we focus on the case with a scalar forcing variable. In Section 5 we discuss the case with additional covariates. Define

$$m(x) = \mathbb{E}[Y_i | X_i = x],$$

to be the conditional expectation of the outcome given the forcing variable. The idea behind the Sharp Regression Discontinuity (SRD) design is that the treatment  $W_i$  is determined solely by the value of the forcing variable  $X_i$  being on either side of a fixed, known threshold  $c$ , or:

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

In Section 5 we extend the SRD setup to the case with additional covariates and to the Fuzzy Regression Discontinuity (FRD) design, where the probability of receiving the treatment jumps discontinuously at the threshold for the forcing variable, but not necessarily from zero to one.

In the SRD design the focus is on average effect of the treatment for units with covariate values equal to the threshold:

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

Now suppose that the conditional distribution functions  $F_{Y(0)|X}(y|x)$  and  $F_{Y(1)|X}(y|x)$  are continuous in  $x$  for all  $y$ , and that the conditional first moments  $\mathbb{E}[Y_i(1) | X_i = x]$  and  $\mathbb{E}[Y_i(0) | X_i = x]$

exist, and are continuous at  $x = c$ . Then

$$\tau_{\text{SRD}} = \mu_+ - \mu_-, \quad \text{where } \mu_+ = \lim_{x \downarrow c} m(x), \quad \text{and } \mu_- = \lim_{x \uparrow c} m(x).$$

Thus, the estimand is the difference of two regression functions evaluated at boundary points.

We focus on estimating  $\tau_{\text{SRD}}$  by local linear regressions on either side of the threshold. Local nonparametric methods are attractive in this setting because of the need to estimate regression functions consistently at a point. Furthermore, in the RD setting local linear regression estimators are preferred to the standard Nadaraya-Watson kernel estimator, because local linear methods have been shown to have attractive bias properties in estimating regression functions at the boundary (Fan and Gijbels, 1992), and enjoy rate optimality (Porter, 2003). To be explicit, we estimate the regression function  $m(\cdot)$  at  $x$  as

$$\hat{m}_h(x) = \begin{cases} \hat{\alpha}_-(x) & \text{if } x < c, \\ \hat{\alpha}_+(x) & \text{if } x \geq c. \end{cases} \quad (2.1)$$

where,

$$(\hat{\alpha}_-(x), \hat{\beta}_-(x)) = \arg \min_{\alpha, \beta} \sum_{i=1}^N \mathbf{1}_{X_i < x} \cdot (Y_i - \alpha - \beta(X_i - x))^2 \cdot K\left(\frac{X_i - x}{h}\right),$$

and

$$(\hat{\alpha}_+(x), \hat{\beta}_+(x)) = \arg \min_{\alpha, \beta} \sum_{i=1}^N \mathbf{1}_{X_i > x} \cdot (Y_i - \alpha - \beta(X_i - x))^2 \cdot K\left(\frac{X_i - x}{h}\right),$$

Then

$$\hat{\tau}_{\text{SRD}} = \hat{\mu}_+ - \hat{\mu}_-,$$

where

$$\hat{\mu}_- = \lim_{x \uparrow c} \hat{m}_h(x) = \hat{\alpha}_-(c) \quad \text{and} \quad \hat{\mu}_+ = \lim_{x \downarrow c} \hat{m}_h(x) = \hat{\alpha}_+(c).$$

### 3 Error Criterion and Infeasible Optimal Bandwidth Choice

In this section we discuss the objective function, and derive the optimal bandwidth under that criterion.

#### 3.1 Error Criteria

The primary question studied in this paper concerns the optimal choice of the bandwidth  $h$ . In the current empirical literature researchers often choose the bandwidth by either crossvalidation

or *ad hoc* methods. See Härdle (1992), Fan and Gijbels (1992), and Wand and Jones (1994) for textbook discussion of crossvalidation and related methods, and see Lee and Lemieux (2009) for a comprehensive discussion of current practice in RD settings. Conventional crossvalidation yields a bandwidth that is optimal for fitting a curve over the entire support of the data.<sup>2</sup> In other words, it attempts to choose the bandwidth to minimize an approximation to the mean integrated squared error criterion (MISE),

$$\text{MISE}(h) = \mathbb{E} \left[ \int_x (\hat{m}_h(x) - m(x))^2 f(x) dx \right],$$

here  $f(x)$  is the density of the forcing variable. This criterion is not directly relevant for the problem at hand: we wish to choose a bandwidth that is optimal for estimating  $\tau_{\text{SRD}}$ . This estimand has a number two special features. First, it depends on  $m(x)$  only through two values, and specifically their difference. Second, both these values are boundary values.

Our proposed criterion is based on the expectation of the asymptotic expansion, around  $h = 0$ , of the squared error  $(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}})^2$ . First, define the expected squared error:

$$\text{MSE}(h) = \mathbb{E} \left[ \left( \hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} \right)^2 \right] = \mathbb{E} \left[ \left( (\hat{\mu}_+ - \mu_+) - (\hat{\mu}_- - \mu_-) \right)^2 \right]. \quad (3.2)$$

and let  $h^*$  be the optimal bandwidth that minimizes this criterion:

$$h^* = \arg \min \text{MSE}(h). \quad (3.3)$$

This criterion is difficult to work with directly. The problem is that in many cases, as the sample size  $N$  goes to infinity, the optimal bandwidth  $h^*$  will not converge to zero, because biases in different parts of the regression function may be offsetting.<sup>3</sup> In such cases the optimal bandwidth is very sensitive to the actual distribution and regression function, and it is difficult to see how one could exploit such knife-edge cases.

A second comment concerns our focus on a single bandwidth. Because the estimand,  $\tau_{\text{SRD}}$ , is a function of the regression function at two points, an alternative would be to allow for a different bandwidth for these two points,  $h_-$  for estimating  $\mu_-$ , and  $h_+$  for estimating  $\mu_+$  and and focus on an objective function that is an approximation to

$$\text{MSE}(h_-, h_+) = \mathbb{E} \left[ \left( (\hat{\mu}_+(h_+) - \mu_+) - (\hat{\mu}_-(h_-) - \mu_-) \right)^2 \right]. \quad (3.4)$$

---

<sup>2</sup>See Ludwig and Miller (2005) and Lee and Lemieux (2009) for a discussion of crossvalidation methods designed more specifically for the RD setting. These methods are discussed in more detail in Section 4.5.

<sup>3</sup>To be explicit, consider a simple example where we are interested in estimating a regression function  $g(x)$  at a single point, say  $g(0)$ . Suppose the covariate  $X$  has a uniform distribution on  $[0, 1]$ . Suppose the regression function is  $g(x) = (x - 1/4)^2 - 1/16$ . With a uniform kernel the estimator for  $g(0)$  is, for a bandwidth  $h$ , equal to  $\sum_{i: X_i < h} X_i / \sum_{i: X_i < h} 1$ . As a function of the bandwidth  $h$  the bias is equal to  $h^2/3 - h/4$ , conditional on  $\sum_{i: X_i < h} 1$ . Thus, the bias is zero at  $h = 3/4$ , and if we minimize the expected squared error, the optimal bandwidth will converge to  $3/4$ .

Doing so would raise an important issue. We focus on minimizing mean squared error, equal to variance plus bias squared. Suppose that for both estimators the bias,  $\mathbb{E}[\hat{\mu}_-(h_-)]$  and  $\mathbb{E}[\hat{\mu}_+(h_+)]$  are strictly increasing (or both strictly decreasing) functions of the bandwidth. Then there is a function  $h_+(h_-)$  such that the biases cancel out:  $\mathbb{E}[\hat{\mu}_-(h_-)] - \mathbb{E}[\hat{\mu}_+(h_+(h_-))] = 0$ . Hence we can minimize the mean-squared-error by letting  $h_-$  get large (the variance is generally a decreasing function of the bandwidth), and choosing  $h_+ = h_+(h_-)$ . Even if this does not hold exactly, the point is that a problem may arise that even for large bandwidths the difference in bias may be close to zero. In practice it is unlikely that one can effectively exploit the cancellation of biases for large bandwidths. However, it would make it difficult to construct practical bandwidth algorithms. But, to avoid this problem, we focus in this discussion on a single bandwidth choice. An alternative would be to change the criterion and add the mean-squared-errors on the left and the right, e.g.,  $\mathbb{E}[(\hat{\mu}_+ - \mu_+)^2 + (\hat{\mu}_- - \mu_-)^2]$ , rather than focusing on the mean-squared-error of the difference,  $\mathbb{E}[(\hat{\mu}_+ - \mu_+) - (\hat{\mu}_- - \mu_-)]^2$ .

### 3.2 An Asymptotic Expansion of the Expected Error

The next step is to derive an asymptotic expansion of (3.2). First we state the key assumptions. Not all of these will be used immediately, but for convenience we state them all here.

**Assumption 3.1:**  $(Y_i, X_i)$ , for  $i = 1, \dots, N$ , are independent and identically distributed.

**Assumption 3.2:** The marginal distribution of the forcing variable  $X_i$ , denoted  $f(\cdot)$ , is continuous and bounded away from zero at the discontinuity,  $c$ .

**Assumption 3.3:** The conditional mean  $m(x) = \mathbb{E}[Y_i | X_i = x]$  has at least three continuous derivatives in an open neighborhood of  $X = c$ . The right and left limits of the  $k^{\text{th}}$  derivative of  $m(x)$  at the threshold  $c$  are denoted  $m_+^{(k)}(c)$  and  $m_-^{(k)}(c)$ .

**Assumption 3.4:** The kernel  $K(\cdot)$  is nonnegative, bounded, differs from zero on a compact interval  $[0, a]$ , and is continuous on  $(0, a)$ .

**Assumption 3.5:** The conditional variance function  $\sigma^2(x) = \text{Var}(Y_i | X_i = x)$  is bounded in an open neighborhood of  $X = c$ , and right and left continuous at  $c$ . The right and left limit are denoted by  $\sigma_+^2(c)$  and  $\sigma_-^2(c)$  respectively.

**Assumption 3.6:** The second derivatives at the right and left,  $m_+^{(2)}(x)$  and  $m_-^{(2)}(x)$ , differ at the threshold:  $m_+^{(2)}(c) \neq m_-^{(2)}(c)$ .

Now define the Asymptotic Mean Squared Error (AMSE) as a function of the bandwidth:

$$\text{AMSE}(h) = C_1 \cdot h^4 \cdot \left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2 + \frac{C_2}{N \cdot h} \cdot \left(\frac{\sigma_+^2(c)}{f(c)} + \frac{\sigma_-^2(c)}{f(c)}\right). \quad (3.5)$$

The constants  $C_1$  and  $C_2$  in this approximation are functions of the kernel:

$$C_1 = \frac{1}{4} \left(\frac{\nu_2^2 - \nu_1 \nu_3}{\nu_2 \nu_0 - \nu_1^2}\right)^2, \quad \text{and} \quad C_2 = \frac{\nu_2^2 \pi_0 - 2\nu_1 \nu_2 \pi_1 + \nu_1^2 \pi_2}{(\nu_2 \nu_0 - \nu_1^2)^2}, \quad (3.6)$$

where

$$\nu_j = \int_0^\infty w^j K(u) du, \quad \text{and} \quad \pi_j = \int_0^\infty w^j K^2(u) du.$$

The first term in (3.5) corresponds to the square of the bias, and the second term corresponds to the variance. This expression clarifies the role that Assumption 3.6 will play. If the left and right limits of the second derivative are equal, then the leading term in the expansion of the square of the bias is not of the order  $h^4$ . Instead the leading bias term would be of lower order. It is difficult to exploit the improved convergence rate that would result from this in practice, because it would be difficult to establish sufficiently fast that this difference is indeed zero, and so we focus on optimality results given Assumption 3.6. Note however, that even if the second derivatives are identical, our estimator for  $\tau_{\text{SRD}}$  will be consistent.

An alternative approach would be to focus on a bandwidth choice that is optimal if the second derivatives from the left and right are identical. It is possible to construct such a bandwidth choice, and still maintain consistency of the resulting estimator for  $\tau_{\text{SRD}}$  irrespective of the difference in second derivatives. However, such a bandwidth choice would generally not be optimal if the difference in second derivatives is nonzero. Thus there is a choice between a bandwidth choice that is optimal under  $m_+^{(2)}(c) \neq m_-^{(2)}(c)$  and a bandwidth choice that is optimal under  $m_+^{(2)}(c) = m_-^{(2)}(c)$ . In the current paper we choose to focus on the first case. The reason is that if the second derivatives are in fact equal, the leading term in the bias vanishes. Ignoring the leading term when in fact it is present appears in our view to be a bigger concern than taking it into account when it is not.

**Lemma 3.1:** (MEAN SQUARED ERROR APPROXIMATION AND OPTIMAL BANDWIDTH)

(i) Suppose Assumptions 3.1-3.5 hold. Then

$$\text{MSE}(h) = \text{AMSE}(h) + o\left(h^4 + \frac{1}{N \cdot h}\right).$$

(ii) Suppose Assumptions 3.1-3.6 hold. Then

$$h_{\text{opt}} = \arg \min_h \text{AMSE}(h) = C_K \cdot \left(\frac{\sigma_+^2(c) + \sigma_-^2(c)}{f(c) \cdot \left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2}\right)^{1/5} \cdot N^{-1/5}, \quad (3.7)$$

where  $C_K = (C_2/(4 \cdot C_1))^{1/5}$ , indexed by the kernel  $K(\cdot)$ .

For the edge kernel, with  $K(u) = \mathbf{1}_{|u| \leq 1}(1 - |u|)$ , shown by Cheng, Fan and Marron (1997) to have AMSE-minimizing properties for boundary estimation problems, the constant is  $C_K \approx 3.4375$ .

## 4 Feasible Optimal Bandwidth Choice

In this section we discuss the proposed bandwidth, provide a full data-dependent estimator for the bandwidth, and discuss its properties.

### 4.1 Proposed bandwidth

A natural choice for the estimator for the optimal bandwidth estimator is to replace the six unknown quantities in the expression for the optimal bandwidth  $h_{\text{opt}}$ , given in (4.16) by non-parametric estimators, leading to

$$\tilde{h}_{\text{opt}} = C_K \cdot \left( \frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c) \cdot (\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c))^2} \right)^{1/5} \cdot N^{-1/5}. \quad (4.8)$$

We make one modification to this approach, motivated partly by the desire to reduce the variance of the estimated bandwidth  $\hat{h}_{\text{opt}}$ , and partly by considerations regarding the structure of the problem. More precisely, the concern is that the precision with which we estimate the second derivatives  $m_+^{(2)}(c)$  and  $m_-^{(2)}(c)$  may be so low, that the estimated optimal bandwidth  $\tilde{h}_{\text{opt}}$  will occasionally be very large, even when the data are consistent with a substantial degree of curvature. To address this problem we add a regularization term to the denominator in (4.8). This regularization term will be chosen carefully to decrease with the sample size, therefore not compromising asymptotic optimality. Including this regularization term guards against unrealistically large bandwidth choices when the curvature of the regression function is imprecisely estimated.

We use as the regularization term the approximate variance of the estimated curvature. This allows the regularization term to be invariant to the scale of the data. To be explicit, we estimate the second derivative  $m_+^{(2)}(c)$  by fitting to the observations with  $X_i \in [c, c + h]$  a quadratic function. The bandwidth  $h$  here may be different from the bandwidth  $\hat{h}_{\text{opt}}$  used in the estimation of  $\tau_{\text{SRD}}$ , and its choice will be discussed in Section 4.2. Let  $N_{h,+}$  be the number of units with covariate values in this interval. We assume homoskedasticity with error variance

$\sigma^2(c)$  in this interval. Let

$$\hat{\mu}_{j,h,+} = \frac{1}{N_{h,+}} \sum_{c \leq X_i \leq c+h} (X_i - \bar{X})^j, \quad \text{where } \bar{X} = \frac{1}{N_{h,+}} \sum_{c \leq X_i \leq c+h} X_i,$$

be the  $j$ -th (centered) moment of the  $X_i$  in this interval to the right of the threshold. We can derive the following explicit formula for the conditional variance of the curvature (viz. twice the coefficient on the quadratic term), denoted by  $r_+$ , in terms of these moments:

$$r_+ = \frac{4}{N_{h,+}} \left( \frac{\sigma_+^2(c)}{\hat{\mu}_{4,h,+} - (\hat{\mu}_{2,h,+})^2 - (\hat{\mu}_{3,h,+})^2 / \hat{\mu}_{2,h,+}} \right)$$

However, to avoid estimating fourth moments, we approximate this expression exploiting the fact that for small  $h$ , the distribution of the forcing variable, normalized to have unit variance, can be approximated by a uniform distribution on  $[c, c+h]$ , so that  $\hat{\mu}_{2,h,+} \approx h^2/12$ ,  $\hat{\mu}_{3,h,+} \approx 0$ , and  $\hat{\mu}_{4,h,+} \approx h^4/60$ . After substituting  $\hat{\sigma}_-^2(c)$  for  $\sigma_-^2(c)$  and  $\hat{\sigma}_+^2(c)$  for  $\sigma_+^2(c)$  this leads to

$$\hat{r}_+ = \frac{720 \cdot \hat{\sigma}_+^2(c)}{N_{h,+} \cdot h^4}, \quad \text{and similarly } \hat{r}_- = \frac{720 \cdot \hat{\sigma}_-^2(c)}{N_{h,-} \cdot h^4}.$$

The proposed bandwidth is now obtained by adding the regularization terms to the curvatures in the bias term of MSE expansion:

$$\hat{h}_{\text{opt}} = C_K \cdot \left( \frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c) \left( \left( \hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + (\hat{r}_+ + \hat{r}_-) \right)} \right)^{1/5} \cdot N^{-1/5}, \quad (4.9)$$

To operationalize this proposed bandwidth, we need specific estimators  $\hat{f}(c)$ ,  $\hat{\sigma}_-^2(c)$ ,  $\hat{\sigma}_+^2(c)$ ,  $\hat{m}_-^{(2)}(c)$ , and  $\hat{m}_+^{(2)}(c)$ . We provide a specific proposal for this in the next section.

## 4.2 Algorithm for bandwidth selection

The reference bandwidth  $\hat{h}_{\text{opt}}$  is a function of the outcome variable  $\mathbf{Y} = (Y_1, \dots, Y_N)$ , the forcing variable  $\mathbf{X} = (X_1, \dots, X_N)$  and the chosen kernel; i.e.  $\hat{h}_{\text{opt}} = h(\mathbf{Y}, \mathbf{X})$ . We give below a general algorithm for a specific implementation. In practice we recommend using the edge optimal kernels, where  $K(u) = 1_{|u| \leq 1} \cdot (1 - |u|)$ , although the algorithm is easily modified for other kernels by changing the kernel-specific constant  $C_K$ .

To calculate the bandwidth we need estimators for the density at the threshold,  $f(c)$ , the conditional variances at the threshold,  $\sigma_-^2(c)$  and  $\sigma_+^2(c)$ , and the limits of the second derivatives at the threshold from the right and the left,  $m_+^{(2)}(c)$ ,  $m_-^{(2)}(c)$ . (The other components of (4.9),  $\hat{r}_-$  and  $\hat{r}_+$  are functions of these four components.) The first two functionals are calculated

in step 1, the second two in step 2. Step 3 puts these together with the appropriate kernel constant  $C_K$  to produce the reference bandwidth.

Step 1: Estimation of density  $f(c)$  and conditional variances  $\sigma_-^2(c)$  and  $\sigma_+^2(c)$

First calculate the sample variance of the forcing variable,  $S_X^2 = \sum (X_i - \bar{X})^2 / (N - 1)$ . We now use the Silverman rule to get a pilot bandwidth for calculating the density and variance at  $c$ . The standard Silverman rule of  $h = 1.06 \cdot S_X \cdot N^{-1/5}$  is based on a normal kernel and a normal reference density. We modify this for the uniform kernel on  $[-1, 1]$  and the normal reference density, and calculate the pilot bandwidth  $h_1$  as:

$$h_1 = 1.84 \cdot S_X \cdot N^{-1/5}.$$

Calculate the number of units on either side of the threshold, and the average outcomes on either side as

$$N_{h_1,-} = \sum_{i=1}^N \mathbf{1}_{c-h_1 \leq X_i < c}, \quad N_{h_1,+} = \sum_{i=1}^N \mathbf{1}_{c \leq X_i \leq c+h_1},$$

$$\bar{Y}_{h_1,-} = \frac{1}{N_{h_1,-}} \sum_{i:c-h_1 \leq X_i < c} Y_i, \quad \text{and} \quad \bar{Y}_{h_1,+} = \frac{1}{N_{h_1,+}} \sum_{i:c \leq X_i \leq c+h_1} Y_i.$$

Now estimate the density of  $X_i$  at  $c$  as

$$\hat{f}_X(c) = \frac{N_{h_1,-} + N_{h_1,+}}{2 \cdot N \cdot h_1}, \tag{4.10}$$

and estimate the limit of the conditional variances of  $Y_i$  given  $X_i = x$ , at  $x = c$ , from the left and the right, as

$$\hat{\sigma}_-^2(c) = \frac{1}{N_{h_1,-} - 1} \sum_{i:c-h_1 \leq X_i < c} (Y_i - \bar{Y}_{h_1,-})^2, \tag{4.11}$$

and

$$\hat{\sigma}_+^2(c) = \frac{1}{N_{h_1,+} - 1} \sum_{i:c \leq X_i \leq c+h_1} (Y_i - \bar{Y}_{h_1,+})^2. \tag{4.12}$$

The main property we will need for these estimators is that they are consistent for the density and the conditional variance respectively. They need not be efficient. For consistency of the density and conditional variance estimators Assumptions 3.2 and 3.5 are sufficient, given that the bandwidth goes to zero at rate  $N^{-1/5}$ .

Step 2: Estimation of second derivatives  $\hat{m}_+^{(2)}(c)$  and  $\hat{m}_-^{(2)}(c)$

First we need a pilot bandwidth  $h_{2,+}$ . We base this on a simple, not necessarily consistent, estimator of the third derivative of  $m(\cdot)$  at  $c$ . Let  $N_-$  and  $N_+$  be the number of observations to the left and right of the threshold, respectively. Now fit a third order polynomial to the data, including an indicator for  $X_i \geq 0$ . Thus, estimate the regression function

$$Y_i = \gamma_0 + \gamma_1 \cdot 1_{X_i \geq c} + \gamma_2 \cdot (X_i - c) + \gamma_3 \cdot (X_i - c)^2 + \gamma_4 \cdot (X_i - c)^3 + \varepsilon_i, \quad (4.13)$$

and estimate  $m^{(3)}(c)$  as  $\hat{m}^{(3)}(c) = 6 \cdot \hat{\gamma}_4$ . This will be our estimate of the third derivative of the regression function. Note that  $\hat{m}^{(3)}(c)$  is in general not a consistent estimate of  $m^{(3)}(c)$  but will converge to some constant at a parametric rate. Let  $m_3 = 6 \cdot \text{plim}(\hat{\gamma}_4)$  denote this constant. However we do not need a consistent estimate of the third derivative at  $c$  here to achieve what we ultimately need: a consistent estimate of the constant in the reference bandwidth. Calculate  $h_{2,+}$ , using the  $\hat{\sigma}_-^2(c)$ ,  $\hat{\sigma}_+^2(c)$  and  $\hat{f}(c)$  from Step 1, as

$$h_{2,+} = 3.56 \left( \frac{\hat{\sigma}_+^2(c)}{\hat{f}(c) (\hat{m}^{(3)}(c))^2} \right)^{1/7} N_+^{-1/7}, \quad (4.14)$$

and

$$h_{2,-} = 3.56 \left( \frac{\hat{\sigma}_-^2(c)}{\hat{f}(c) (\hat{m}^{(3)}(c))^2} \right)^{1/7} N_-^{-1/7}.$$

These bandwidths,  $h_{2,-}$  and  $h_{2,+}$ , are estimates of the optimal bandwidth for calculation of the second derivative at the boundary using a local quadratic. See the Appendix for details.

Given this pilot bandwidth  $h_{2,+}$ , we estimate the curvature  $m^{(2)}(c)$  by a local quadratic fit. To be precise, temporarily discard the observations other than the  $N_{2,+}$  observations with  $c \leq X_i \leq c + h_{2,+}$ . Label the new data  $\hat{\mathbf{Y}}_+ = (Y_1, \dots, Y_{N_{2,+}})$  and  $\hat{\mathbf{X}}_+ = (X_1, \dots, X_{N_{2,+}})$  each of length  $N_{2,+}$ . Fit a quadratic to the new data. I.e. let  $\mathbf{T} = [\iota \quad \mathbf{T}_1 \quad \mathbf{T}_2]$  where  $\iota$  is a column vector of ones, and  $\mathbf{T}'_j = ((X_1 - c)^j, \dots, (X_{N_{2,+}} - c)^j)$ , for  $j = 1, 2$ . Estimate the three dimensional regression coefficient vector,  $\hat{\lambda} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\hat{\mathbf{Y}}$ . Calculate the curvature as  $\hat{m}_+^{(2)}(c) = 2 \cdot \hat{\lambda}_3$ . This is a consistent estimate of  $m_+^{(2)}(c)$ . To estimate  $m_-^{(2)}(c)$  follow the same procedure using the data with  $c - h_{2,-} \leq X_i < c$ .

### Step 3: Calculation of Regularization Terms $\hat{r}_-$ and $\hat{r}_+$ , and Calculation of $\hat{h}_{\text{opt}}$

Given the previous steps, the regularization terms are calculated as

$$\hat{r}_+ = \frac{720 \cdot \hat{\sigma}_+^2(c)}{N_{2,+} \cdot h_{2,+}^4}, \quad \text{and} \quad \hat{r}_- = \frac{720 \cdot \hat{\sigma}_-^2(c)}{N_{2,-} \cdot h_{2,-}^4}. \quad (4.15)$$

We now have all the pieces to calculate the proposed bandwidth:

$$\hat{h}_{\text{opt}} = C_K \cdot \left( \frac{\hat{\sigma}_-^2(c) + \hat{\sigma}_+^2(c)}{\hat{f}(c) \cdot \left( \left( \hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + (\hat{r}_+ + \hat{r}_-) \right)} \right)^{1/5} \cdot N^{-1/5}. \quad (4.16)$$

where  $C_K$  is, as in Lemma 3.1, a constant that depends on the kernel used. For the edge kernel, with  $K(u) = (1 - |u|) \cdot \mathbf{1}_{|u| \leq 1}$ , the constant is  $C_K \approx 3.4375$ .

Given  $\hat{h}_{\text{opt}}$ , we estimate  $\tau_{\text{SRD}}$  as

$$\hat{\tau}_{\text{SRD}} = \lim_{x \downarrow c} \hat{m}_{\hat{h}_{\text{opt}}}(x) - \lim_{x \uparrow c} \hat{m}_{\hat{h}_{\text{opt}}}(x),$$

where  $\hat{m}_h(x)$  is the local linear regression estimator as defined in (2.1).

### 4.3 Properties of algorithm

For this algorithm we establish certain optimality properties. First, the resulting RD estimator is consistent at the best rate for nonparametric regression functions at a point (Stone, 1982). Second, as the sample size increases, the estimated constant term in the reference bandwidth converges to the best constant. Third, we also have an ‘‘asymptotic no-regret’’ or Li (1987) type consistency result for the mean squared error and consistency at the optimal rate for the RD estimate.

**Theorem 4.1:** (PROPERTIES OF  $\hat{h}_{\text{opt}}$ )

*Suppose Assumptions 3.1-3.5 hold. Then:*

(i)

$$\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p \left( N^{-12/35} \right), \quad (4.17)$$

(ii) *Suppose also Assumption 3.6 holds. Then:*

$$\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p \left( N^{-2/5} \right), \quad (4.18)$$

(iii)

$$\frac{\hat{h}_{\text{opt}} - h_{\text{opt}}}{h_{\text{opt}}} = o_p(1), \quad (4.19)$$

and (iv):

$$\frac{\text{MSE}(\hat{h}_{\text{opt}}) - \text{MSE}(h_{\text{opt}})}{\text{MSE}(h_{\text{opt}})} = o(1). \quad (4.20)$$

Note that if Assumption 3.6 holds, the convergence rate is  $N^{-14/35}$ , and if it fails, the convergence rate for  $\hat{\tau}_{\text{SRD}}$  is slower, namely  $N^{-12/35}$ . This may be somewhat counterintuitive, because failure of Assumption (3.6) implies that the leading term of the bias vanishes, which, one might expect, would improve convergence. Here is the explanation. If Assumption 3.6 fails, the leading term in the bias vanishes, and the square of the bias becomes of order  $O(h^6)$ . Because the variance remains of order  $O((Nh)^{-1})$ , the optimal rate for the bandwidth, based on balancing the bias-squared and the variance, becomes  $N^{-1/7}$ . As a result the optimal rate for the MSE becomes  $N^{-6/7}$  and thus the optimal rate for  $\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}}$  becomes  $N^{-3/7}$ . This is better than the convergence rate of  $N^{-2/5}$  that we have when Assumption 3.6 holds. The reason this  $N^{-3/7}$  convergence rate does not show up in the theorem is that the proposed optimal bandwidth does not adapt to the vanishing of the difference in second derivatives. If Assumption 3.6 fails, the proposed bandwidth goes to zero as  $N^{-4/35}$  (instead of the optimal rate  $N^{-1/7}$ ), and so the MSE becomes  $N^{-24/35}$ , leading to  $\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p(N^{-12/35})$ , slower than the optimal rate of  $N^{-3/7}$ , and even slower than the rate we achieve when Assumption 3.6 holds (namely,  $N^{-2/5}$ ). One could modify the regularization term to take account of this, but in practice it is unlikely to make a difference.

#### 4.4 DeJardins-McCall Bandwidth Selection

DesJardins and McCall (2008) use an alternative method for choosing the bandwidth. They focus separately on the limits of the regression function to the left and the right, rather than on the difference in the limits. This leads them to minimize the sum of the squared differences between  $\hat{\mu}_-$  and  $\mu_-$ , and between  $\hat{\mu}_+$  and  $\mu_+$ :

$$\mathbb{E}[(\hat{\mu}_+ - \mu_+)^2 + (\hat{\mu}_- - \mu_-)^2],$$

instead of our criterion,

$$\mathbb{E}[(\hat{\mu}_+ - \mu_+) - (\hat{\mu}_- - \mu_-)]^2.$$

The single optimal bandwidth based on this criterion is<sup>4</sup>

$$h_{DM} = C_K \cdot \left( \frac{\sigma_+^2(c) + \sigma_-^2(c)}{f(c) \cdot (m_+^{(2)}(c)^2 + m_-^{(2)}(c)^2)} \right)^{1/5} \cdot N^{-1/5}.$$

---

<sup>4</sup>DesJardins and McCall actually allow for different bandwidths on the left and the right, and also use a Epanechnikov kernel instead of the optimal edge kernel. In the calculations below we use the edge kernel to make the results more comparable.

This will lead to a smaller bandwidth than our proposed bandwidth choice if the second derivatives are of the same sign. We include the DesJardins-McCall bandwidth in our bandwidth comparisons below.

#### 4.5 Ludwig-Miller Crossvalidation

In this section we briefly describe the crossvalidation method proposed by Ludwig and Miller (2005, LM from hereon), which we will compare to our proposed bandwidth in the application and simulations. See also Lee and Lemieux (2009). The LM bandwidth is the only proposed bandwidth selection procedure in the literature that is specifically aimed at providing a bandwidth in a regression discontinuity setting. Let  $N_-$  and  $N_+$  be the number of observations with  $X_i < c$  and  $X_i \geq c$  respectively. For  $\delta \in (0, 1)$ , let  $\theta_-(\delta)$  and  $\theta_+(\delta)$  be the  $\delta$ -th quantile of the  $X_i$  among the subsample of observations with  $X_i < c$  and  $X_i \geq c$  respectively, so that

$$\theta_-(\delta) = \arg \min_a \left\{ a \left| \left( \sum_{i=1}^N 1_{X_i \leq a} \right) \geq \delta \cdot N_- \right. \right\},$$

and

$$\theta_+(\delta) = \arg \min_a \left\{ a \left| \left( \sum_{i=1}^N 1_{c \leq X_i \leq a} \right) \geq \delta \cdot N_+ \right. \right\}.$$

Now the LM crossvalidation criterion we use is of the form:

$$CV_\delta(h) = \sum_{i=1}^N 1_{\theta_-(1-\delta) \leq X_i \leq \theta_+(\delta)} \cdot (Y_i - \hat{m}_h(X_i))^2.$$

(In fact, LM use a slightly different criterion function, where they sum up over all observations within a distance  $h_0$  from the threshold.) The estimator for the regression function here is  $\hat{m}_h(x)$  defined in equation (2.1). A key feature of this estimator is that for values of  $x < c$ , it only uses observations with  $X_i < x$  to estimate  $m(x)$ , and for values of  $x \geq c$ , it only uses observations with  $X_i > x$  to estimate  $m(x)$ , so that  $\hat{m}_h(X_i)$  does not depend on  $Y_i$ , as is necessary for crossvalidation. By using a value for  $\delta$  close to zero, we only use observations close to the threshold to evaluate the cross-validation criterion. The only concern is that by using too small value of  $\delta$ , we may not get a precisely estimated crossvalidation bandwidth. In a minor modification of the LM proposal we use the edge kernel instead of the Epanechnikov kernel they suggest. In our calculations we use  $\delta = 0.5$ . Any fixed value for  $\delta$  is unlikely to lead to an optimal bandwidth in general. Moreover, the criterion focuses implicitly on minimizing a criterion more akin to  $\mathbb{E} [(\hat{\mu}_+ - \mu_+)^2 - (\hat{\mu}_- - \mu_-)^2]$ , (with the errors in estimating  $\mu_-$  and  $\mu_+$

squared before adding them up), rather than rather than  $\text{MSE}(h) = \mathbb{E}[(\hat{\mu}_+ - \mu_+) - (\hat{\mu}_- - \mu_-)]^2$  in (4.19), where the error in the difference  $\mu_+ - \mu_-$  is squared. As a result even letting  $\delta \rightarrow 0$  with the sample size in the crossvalidation procedure is unlikely to result in an optimal bandwidth.

## 5 Extensions

In this section we discuss two extensions. First the fuzzy regression discontinuity design, and second the presence of covariates.

### 5.1 The Fuzzy Regression Design

In the Fuzzy Regression Discontinuity Design (FRD) the treatment  $W_i$  is not a deterministic function of the forcing variable. Instead the probability  $\Pr(W_i = 1|X_i = x)$  changes discontinuously at the threshold  $c$ . The focus is on the ratio

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

In an important theoretical paper Hahn, Todd and VanderKlaauw (2001) discuss identification in this setting, and show that in settings with heterogenous effects the estimand has an interpretation as a local average treatment effects (Imbens and Angrist, 1994). Now we need to estimate two regression functions, each at two boundary points: the expected outcome given the forcing variable  $\mathbb{E}[Y_i|X_i = x]$  to the right and left of the threshold  $c$  and the expected value of the treatment variable given the forcing variable  $\mathbb{E}[W_i|X_i = x]$ , again both to the right and left of  $c$ . Again we focus on a single bandwidth, now the bandwidth that minimize the mean squared error to this ratio. Define

$$\tau_Y = \lim_{x \downarrow c} \mathbb{E}[Y_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i|X_i = x], \quad \text{and} \quad \tau_W = \lim_{x \downarrow c} \mathbb{E}[W_i|X_i = x] - \lim_{x \uparrow c} \mathbb{E}[W_i|X_i = x],$$

with  $\hat{\tau}_Y$  and  $\hat{\tau}_W$  denoting the corresponding estimators, so that  $\tau_{\text{FRD}} = \tau_Y/\tau_W$ , and  $\hat{\tau}_{\text{FRD}} = \hat{\tau}_Y/\hat{\tau}_W$ . In large samples we can approximate the difference  $\hat{\tau}_{\text{FRD}} - \tau_{\text{FRD}}$  by

$$\hat{\tau}_{\text{FRD}} - \tau_{\text{FRD}} = \frac{1}{\tau_W}(\hat{\tau}_Y - \tau_Y) - \frac{\tau_Y}{\tau_W^2}(\hat{\tau}_W - \tau_W) + o_p((\hat{\tau}_Y - \tau_Y) + (\hat{\tau}_W - \tau_W)).$$

This is the basis for the asymptotic approximation to the MSE around  $h = 0$ :

$$\begin{aligned} \text{AMSE}_{\text{fuzzy}}(h) &= C_1 h^4 \left( \frac{1}{\tau_W} \left( m_{Y,+}^{(2)}(c) - m_{Y,-}^{(2)}(c) \right) - \frac{\tau_Y}{\tau_W^2} \left( m_{W,+}^{(2)}(c) - m_{W,-}^{(2)}(c) \right) \right)^2 \quad (5.21) \\ &+ \frac{C_2}{Nh f(c)} \left( \frac{1}{\tau_W^2} (\sigma_{Y,+}^2(c) + \sigma_{Y,-}^2(c)) + \frac{\tau_Y^2}{\tau_W^4} (\sigma_{W,+}^2(c) + \sigma_{W,-}^2(c)) - \frac{2\tau_Y}{\tau_W^3} (\sigma_{YW,+}(c) + \sigma_{YW,-}(c)) \right). \end{aligned}$$

In this expression the constants  $C_1$  and  $C_2$  are the same as before in Equation (3.6). The second derivatives of the regression functions,  $m_{Y,-}^{(2)}(c)$ ,  $m_{Y,+}^{(2)}(c)$ ,  $m_{W,-}^{(2)}(c)$ , and  $m_{W,+}^{(2)}(c)$ , are now defined separately for the treatment  $W$  and the outcome  $Y$ . In addition the conditional variances are indexed by the treatment and outcome. Finally the AMSE also depends on the right and left limit of the covariance of  $W$  and  $Y$  conditional on the forcing variable, at the threshold, denoted by  $\sigma_{YW,+}(c)$  and  $\sigma_{YW,-}(c)$  respectively.

The bandwidth that minimizes the AMSE in the fuzzy design is

$$h_{\text{opt,fuzzy}} = C_K \cdot N^{-1/5} \tag{5.22}$$

$$\times \left( \frac{\left( \sigma_{Y,+}^2(c) + \sigma_{Y,-}^2(c) \right) + \tau_{\text{FRD}}^2 \left( \sigma_{W,+}^2(c) + \sigma_{W,-}^2(c) \right) - 2\tau_{\text{FRD}} \left( \sigma_{YW,+}(c) + \sigma_{YW,-}(c) \right)}{f(c) \cdot \left( \left( m_{Y,+}^{(2)}(c) - m_{Y,-}^{(2)}(c) \right) - \tau_{\text{FRD}} \left( m_{W,+}^{(2)}(c) - m_{W,-}^{(2)}(c) \right) \right)^2} \right)^{1/5}.$$

The analogue of the bandwidth proposed for the sharp regression discontinuity is

$$\hat{h}_{\text{opt,fuzzy}} = C_K \cdot N^{-1/5}$$

$$\times \left( \frac{\left( \hat{\sigma}_{Y,+}^2(c) + \hat{\sigma}_{Y,-}^2(c) \right) + \hat{\tau}_{\text{FRD}}^2 \left( \hat{\sigma}_{W,+}^2(c) + \hat{\sigma}_{W,-}^2(c) \right) - 2\hat{\tau}_{\text{FRD}} \left( \hat{\sigma}_{YW,+}(c) + \hat{\sigma}_{YW,-}(c) \right)}{\hat{f}(c) \cdot \left( \left( \hat{m}_{Y,+}^{(2)}(c) - \hat{m}_{Y,-}^{(2)}(c) \right) - \hat{\tau}_{\text{FRD}} \left( \hat{m}_{W,+}^{(2)}(c) - \hat{m}_{W,-}^{(2)}(c) \right) \right)^2 + \hat{r}_{Y,+} + \hat{r}_{Y,-} + \hat{\tau}_{\text{FRD}} \left( \hat{r}_{W,+} + \hat{r}_{W,-} \right)} \right)^{1/5}.$$

We can implement this as follows. First, using the algorithm described for the sharp RD case separately for the treatment indicator and the outcome, estimate  $\tau_{\text{FRD}}$ ,  $\hat{\sigma}_{Y,+}^2$ ,  $\hat{\sigma}_{Y,-}^2$ ,  $\hat{\sigma}_{W,+}^2$ ,  $\hat{\sigma}_{W,-}^2$ ,  $\hat{m}_{Y,+}^{(2)}(c)$ ,  $\hat{m}_{Y,-}^{(2)}(c)$ ,  $\hat{m}_{W,+}^{(2)}(c)$ ,  $\hat{m}_{W,-}^{(2)}(c)$ ,  $\hat{r}_{Y,+}$ ,  $\hat{r}_{Y,-}$ ,  $\hat{r}_{W,+}$ , and  $\hat{r}_{W,-}$ . Second, using the initial Silverman bandwidth use the deviations from the means to estimate the conditional covariances  $\hat{\sigma}_{YW,+}(c)$  and  $\hat{\sigma}_{YW,-}(c)$ . Then substitute everything into the expression for the bandwidth. By the same argument as for the sharp RD case the resulting bandwidth has the asymptotic no-regret property.

## 5.2 Additional covariates

Typically the presence of additional covariates does not affect the regression discontinuity analyses very much. In most cases the distribution of the additional covariates does not exhibit any discontinuity around the threshold for the forcing variable, and as a result those covariates are approximately independent of the treatment indicator for samples constructed to be close to the threshold. In that case the covariates only affect the precision of the estimator, and one can modify the previous analysis using the conditional variance of  $Y_i$  given all covariates at the threshold,  $\sigma_-^2(c|x)$  and  $\sigma_+^2(c|x)$  instead of the unconditional variances  $\sigma_-^2(c)$  and  $\sigma_+^2(c)$ .

In practice this does not affect the optimal bandwidth much unless the additional covariates have great explanatory power (recall that the variance enters to the power  $1/5$ ), and the basic algorithm is likely to perform adequately even in the presence of covariates. For example, if the conditional variances are half the size of the unconditional ones, the bandwidth will change only by a factor  $1/2^{1/5}$ , or approximately 0.83.

## 6 An Illustration and Some Simulations

### 6.1 Data

To illustrate the implementation of these methods we use a data set previously analyzed by Lee (2008) in one of the most convincing applications of regression discontinuity designs. Lee studies the incumbency advantage in elections. His identification strategy is based on the discontinuity generated by the rule that the party with a majority vote share wins. The forcing variable  $X_i$  is the difference in vote share between the Democratic and Republican parties in one election, with the threshold  $c = 0$ . The outcome variable  $Y_i$  is vote share at the second election. There are 6558 observations (districts) in this data set, 3818 with  $X_i > 0$ , and 2740 with  $X_i < 0$ . The difference in voting percentages at the last election for the Democrats was 0.13, with a standard deviation of 0.46. Figure 1 plots the density of the forcing variable, in bins with width 0.05. Figure 2 plots the average value of the outcome variable, in 40 bins with width 0.05, against the forcing variable. The discontinuity is clearly visible in the raw data, lending credibility to any positive estimate of the treatment effect. The vertical line indicate the optimal bandwidth calculated below.

### 6.2 IK algorithm on Lee Data

In this section we implement our proposed bandwidth on the Lee dataset. For expositional reasons we gave all the intermediate steps.

Step 1: Estimation of density  $f(0)$  and conditional variance  $\sigma^2(0)$

We start with the modified Silverman bandwidth,

$$h_1 = 1.84 \cdot S_X \cdot N^{-1/5} = 1.84 \cdot 0.4553 \cdot 6558^{-1/5} = 0.1445.$$

There are  $N_{h_1,-} = 836$  units with values for  $X_i$  in the interval  $[-h_1, 0)$ , with an average outcome of  $\bar{Y}_{h_1,-} = 0.4219$  and a sample variance of  $S_{Y,h_1,-}^2 = 0.1047^2$ , and  $N_{h_1,+} = 862$  units with values for  $X_i$  in the interval  $[0, h_1]$ , with an average outcome of  $\bar{Y}_{h_1,+} = 0.5643$  and a sample

variance of  $S_{Y,h_{1,+}}^2 = 0.1202^2$ . This leads to

$$\hat{f}(0) = \frac{N_{h_{1,-}} + N_{h_{1,+}}}{2 \cdot N \cdot h_1} = \frac{836 + 862}{2 \cdot 6558 \cdot 0.1445} = 0.8962,$$

and

$$\hat{\sigma}_-^2(0) = S_{Y,h_{1,-}}^2 = 0.1047^2 \quad \text{and} \quad \hat{\sigma}_+^2(0) = S_{Y,h_{1,+}}^2 = 0.1202^2.$$

Step 2: Estimation of second derivatives  $\hat{m}_+^{(2)}(0)$  and  $\hat{m}_-^{(2)}(0)$

To estimate the curvature at the threshold, we first need to choose bandwidths  $h_{2,+}$  and  $h_{2,-}$ . We choose these bandwidths based on an estimate of  $\hat{m}^{(3)}(0)$ , obtained by fitting a global cubic with a jump at the threshold:

$$Y_i = \gamma_0 + \gamma_1 \cdot 1_{X_i \geq c} + \gamma_2 \cdot (X_i - c) + \gamma_3 \cdot (X_i - c)^2 + \gamma_4 \cdot (X_i - c)^3 + \varepsilon_i,$$

The least squares estimate for  $\gamma_4$  is  $\hat{\gamma}_4 = -0.1686$ , and thus the third derivative at the threshold is estimated as  $\hat{m}^{(3)}(0) = 6 \cdot \hat{\gamma}_4 = -1.0119$ . This leads to the two bandwidths

$$h_{2,+} = 3.56 \cdot \left( \frac{\hat{\sigma}_+^2(0)}{\hat{f}(0) \cdot (\hat{m}^{(3)}(0))^2} \right)^{1/7} \cdot N_+^{-1/7} = 0.6057, \quad \text{and} \quad h_{2,-} = 0.6105.$$

The two pilot bandwidths are used to fit two quadratics. The quadratic to the right of 0 is fitted on  $[0, 0.6057]$ , yielding  $\hat{m}_+^{(2)}(0) = 0.0455$  and the quadratic to the left is fitted on  $[-0.6105, 0]$  yielding  $\hat{m}_-^{(2)}(0) = -0.8471$ .

Step 3: Calculation of Regularization Terms  $\hat{r}_-$  and  $\hat{r}_+$ , and Calculation of  $\hat{h}_{\text{opt}}$

Next, the regularization terms are calculated. We obtain

$$\hat{r}_+ = \frac{720 \cdot \hat{\sigma}_+^2(0)}{N_{2,+} h_{2,+}^4} = \frac{720 \cdot 0.11202^2}{1983 \cdot 0.6057^4} = 0.0275 \quad \text{and} \quad \hat{r}_- = \frac{720 \cdot \hat{\sigma}_-^2(0)}{N_{2,-} h_{2,-}^4} = 0.0225.$$

Now we have all the ingredients to calculate the optimal bandwidth under different kernels and the corresponding RD estimates. Using the edge kernel with  $C_K = 3.4375$ , we obtain

$$\hat{h}_{\text{opt}} = C_K \left( \frac{\hat{\sigma}_-^2(0) + \hat{\sigma}_+^2(0)}{\hat{f}(0) \cdot \left[ (\hat{m}_+^{(2)}(0) - \hat{m}_-^{(2)}(0))^2 + (\hat{r}_+ + \hat{r}_-) \right]} \right)^{1/5} N^{-1/5} = 0.3005.$$

### 6.3 Eleven Estimates for the Lee Data

Here we calculate fourteen estimates of the ultimate object of interest, the size of the discontinuity in  $m(x)$  at zero. The first eight are based on local linear regression, and the last five on global polynomial regressions. The first is based on our proposed bandwidth. The second drops the regularization terms. The third uses a normal kernel and the corresponding Silverman bandwidth for estimating the density function at the threshold. The fourth estimates separate cubic regressions on the left and the right of the threshold to derive the bandwidth for estimating the second derivatives. The fifth estimates the conditional variance at the threshold assuming its left and right limit are identical. The sixth uses a uniform kernel instead of the optimal edge kernel. The seventh bandwidth is based on the DeJardin-McCall criterion. The eighth bandwidth is based on the Ludwig-Miller crossvalidation. The last five are based on global linear, quadratic, cubic, quartic, and quintic regressions. The point estimates and standard errors are presented in Table 1. To investigate the overall sensitivity of the point estimates to the bandwidth choice, Figure 3 plots the RD estimates, and the associated 95% confidence intervals, as a function of the bandwidth, for  $h$  between 0 and 1. The solid vertical line indicates the optimal bandwidth ( $\hat{h}_{\text{opt}} = 0.3005$ ).

### 6.4 A Small Simulation Study

Next we conduct a small Monte Carlo study assess the properties of the proposed bandwidth selection rule in practice. We consider three designs, the first based on the Lee data, the second based on the Ludwig-Miller data, and the last a modified Lee design.

In the first design, based on the Lee data, we use a Beta distribution for the forcing variable. Let  $Z$  have a beta distribution with parameters  $\alpha = 5$  and  $\beta = 5$ , then the forcing variable is  $X = 2 \cdot Z - 1$ . The regression function is a 5-th order polynomial, with separate coefficients for  $X_i < 0$  and  $X_i > 0$ , with the coefficients estimated on the Lee data, leading to

$$m_{\text{Lee}}(x) = \begin{cases} 0.52 + 0.76x - 2.29x^2 + 5.66x^3 - 5.87x^4 + 2.09x^5 & \text{if } x < 0, \\ 0.48 + 1.43x + 8.69x^2 + 25.50x^3 + 29.16x^4 + 11.13x^5 & \text{if } x \geq 0. \end{cases}$$

The error variance is  $\sigma_\varepsilon^2 = 0.1356^2$ . We use data sets of size 500 (smaller than the Lee data set with 6558 observations, but more in line with common sample sizes).

In the second design we use the same distribution for the forcing variable as in the first design. We again have 500 observations per sample, and the true regression function is quadratic both to the left and to the right of the threshold, but with different coefficients:

$$m_{\text{quad}}(x) = \begin{cases} 3x^2 & \text{if } x < 0, \\ 4x^2 & \text{if } x \geq 0, \end{cases}$$

implying the data generating process is close to the point where the bandwidth  $h_{\text{opt}}$  is infinite (because the left and right limit of the second derivative are 6 and 8 respectively), and one may expect substantial effect from the regularization. The error variance is the same as in the first design,  $\sigma_\varepsilon^2 = 0.1356^2$ .

In Table 2 we report results for the same estimators as we reported in Table 1 for the real data. We include one additional bandwidth choice, namely the infeasible optimal bandwidth  $h_{\text{opt}}$ , which can be derived because we know the data generating process. In Table 2 we present for the both designs, the mean (Mean) and standard deviation (Std) of the bandwidth choices, and the bias (Bias) and the root-mean-squared-error (RMSE) of the estimator for  $\tau$ .

First consider the design motivated by the Lee data. All bandwidth selection methods combined with local linear estimation perform fairly similarly under this design. There is considerably more variation in the performance of the global polynomial estimators. The quadratic estimator performs very well, but adding a third order term more than doubles the RMSE. The quintic approximation does very well, not surprisingly given the data generating process that involves a fifth order polynomial.

In the second design the regularization matters, and the bandwidth choices based on different criterion functions perform worse than the proposed bandwidth in terms of RMSE, increasing it by about 28%. The global quadratic estimator obviously performs well here because it corresponds to the data generating process, but it is interesting that the local linear estimators have a RMSE very similar to the global quadratic estimator.

## 7 Conclusion

In this paper we propose a fully data-driven, asymptotically optimal bandwidth choice for regression discontinuity settings. Although this choice has asymptotic optimality properties, it still relies on somewhat arbitrary initial bandwidth choices. Rather than relying on a single bandwidth, we therefore encourage researchers to use this bandwidth choice as a reference point for assessing sensitivity to bandwidth choice in regression discontinuity settings. The proposed procedure is the first available procedure with optimality properties. The bandwidth selection procedures commonly used in this literature are typically based on different objectives, for example on global measures, not tailored to the specific features of the regression discontinuity setting. We compare our proposed bandwidth selection procedure to the crossvalidation procedure developed by Ludwig and Miller (2005), which is tailored to the regression discontinuity setting, but which requires the researcher to specify an additional tuning parameter, as well

as to the procedure proposed by DeJardins and McCall (2008). We find that our proposed method works well in realistic settings, including one motivated by data previously analyzed by Lee (2008).

## Appendix

To obtain the MSE expansions for the RD estimand, we first obtain the bias and variance estimates from estimating a regression function at a boundary point. Fan and Gijbels (1992) derive the same claim but under weaker assumptions (such as thin tailed kernels rather than compact kernels) and hence their proof is less transparent and not easily generalizable to multiple dimensions and derivatives. The proof we outline is based on Ruppert and Wand (1994) but since they only cursorily indicate the approach for a boundary point in multiple dimensions, we provide a simple proof for our case.

**Lemma A.1:** (MSE FOR ESTIMATION OF A REGRESSION FUNCTION AT THE BOUNDARY)

Suppose (i) we have  $N$  pairs  $(Y_i, X_i)$ , independent and identically distributed, with  $X_i \geq 0$ , (ii)  $m(x) = \mathbb{E}[Y_i | X_i = x]$  is three times continuously differentiable, (iii) the density of  $X_i$ ,  $f(x)$ , is continuously differentiable at  $x = 0$ , with  $f(0) > 0$ , (iv) the conditional variance  $\sigma^2(x) = \text{Var}(Y_i | X_i = x)$  is bounded, and continuous at  $x = 0$ , (v) we have a kernel  $K : \mathbb{R}^+ \mapsto \mathbb{R}$ , with  $K(u) = 0$  for  $u \geq \bar{u}$ , and  $\int_0^{\bar{u}} K(u) du = 1$ , and define  $K_h(u) = K(u/h)/h$ . Define  $\mu = m(0)$ , and

$$(\hat{\mu}_h, \hat{\beta}_h) = \arg \min_{\mu, \beta} \sum_{i=1}^N (Y_i - \mu - \beta \cdot X_i)^2 \cdot K_h(X_i).$$

Then:

$$\mathbb{E}[\hat{\mu} | X_1, \dots, X_N] - \mu = C_1^{1/2} m^{(2)}(0) h^2 + o_p(h^2), \quad (\text{A.1})$$

$$\mathbb{V}(\hat{\mu} | X_1, \dots, X_N) = C_2 \frac{\sigma^2(0)}{f(0)Nh} + o_p\left(\frac{1}{Nh}\right), \quad (\text{A.2})$$

and

$$\mathbb{E}[(\hat{\mu} - \mu)^2 | X_1, \dots, X_N] = C_1 \left(m^{(2)}(0)\right)^2 h^4 + C_2 \frac{\sigma^2(0)}{f(0)Nh} + o_p\left(h^4 + \frac{1}{Nh}\right), \quad (\text{A.3})$$

where the kernel-specific constants  $C_1$  and  $C_2$  are those given in Lemma 3.1.

Before proving Lemma A.1, we state and prove two preliminary results.

**Lemma A.2:** Define  $F_j = \frac{1}{N} \sum_{i=1}^N K_h(X_i) X_i^j$ . Under the assumptions in Lemma A.1, (i), for nonnegative integer  $j$ ,

$$F_j = h^j f(0) \nu_j + o_p(h^j) \equiv h^j (F_j^* + o_p(1)),$$

with  $\nu_j = \int_0^\infty t^j K(t) dt$  and  $F_j^* \equiv f(0) \nu_j$ , and (ii), If  $j \geq 1$ ,  $F_j = o_p(h^{j-1})$ .

**Proof:**  $F_j$  is the average of independent and identically distributed random variables, so

$$F_j = \mathbb{E}[F_j] + O_p\left(\text{Var}(F_j)^{1/2}\right).$$

The mean of  $F_j$  is, using a change of variables from  $z$  to  $x = z/h$ ,

$$\begin{aligned} \mathbb{E}[F_j] &= \int_0^\infty \frac{1}{h} K\left(\frac{z}{h}\right) z^j f(z) dz = h^j \int_0^\infty K(x) x^j f(hx) dx \\ &= h^j \int_0^\infty K(x) x^j f(0) dx + h^{j+1} \int_0^\infty K(x) x^{j+1} \frac{f(hx) - f(0)}{hx} dx = h^j f(0) \nu_j + O(h^{j+1}). \end{aligned}$$

The variance of  $F_j$  can be bounded by

$$\frac{1}{N} \mathbb{E}\left[(K_h(X_i))^2 X_i^{2j}\right] = \frac{1}{Nh^2} \mathbb{E}\left[\left(K\left(\frac{X_i}{h}\right)\right)^2 \cdot X_i^{2j}\right] = \frac{1}{Nh^2} \int_0^\infty \left(K\left(\frac{z}{h}\right)\right)^2 \cdot z^{2j} f(z) dz.$$

By a change of variables from  $z$  to  $x = z/h$ , this is equal to

$$\frac{h^{2j-1}}{N} \int_0^\infty (K(x))^2 \cdot x^{2j} f(hx) dx = O\left(\frac{h^{2j-1}}{N}\right) = o\left(\left(\frac{h^j}{hN^{1/2}}\right)^2\right) = o\left((h^j)^2\right).$$

Hence

$$F_j = \mathbb{E}[F_j] + o_p(h^j) = h^j f(0) \nu_j + o_p(h^j) = h^j \cdot (f(0) \nu_j + o_p(1)).$$

□

**Lemma A.3:** Let  $G_j = \frac{1}{N} \sum_{i=1}^N K_h^2(X_i) X_i^j \sigma^2(X_i)$ . Under the assumptions from Lemma A.1,

$$G_j = h^{j-1} \sigma^2(0) f(0) \pi_j (1 + o_p(1)), \quad \text{with } \pi_j = \int_0^\infty t^j K^2(t) dt.$$

**Proof:** This claim is proved in a manner exactly like Lemma A.1, here using in addition the differentiability of the conditional variance function.  $\square$

**Proof of Lemma A.1:** Define  $R = [\iota \ X]$ , where  $\iota$  is a  $N$ -dimensional column of ones, define the diagonal weight matrix  $W$  with  $(i, i)$ th element equal to  $K_h(X_i)$ , and define  $e_1 = (1 \ 0)'$ . Then

$$\hat{m}(0) = \hat{\mu} = e_1'(R'WR)^{-1}R'WY.$$

The conditional bias is  $B = \mathbb{E}[\hat{m}(0)|X_1, \dots, X_N] - m(0)$ . Note that  $\mathbb{E}(\hat{m}(0)|X) = e_1'(R'WR)^{-1}R'WM$  where  $M = (m(X_1), \dots, m(X_N))'$ . Let  $m^{(k)}(x)$  denote the  $k$ th derivative of  $m(x)$  with respect to  $x$ . Using Assumption (ii) in Lemma A.1, a Taylor expansion of  $m(X_i)$  yields:

$$m(X_i) = m(0) + m^{(1)}(0)X_i + \frac{1}{2}m^{(2)}(0)X_i^2 + T_i,$$

where

$$|T_i| \leq \sup_x m^{(3)}(x) \cdot X_i^3.$$

Thus we can write the vector  $M$  as

$$M = R \begin{pmatrix} m(0) \\ m^{(1)}(0) \end{pmatrix} + S + T.$$

where the vector  $S$  has  $i$ th element equal to  $S_i = m^{(2)}(0)X_i^2/2$ , and the vector  $T$  has typical element  $T_i$ . Therefore the bias can be written as

$$B = e_1'(R'WR)^{-1}R'WM - m(0) = e_1'(R'WR)^{-1}R'W(S + T).$$

Using Lemma A.2 we have

$$\begin{aligned} \left(\frac{1}{N}R'WR\right)^{-1} &= \begin{pmatrix} F_0 & F_1 \\ F_1 & F_2 \end{pmatrix}^{-1} = \frac{1}{F_0F_2 - F_1^2} \begin{pmatrix} F_2 & -F_1 \\ -F_1 & F_0 \end{pmatrix} \\ &= \begin{pmatrix} \frac{F_2^*}{F_0^*F_2^* - (F_1^*)^2} + o_p(1) & -\frac{1}{h} \left( \frac{F_1^*}{F_0^*F_2^* - (F_1^*)^2} + o_p(1) \right) \\ -\frac{1}{h} \left( \frac{F_1^*}{F_0^*F_2^* - (F_1^*)^2} + o_p(1) \right) & \frac{1}{h^2} \left( \frac{F_0^*}{F_0^*F_2^* - (F_1^*)^2} + o_p(1) \right) \end{pmatrix} \\ &= \begin{pmatrix} \frac{\nu_2}{(\nu_0\nu_2 - \nu_1^2)f(c)} + o_p(1) & -\frac{\nu_1}{(\nu_0\nu_2 - \nu_1^2)f(c)h} + o_p\left(\frac{1}{h}\right) \\ -\frac{\nu_1}{(\nu_0\nu_2 - \nu_1^2)f(c)h} + o_p\left(\frac{1}{h}\right) & O_p\left(\frac{1}{h^2}\right) \end{pmatrix} \\ &= \begin{pmatrix} O_p(1) & O_p\left(\frac{1}{h}\right) \\ O_p\left(\frac{1}{h}\right) & O_p\left(\frac{1}{h^2}\right) \end{pmatrix}. \end{aligned}$$

Next

$$\left|\frac{1}{N}R'WT\right| = \sup_x m^{(3)}(x) \cdot \begin{pmatrix} F_3 \\ F_4 \end{pmatrix} = \begin{pmatrix} o_p(h^2) \\ o_p(h^3) \end{pmatrix}.$$

Thus

$$e_1'(R'WR)^{-1}R'WT = O_p(1) \cdot o_p(h^2) + O_p\left(\frac{1}{h}\right) \cdot o_p(h^3) = o_p(h^2),$$

implying

$$B = e_1'(R'WR)^{-1}R'WS + o_p(h^2).$$

Similarly:

$$\frac{1}{N}(R'WS) = \frac{1}{2}m^{(2)}(0) \begin{pmatrix} \frac{1}{N} \sum_{i=1}^N K_h(X_i) X_i^2 \\ \frac{1}{N} \sum_{i=1}^N K_h(X_i) X_i^3 \end{pmatrix} = \frac{1}{2}m^{(2)}(0)f(0) \begin{pmatrix} \nu_2 h^2 + o_p(h^2) \\ \nu_3 h^3 + o_p(h^3) \end{pmatrix}.$$

Therefore:

$$B = e_1'(R'WR)^{-1}R'WS + o_p(h^2) = \frac{1}{2}m^{(2)}(c) \left( \frac{\nu_2^2 - \nu_3\nu_1}{\nu_0\nu_2 - \nu_1^2} \right) h^2 + o_p(h^2).$$

This finishes the proof for the first part of the result in Lemma A.1, equation (A.1).

Next, we consider the expression for the conditional variance in (A.2).

$$V = \mathbb{V}(\hat{m}(0)|X_1, \dots, X_N) = e_1'(R'WR)^{-1}R'W\Sigma WR(R'WR)^{-1}e_1,$$

where  $\Sigma$  is the diagonal matrix with  $(i, i)$ th element equal to  $\sigma^2(X_i)$ .

Consider the middle term

$$\frac{1}{N}R'W\Sigma WR = \begin{pmatrix} \frac{1}{N}\sum_i K_h^2(X_i)\sigma^2(X_i) & \frac{1}{N}\sum_i K_h^2(X_i)X_i\sigma^2(X_i) \\ \frac{1}{N}\sum_i K_h^2(X_i)X_i\sigma^2(X_i) & \frac{1}{N}\sum_i K_h^2(X_i)X_i^2\sigma^2(X_i) \end{pmatrix} = \begin{pmatrix} G_0 & G_1 \\ G_1 & G_2 \end{pmatrix}.$$

Thus we have:

$$\begin{aligned} NV &= \frac{1}{(F_0F_2 - F_1^2)^2} e_1' \begin{pmatrix} F_2 & -F_1 \\ -F_1 & F_0 \end{pmatrix} \begin{pmatrix} G_0 & G_1 \\ G_1 & G_2 \end{pmatrix} \begin{pmatrix} F_2 & -F_1 \\ -F_1 & F_0 \end{pmatrix} e_1 \\ &= \frac{F_2^2G_0 - 2F_1F_2G_1 + F_1^2G_2}{(F_0F_2 - F_1^2)^2} \end{aligned}$$

Applying lemmas A.1 and A.2 this leads to

$$V = \frac{\sigma^2(0)}{f(0)Nh} \cdot \left( \frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_0\nu_2 - \nu_1^2)^2} \right) + o_p\left(\frac{1}{Nh}\right).$$

This finishes the proof for the statement in (A.2). The final result in (A.3) follows directly from the first two results.  $\square$

**Proof of Lemma 3.1:** Applying Lemma A.1 to the  $N_+$  units with  $X_i \geq c$ , implies that

$$\mathbb{E}[\hat{\mu}_+ - \mu_+ | X_1, \dots, X_N] = C_1^{1/2}m_+^{(2)}(c)h^2 + o_p(h^2),$$

and

$$\mathbb{V}(\hat{\mu}_+ - \mu_+ | X_1, \dots, X_N) = C_2 \frac{\sigma_+^2(c)}{f_{X|X \geq c}(c)N_+h} + o_p\left(\frac{1}{N_+h}\right).$$

Because  $N_+/N = \text{pr}(X_i \geq c) + O(1/N)$ , and  $f_{X|X \geq c}(x) = f(x)/\text{Pr}(X_i \geq c)$  (and thus  $f_{X|X \geq c}(c) = f_+(c)/\text{Pr}(X_i \geq c)$ ), it follows that

$$\mathbb{V}(\hat{\mu}_+ - \mu_+ | X_1, \dots, X_N) = C_2 \frac{\sigma_+^2(c)}{f_+(c)Nh} + o_p\left(\frac{1}{Nh}\right).$$

Conditional on  $X_1, \dots, X_N$  the covariance between  $\hat{\mu}_+$  and  $\hat{\mu}_-$  is zero, and thus, combining the results from applying Lemma A.1 also to the units with  $X_i < c$ , we find

$$\begin{aligned} \mathbb{E}[(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}})^2 | X_1, \dots, X_N] &= \mathbb{E}[(\hat{\mu}_+ - \hat{\mu}_- - (\hat{\mu}_+ - \hat{\mu}_-))^2 | X_1, \dots, X_N] \\ &= \mathbb{E}[(\hat{\mu}_+ - \mu_+)^2 | X_1, \dots, X_N] + \mathbb{E}[(\hat{\mu}_- - \mu_-)^2 | X_1, \dots, X_N] \\ &\quad - 2 \cdot \mathbb{E}[\hat{\mu}_+ - \mu_+ | X_1, \dots, X_N] \cdot \mathbb{E}[\hat{\mu}_- - \mu_- | X_1, \dots, X_N] \\ &= C_1 \cdot h^4 \cdot (m_+^{(2)}(c) - m_-^{(2)}(c))^2 + \frac{C_2}{N \cdot h} \cdot \left( \frac{\sigma_+^2(c)}{f_+(c)} + \frac{\sigma_-^2(c)}{f_-(c)} \right) + o_p\left(h^4 + \frac{1}{N \cdot h}\right), \end{aligned}$$

proving the first result in Lemma 3.1.

For the second part of Lemma 3.1, solve

$$h_{\text{opt}} = \arg \min_h \left( C_1 h^4 (m_+^{(2)}(c) - m_-^{(2)}(c))^2 + C_2 \left( \frac{\sigma_+^2(c)}{f_+(c)Nh} + \frac{\sigma_-^2(c)}{f_-(c)Nh} \right) \right),$$

which leads to

$$h_{\text{opt}} = \left( \frac{C_2}{4C_1} \right)^{1/5} \left( \frac{\frac{\sigma_+^2(c)}{f_+(c)} + \frac{\sigma_-^2(c)}{f_-(c)}}{(m_+^{(2)}(c) - m_-^{(2)}(c))^2} \right)^{1/5} N^{-1/5}.$$

□

**Motivation for the Bandwidth Choice in Equation (4.14) in Step 2 of bandwidth algorithm**

Fan and Gijbels (1996 Theorem 3.2) give an asymptotic approximation to the MSE for an estimator of the  $\nu$ -th derivative of a regression function at a boundary point, using a  $p$ -th order local polynomial (using the notation in Fan and Gijbels). Specializing this to our case, with the boundary point  $c$ , a uniform one-sided kernel  $K(t) = 1_{0 \leq t \leq 1}$ , and interest in the 2-nd derivative using a local quadratic approximation ( $\nu = p = 2$ , their MSE formula simplifies to

$$MSE = \left( \frac{1}{9} K_1^2 \left( m_+^{(3)}(c) \right)^2 h^2 + 4K_2 \frac{1}{Nh^5} \frac{\sigma_+^2(c)}{f_+(c)} \right) (1 + o_p(1))$$

Here

$$K_1 = \int t^3 K^*(t) dt \quad K_2 = \int (K^*(t))^2 dt,$$

where

$$K^*(t) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}' \begin{pmatrix} \mu_0 & \mu_1 & \mu_2 \\ \mu_1 & \mu_2 & \mu_3 \\ \mu_2 & \mu_3 & \mu_4 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix} \cdot K(t), \quad \text{with } \mu_k = \int q^k K(q) dq = \frac{1}{(k+1)},$$

so that

$$K^*(t) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}' \begin{pmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ t \\ t^2 \end{pmatrix} \cdot K(t) = (30 - 180t + 180t^2) \cdot \mathbf{1}_{[0,1]},$$

and therefore,  $K_1 = 1.5$  and  $K_2 = 180$ . Thus

$$MSE = \left( \frac{1}{4} \left( m_+^{(3)}(c) \right)^2 h^2 + 720 \frac{1}{Nh^5} \frac{\sigma_+^2(c)}{f_+(c)} \right) (1 + o_p(1)).$$

Minimizing this over  $h$  leads to

$$h_{2,+} = 7200^{1/7} \cdot \left( \frac{\sigma_+^2(c)}{f_+(c) \left( m_+^{(3)}(c) \right)^2} \right)^{1/7} N_+^{-1/7} \approx 3.56 \cdot \left( \frac{\sigma_+^2(c)}{f_+(c) \left( m_+^{(3)}(c) \right)^2} \right)^{1/7} N_+^{-1/7}.$$

This is the expression in the text for  $h_{2,+}$  except for the addition of the 0.01 term that ensures the necessary properties if the estimate of  $m^{(3)}(c)$  converges to zero. □

**Proof of Theorem 4.1:** Before directly proving the three claims in the theorem, we make some preliminary observations. Write

$$h_{\text{opt}} = C_{\text{opt}} \cdot N^{-1/5}, \quad \text{with } C_{\text{opt}} = C_K \cdot \left( \frac{2\sigma^2(c)}{f(c) \cdot \left( \left( m_+^{(2)}(c) - m_-^{(2)}(c) \right)^2 \right)} \right)^{1/5},$$

and

$$\hat{h}_{\text{opt}} = \hat{C}_{\text{opt}} \cdot N^{-1/5}, \quad \text{with } \hat{C}_{\text{opt}} = C_K \cdot \left( \frac{2\hat{\sigma}^2(c)}{\hat{f}(c) \cdot \left( \left( \hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) + \hat{r}_+ + \hat{r}_- \right)^2 \right)} \right)^{1/5}.$$

First we show that the various estimates of the functionals in  $\hat{C}_{\text{opt}}$ ,  $\hat{\sigma}_-^2(c)$ ,  $\hat{\sigma}_+^2(c)$ ,  $\hat{f}(c)$ ,  $\hat{m}_+^{(2)}(c)$  and  $\hat{m}_-^{(2)}(c)$  converge to their counterparts in  $C_{\text{opt}}$ ,  $\sigma_-^2(c)$ ,  $\sigma_+^2(c)$ ,  $f(c)$ ,  $m_+^{(2)}(c)$  and  $m_-^{(2)}(c)$ . Consider  $\hat{f}(c)$ . This is a histogram estimate of density at  $c$ , with bandwidth  $h = CN^{-1/5}$ . Hence  $\hat{f}(c)$  is consistent for  $f(c)$  if  $f_-(c) = f_+(c) = f(c)$ , if the left and righthand limit are equal, and for  $(f_-(c) + f_+(c))/2$  if they are different.

Next, consider  $\hat{\sigma}_-^2(c)$  (and  $\hat{\sigma}_+^2(c)$ ). Because it is based on a bandwidth  $h = C \cdot N^{-1/5}$  that converges to zero, it is consistent for  $\sigma_-^2(c)$  if  $\sigma_-^2(c) = \sigma_+^2(c) = \sigma^2(c)$ .

Third, consider  $\hat{m}_+^{(2)}(c)$ . This is a local quadratic estimate using a one sided uniform kernel. From Fan and Gijbels (1996), Theorem 3.2, it follows that to guarantee consistency of  $\hat{m}_+^{(2)}(c)$  for  $m_+^{(2)}(c)$  we need both

$$h_{2,+} = o_p(1) \quad \text{and} \quad (Nh_{2,+}^5)^{-1} = o_p(1). \quad (\text{A.4})$$

Let  $m_3$  be the probability limit of  $\hat{m}^{(3)}(c)$ . This probability limit need not be equal to  $m^{(3)}(c)$ , but it will exist under the assumptions in Theorem 4.1. As long as this probability limit differs from zero, then  $h_{2,+} = O_p(N^{-1/7})$ , so that the two conditions in (A.4) are satisfied and  $\hat{m}_+^{(2)}(c)$  is consistent for  $m_+^{(2)}(c)$ .

Fourth, consider  $\hat{r}_+ = 720\hat{\sigma}_+^2(c)/(N_{2,+}h_{2,+}^4)$ . The numerator converges to  $720\hat{\sigma}_+^2(c)$ . The denominator is approximately  $N_{2,+} \cdot h_{2,+}^4 = (C \cdot N \cdot h_{2,+}) \cdot C \cdot N^{-4/7} = C \cdot N^{2/7}$ , so that the ratio is  $C \cdot N^{-2/7} = o_p(1)$ . A similar result holds for  $\hat{r}_-$ .

Now we turn to the statements in Theorem 4.1. We will prove (iii), then (iv), and then (i) and (ii). First consider (iii). If  $m_+^{(2)}(c) - m_-^{(2)}(c)$  differs from zero, then  $C_{\text{opt}}$  is finite. Moreover, in that case  $(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c))^2 + \hat{r}_+ + \hat{r}_-$  converges to  $(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c))^2$ , and  $\hat{C}_{\text{opt}}$  converges to  $C_{\text{opt}}$ . These two implications in turn lead to the result that  $(\hat{h}_{\text{opt}} - h_{\text{opt}})/h_{\text{opt}} = (\hat{C}_{\text{opt}} - C_{\text{opt}})/C_{\text{opt}} = o_p(1)$ , finishing the proof for (iii).

Next, we prove (iv). Because  $h_{\text{opt}} = C_{\text{opt}} \cdot N^{-1/5}$ , it follows that

$$\text{MSE}(h_{\text{opt}}) = \text{AMSE}(h_{\text{opt}}) + o\left(h_{\text{opt}}^4 + \frac{1}{N \cdot h_{\text{opt}}}\right) = \text{AMSE}(h_{\text{opt}}) + o\left(N^{-4/5}\right).$$

Because  $\hat{h}_{\text{opt}} = (\hat{C}_{\text{opt}}/C_{\text{opt}}) \cdot C_{\text{opt}}N^{-1/5}$ , and  $\hat{C}_{\text{opt}}/C_{\text{opt}} \rightarrow 1$  it follows that

$$\text{MSE}(\hat{h}_{\text{opt}}) = \text{AMSE}(\hat{h}_{\text{opt}}) + o\left(N^{-4/5}\right).$$

Therefore

$$N^{4/5} \cdot \left(\text{MSE}(\hat{h}_{\text{opt}}) - \text{MSE}(h_{\text{opt}})\right) = N^{4/5} \cdot \left(\text{AMSE}(\hat{h}_{\text{opt}}) - \text{AMSE}(h_{\text{opt}})\right) + o_p(1),$$

and

$$\begin{aligned} \frac{\text{MSE}(\hat{h}_{\text{opt}}) - \text{MSE}(h_{\text{opt}})}{\text{MSE}(h_{\text{opt}})} &= \frac{N^{4/5} \cdot \left(\text{MSE}(\hat{h}_{\text{opt}}) - \text{MSE}(h_{\text{opt}})\right)}{N^{4/5} \cdot \text{MSE}(h_{\text{opt}})} \\ &= \frac{N^{4/5} \cdot \left(\text{AMSE}(\hat{h}_{\text{opt}}) - \text{AMSE}(h_{\text{opt}})\right) + o_p(1)}{N^{4/5} \cdot \text{AMSE}(h_{\text{opt}}) + o_p(1)} \\ &= \frac{N^{4/5} \cdot \left(\text{AMSE}(\hat{h}_{\text{opt}}) - \text{AMSE}(h_{\text{opt}})\right)}{N^{4/5} \cdot \text{AMSE}(h_{\text{opt}})} + o_p(1). \end{aligned}$$

Because  $N^{4/5} \cdot \text{AMSE}(h_{\text{opt}})$  converges to a nonzero constant, all that is left to prove in order to establish (iii) is that

$$N^{4/5} \cdot \left(\text{AMSE}(\hat{h}_{\text{opt}}) - \text{AMSE}(h_{\text{opt}})\right) = o_p(1). \quad (\text{A.5})$$

Substituting in, we have

$$\begin{aligned} &N^{4/5} \cdot \left(\text{AMSE}(\hat{h}_{\text{opt}}) - \text{AMSE}(h_{\text{opt}})\right) \\ &= C_1 \cdot \left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2 \cdot \left((N^{1/5}h_{\text{opt}})^4 - N^{1/5}\hat{h}_{\text{opt}}^4\right) + \left(\frac{C_2}{N^{1/5} \cdot h_{\text{opt}}} - \frac{C_2}{N^{1/5} \cdot \hat{h}_{\text{opt}}}\right) \cdot \left(\frac{\sigma_+^2(c)}{f_+(c)} + \frac{\sigma_-^2(c)}{f_-(c)}\right) \\ &= o_p(1), \end{aligned}$$

because  $N^{1/5}h_{\text{opt}} - N^{1/5}\hat{h}_{\text{opt}} = C_{\text{opt}} - \hat{C}_{\text{opt}} = o_p(1)$ , so that A.5 holds, and therefore (iv) is proven.

Now we turn to (ii). Under the conditions for (ii),  $\hat{h}_{\text{opt}} = \hat{C}_{\text{opt}}N^{-1/5}$ , with  $\hat{C}_{\text{opt}} \rightarrow C_{\text{opt}}$ , a nonzero constant. Then Lemma 3.1 implies that  $\text{MSE}(\hat{h}_{\text{opt}})$  is  $O_p(\hat{h}_{\text{opt}}^4 + N^{-1}\hat{h}_{\text{opt}}^{-1}) = O_p(N^{-4/5})$  so that  $\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p(N^{-2/5})$ .

Finally, consider (i). If Assumption 3.6 holds, then  $\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p(N^{-2/5})$ , and the result holds. Now suppose Assumption 3.6 does not hold and  $m_+^{(2)}(c) - m_-^{(2)}(c) = 0$ . Because  $h_{2,+} = CN^{-1/7}$ , it follows that  $r_+ = CN^{-1}h^{-4} = CN^{-3/7}$  (with each time different constants  $C$ ), it follows that  $\hat{h}_{\text{opt}} = C(N^{3/7})^{1/5}N^{-1/5} = CN^{-4/35}$ , so that the  $\text{MSE}(h) = CN^{-24/35} + \tilde{C}N^{-31/35} = CN^{-16/35}$  (note that the leading bias term is now  $O(h^3)$  so that the square of the bias is  $O(h^6) = O(N^{-24/25})$ ) and thus  $\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}} = O_p(N^{-12/35})$ .  $\square$

## References

- COOK, T., (2008), ““Waiting for Life to arrive”: A history of the regression-discontinuity design in Psychology, Statistics and Economics,” *Journal of Econometrics*, 142, 636-654.
- CHENG, M.-Y., FAN, J. AND MARRON, J.S., (1997), “On Automatic Boundary Corrections,” *The Annals of Statistics*, 25, 1691-1708.
- DESJARDINS, S., AND B. MCCALL, (2008) “The Impact of the Gates Millennium Scholars Program on the Retention, College Finance- and Work-Related Choices, and Future Educational Aspirations of Low-Income Minority Students,” Unpublished Manuscript.
- FAN, J. AND GIJBELS, I., (1992), “Variable bandwidth and local linear regression smoothers,” *The Annals of Statistics*, 20, 2008-2036.
- FAN, J. AND GIJBELS, I., (1996), *Local polynomial modeling and its implications*, Monographs on Statistics and Applied Probability 66, Chapman and Hall/CRC, Boca Raton, FL.
- FRANDSEN, B., (2008), “A Nonparametric Estimator for Local Quantile Treatment Effects in the Regression Discontinuity Design,” Unpublished Working Paper, Dept of Economics, MIT.
- FRÖLICH, M., (2007), “Regression Discontinuity Design with Covariates,” IZA Discussion Paper 3024 Bonn.
- FRÖLICH, M., AND B. MELLY, (2008), “Quantile Treatment Effects in the Regression Discontinuity Design,” IZA Discussion Paper 3638, Bonn.
- HAHN, J., TODD, P., AND VAN DER KLAUW, W., (2001), “Regression discontinuity,” *Econometrica*, 69(1), 201-209.
- HÄRDLE, W., (1992), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, UK.
- IMBENS, G., AND J. ANGRIST (1994), “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, Vol. 61, No. 2, 467-476.
- IMBENS, G., AND LEMIEUX, T., (2008), “Regression discontinuity designs,” *Journal of Econometrics*, 142, 615-635.
- KALYANARAMAN, K., (2008), “Bandwidth selection for linear functionals of the regression function,” Working Paper, Harvard University Department of Economics, June, 2008.
- LEE, D., (2008), “Randomized experiments from non-random selection in U.S. House elections,” *Journal of Econometrics*, 142, 675-697.
- LEE, D., AND T. LEMIEUX, (2009), “Regression Discontinuity Designs in Economics,” Working Paper, Dept of Economics, Princeton University.
- LI, K.-C., (1987), “Asymptotic Optimality for  $C_p$ ,  $C_L$ , crossvalidation and Generalized crossvalidation: Discrete Index Set,” *The Annals of Statistics*, vol. 15(3), 958-975.
- LUDWIG, J. AND MILLER, D., (2005), “Does Head Start Improve Children’s Life Chances? Evidence from a Regression Discontinuity Design,” NBER Working Paper 11702.
- LUDWIG, J. AND MILLER, D., (2007), “Does Head Start Improve Children’s Life Chances? Evidence from a Regression Discontinuity Design,” *Quarterly Journal of Economics*.
- MCCRARY, J., (2008), “Manipulation of the running variable in the regression discontinuity design: A density test,” *Journal of Econometrics*, 142, 698-714.
- PORTER, J., (2003), “Estimation in the Regression Discontinuity Model,” Working Paper, Harvard University Department of Economics, draft date September 25, 2003.

- POWELL, J., AND T. STOKER, (1996), "Optimal Bandwidth Choice for Density Weighted Averages," *Journal of Econometrics*, Vol 75, 291-316
- RUPPERT, D. AND WAND, M.P., (1994), "Multivariate locally weighted least squares regression," *The Annals of Statistics*, 22, 1346-1370
- SHADISH, W., T. CAMPBELL AND D. COOK, (2002), *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton and Mifflin, Boston.
- STONE, C., (1982), "Optimal global rates of convergence for nonparametric regression," *The Annals of Statistics*, 10, 1040-1053
- THISTLEWAITE, D., AND CAMPBELL, D., (1960), "Regression-discontinuity analysis: an alternative to the ex-post facto experiment," *Journal of Educational Psychology*, 51, 309-317.
- VAN DER KLAUW, W., (2008), "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22(2): 219-245.
- WAND, M. AND M. JONES, (1994), *Kernel Smoothing*, Chapman and Hall.

Table 1: RD ESTIMATES AND BANDWIDTHS FOR LEE DATA

Procedure	$h$	$\hat{\tau}_{\text{SRD}}$	(s.e.)
$\hat{h}_{\text{opt}}$	0.3005	0.0801	0.0083
no regularization	0.3042	0.0802	0.0082
$f(c)$ estimated using normal kernel	0.3004	0.0801	0.0083
third order polynomial separate on left and right	0.2847	0.0795	0.0085
homoskedastic variance	0.3006	0.0801	0.0083
uniform kernel	0.4721	0.0730	0.0098
Desjardin-McCall	0.3105	0.0804	0.0081
Ludwig-Miller cross-validation ( $\delta = 0.5$ )	0.3250	0.0810	0.0080
Linear	global	0.1182	0.0056
Quadratic	global	0.0519	0.0080
Cubic	global	0.1115	0.0105
Quartic	global	0.0766	0.0131
Quintic	global	0.0433	0.0157

Table 2: SIMULATIONS, 5,000 REPLICATIONS

	$\hat{h}$		$\hat{\tau}_{\text{SRD}}$	
	Mean	Std	Bias	RMSE
<u>Lee Design</u>				
$h_{\text{opt}}$ (infeasible)	0.166	0.000	0.019	0.062
$\hat{h}_{\text{opt}}$	0.538	0.094	0.039	0.053
no regularization	0.724	0.650	0.038	0.052
$f(c)$ estimated using normal kernel	0.538	0.094	0.039	0.053
third order polynomial separate on left and right	0.395	0.055	0.042	0.058
homoskedastic variance	0.536	0.094	0.039	0.054
uniform kernel	0.845	0.148	0.042	0.061
Desjardin-McCall	0.551	0.131	0.039	0.052
Ludwig-Miller cross-validation ( $\delta = 0.5$ )	0.405	0.071	0.039	0.056
Linear		global	0.049	0.057
Quadratic		global	-0.018	0.043
Cubic		global	0.089	0.102
Quartic		global	0.029	0.069
Quintic		global	0.003	0.076
<u>Quadratic Design</u>				
$h_{\text{opt}}$ (infeasible)	0.371	0	0.001	0.041
$\hat{h}_{\text{opt}}$	0.452	0.098	0.012	0.039
no regularization	0.488	0.314	0.018	0.049
$f(c)$ estimated using normal kernel	0.452	0.098	0.012	0.039
third order polynomial separate on left and right	0.410	0.080	0.008	0.040
homoskedastic variance	0.450	0.097	0.011	0.039
uniform kernel	0.709	0.154	-0.044	0.071
Desjardin-McCall	0.227	0.010	-0.002	0.051
Ludwig-Miller cross-validation ( $\delta = 0.5$ )	0.224	0.024	-0.000	0.052
Linear		global	0.246	0.251
Quadratic		global	0.000	0.039
Cubic		global	0.000	0.051
Quartic		global	0.001	0.063
Quintic		global	0.000	0.078

Fig 1: Density for Forcing Variable

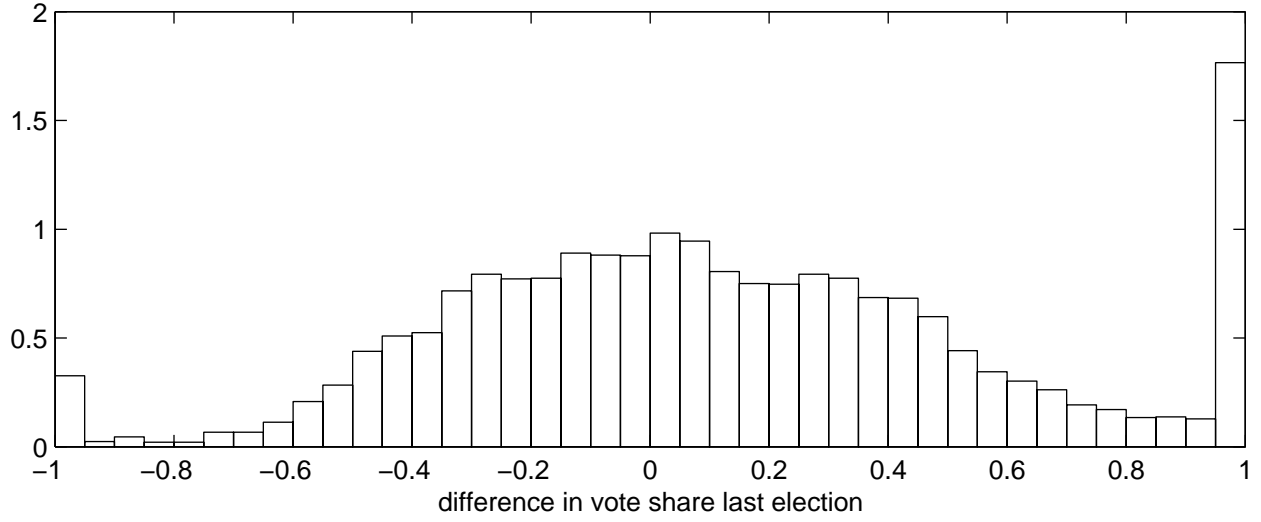


Fig 2: Regression Function for Margin

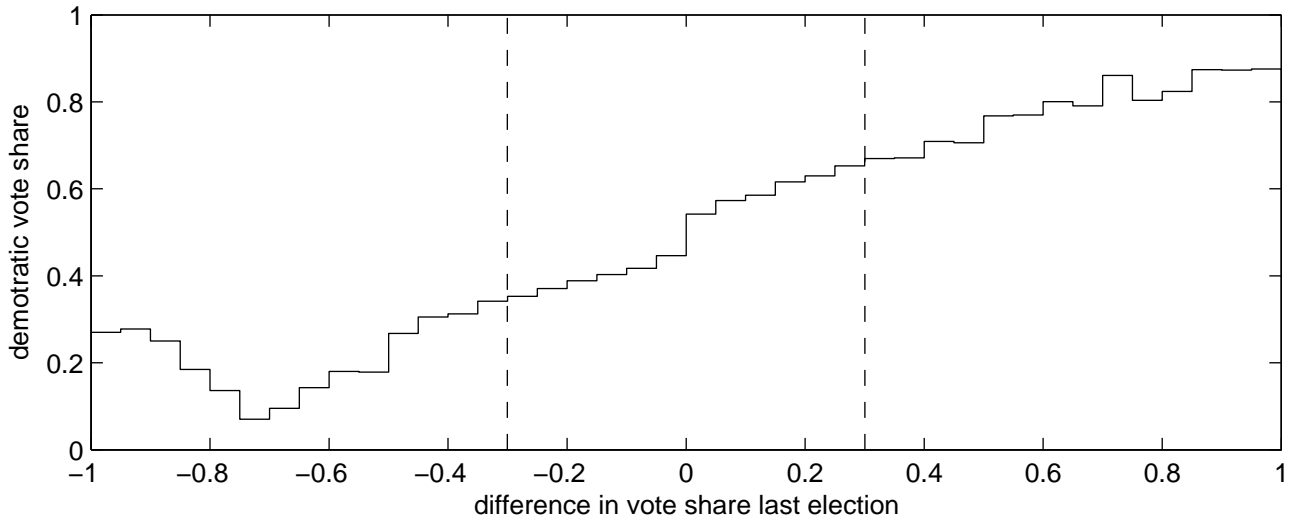


Fig 3: RD Estimates and Confidence Intervals for Lee Data by Bandwidth

