Imbens, Guido W.; Kalyanaraman, Karthik

Working Paper

Optimal bandwidth choice for the regression discontinuity estimator

IZA Discussion Papers, No. 3995

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Imbens, Guido W.; Kalyanaraman, Karthik (2009) : Optimal bandwidth choice for the regression discontinuity estimator, IZA Discussion Papers, No. 3995, Institute for the Study of Labor (IZA), Bonn,
https://nbn-resolving.de/urn:nbn:de:101:1-20090304650

This Version is available at:
https://hdl.handle.net/10419/35551

DISCUSSION PAPER SERIES

IZA DP No. 3995

Optimal Bandwidth Choice for the Regression Discontinuity Estimator

Guido Imbens
Karthik Kalyanaraman

February 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Optimal Bandwidth Choice for the Regression Discontinuity Estimator

Guido Imbens
Harvard University, NBER and IZA

Karthik Kalyanaraman
Harvard University

Discussion Paper No. 3995
February 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3995
February 2009

ABSTRACT

Optimal Bandwidth Choice for the Regression Discontinuity Estimator*

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). Investigation of an expected-squared-error-loss criterion reveals the need for regularization. We propose an optimal, data dependent, bandwidth choice rule. We illustrate the proposed bandwidth choice using data previously analyzed by Lee (2008), as well as in a simulation study based on this data set. The simulations suggest that the proposed rule performs well.

JEL Classification: C14

Keywords: optimal bandwidth selection, local linear regression, regression discontinuity designs

Corresponding author:

Guido Imbens
Department of Economics
Harvard University
1805 Cambridge Street
Cambridge, MA 02138
USA
E-mail: imbens@harvard.edu

* 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, and Doug Miller for comments.

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), and incumbency on reelection (Lee, 2008). 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. Since the focus is solely on the change in the value of the regression function at the threshold, standard plug-in methods and cross-validation methods, which choose a bandwidth that is optimal for estimating the regression function over the entire support, do not yield an optimal bandwidth here. The two contributions of this paper are (i), the derivation of the optimal bandwidth for this setting, and (ii), a data-dependent method for choosing the bandwidth that is asymptotically optimal.[1] 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, $i = 1, \ldots, N$, 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

[1] Software for implementing this bandwidth rule is available on the website http://www.economics.harvard.edu/faculty/imbens/imbens.html. This is at the moment limited to Matlab. In the near future a STATA version will also be available.

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. 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:

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

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

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

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

We focus on estimating τ_{RD} 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

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

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

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

where

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

3 Error Criterion and Infeasible Optimal Bandwidth Choice

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) for a textbook discussion of cross-validation 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.[2] In other words, it attempts to choose the bandwidth to minimize an approximation to the mean integrated squared error criterion (MISE),

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

This criterion is not directly relevant for the problem at hand: we wish to choose a bandwidth that is optimal for estimating τ_{RD}. 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

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

[2] 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 4.4.

Let h^* be the optimal bandwidth that minimizes this criterion:

$$h^* = \arg\min \mathrm{MSE}(h). \tag{3.3}$$

This bandwidth is obviously not feasible, and so we will use an approximations to this oracle bandwidth. The first part of the approximation is that we focus on values of h close to zero, at least asymptotically. In principle, for a specific regression function in combination with a specific distribution for the forcing variable it may well be that the optimal bandwidth does not converge to zero with the sample size. 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.

The next step is to derive an asymptotic expansion of (4.16). 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 $i = 1, \ldots, N$, are independent and identically distributed.*

Assumption 3.2: *The marginal distribution of the forcing variable X_i, denoted $f(\cdot)$, is right and left continuous at the discontinuity, c, with limits $\lim_{x \downarrow c} f(x) = f_+(c) > 0$ and $\lim_{x \uparrow c} f(x) = f_-(c) > 0$ respectively.*

Assumption 3.3: *The conditional mean $m(x) = \mathbb{E}[Y_i | X_i = x]$ has $p \geq$ continuous derivatives almost everywhere. The right and left limits of the k^{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 smooth and has compact support.*

Assumption 3.5: *The conditional variance function $\sigma^2(x) = \mathrm{Var}(Y_i | X_i = x)$ is bounded everywhere, 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:

$$\mathrm{AMSE}(h) = C_1 \cdot h^4 \cdot \left(m_+^{(2)}(c) - m_-^{(2)}(c) \right)^2 + \frac{C_2}{N \cdot h} \cdot \left(\frac{\sigma_+^2(c)}{f_+(c)} + \frac{\sigma_-^2(c)}{f_-(c)} \right). \tag{3.4}$$

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

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

where

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

The first term in (3.4) 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 we will not rely on Assumption 3.6 for consistency of the estimator for the average treatment effect.

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

$$h_{\text{opt}} = \arg\min_h \text{AMSE}(h) = C_K \cdot \left(\frac{\frac{\sigma_+^2(c)}{f_+(c)} + \frac{\sigma_-^2(c)}{f_-(c)}}{\left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2}\right)^{1/5} \cdot N^{-1/5}, \tag{3.5}$$

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_{opt}, given in (4.13) by non-parametric estimators. We make three modifications 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.

The first two modification involve estimating a single density $f(c)$ and a single conditional variance $\sigma^2(c)$, rather than allowing the density and conditional variance functions to have discontinuities at the threshold. For the density the motivation is largely the concern that if the density is discontinuous at the threshold, the validity of the design is typically questioned. (In fact, this is the basis for the test proposed by McCrary, 2008.) For the use of a single conditional variance the motivation is largely that in practice the degree of heteroskedasticity tends to be modest, and so any estimated differences will tend to be largely due to uncertainty in the estimates rather than actual differences. These two modifications suggest using

$$\tilde{h}_{\text{opt}} = C_K \cdot \left(\frac{2 \cdot \hat{\sigma}^2(c)/\hat{f}(c)}{\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c)\right)^2} \right)^{1/5} \cdot N^{-1/5}. \tag{4.6}$$

We introduce one more modification. The motivation for the third modification is the concern 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 (4.6). This regularization term will be choosen 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 τ_{RD}, 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

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

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 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 $\sigma^2(c)$ this leads to

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

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

$$\hat{h}_{\text{opt}} = C_K \cdot \left(\frac{2 \cdot \hat{\sigma}^2(c) / \hat{f}(c)}{\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + (\hat{r}_+ + \hat{r}_-)} \right)^{1/5} \cdot N^{-1/5}, \tag{4.7}$$

To operationalize this proposed bandwidth, we need specific estimators $\hat{f}(c)$, $\hat{\sigma}^2(c)$, $\hat{m}_-^{(2)}(c)$, 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 $\mathbf{Y} = (Y, , \ldots, Y_N)$, the forcing variable $\mathbf{X} = (X_1, \ldots, X_N)$ 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 $K(u) = 1_{|u| \le 1} \cdot (1 - |u|)$, although the algoritm 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 variance at the threshold, $\sigma^2(c)$, 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 (4.7), $\hat{r}_-$ 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 variance $\sigma^2(c)$

First calculate the sample variance of the forcing variable, $S_X^2 = \sum \left(X_i - \overline{X}\right)^2/(N-1)$. 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 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

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

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

Now estimate the density of X_i at c as

$$\hat{f}_X(c) = \frac{N_{h_1,-} + N_{h_1,+}}{N \cdot h_1}, \tag{4.8}$$

and estimate the conditional variance of Y_i given $X_i = x$, at $x = c$, as

$$\hat{\sigma}^2(c) = \frac{1}{N_{h_1,-} + N_{h_1,+}} \left(\sum_{i:c-h_1 \leq X_i < c} \left(Y_i - \overline{Y}_{h_1,-}\right)^2 + \sum_{i:c \leq X_i < c+h_1} \left(Y_i - \overline{Y}_{h_1,+}\right)^2 \right). \tag{4.9}$$

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.

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. First, calculate the median of X_i among the observations with $X_i \geq c$, call this median($\mathbf{X}_+$), and the same for the median of X_i among the observations with $X_i < c$, call this median($\mathbf{X}_-$). To be precise, if the number of observations with $X_i \geq 0$ is even, we define the median to be the average of the middle two observations. Temporarily discard the observations with $X_i <$ median($\mathbf{X}_-$), and the observations with $X_i >$

median($\mathbf{X}_+$). Now fit a third order polynomial to the remaining data, including an indicator for $X_i \geq 0$. Thus, estimate the regression function

$$Y_i = \gamma_0 + \gamma_1 \cdot 1_{X_i \geq c} + \gamma_2 \cdot (X_i - c) + \gamma_3 \cdot (X_i - c)^2 + \gamma_4 \cdot (X_i - c)^3 + \varepsilon_i, \tag{4.10}$$

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 a 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 here to achieve what we ultimately need: a consistent estimate of the constant in the reference bandwidth. Calculate $h_{2,+}$, using the $\hat{\sigma}^2(c)$ and $\hat{f}(c)$ from Step 1, as

$$h_{2,+} = 3.56 \left(\frac{\hat{\sigma}^2(c)}{\hat{f}(c) \max\left(\left(\hat{m}^{(3)}(c)\right)^2, 0.01 \right)} \right)^{1/7} N_+^{-1/7}, \tag{4.11}$$

and

$$h_{2,-} = 3.56 \left(\frac{\hat{\sigma}^2(c)}{\hat{f}(c) \max\left(\left(\hat{m}^{(3)}(c)\right)^2, 0.01 \right)} \right)^{1/7} N_-^{-1/7}.$$

The motivation for taking the maximum of $(\hat{m}^{(3)}(c))^2$ and 0.01 is to avoid problems if $m_3 = 6 \cdot \text{plim}(\hat{\gamma}_4)$ is in fact equal to zero. In practice this is unlikely to be a problem, and for the formal arguments the constant 0.01 can be replaced by any positive number. Without this constant, $h_{2,+}$ is in fact an estimate 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. I.e. temporarily discard the observations other than the $N_{2,+}$ observations with $c \leq X_i \leq c + h_{2,+}$. Label the new data $\hat{\mathbf{Y}}_+ = (Y_1, \ldots, Y_{N_{2,+}})$ and $\hat{\mathbf{X}}_+ = (X_1, \ldots, X_{N_{2,+}})$ 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 ι is a column vector of ones, and $\mathbf{T}_j' = \left((X_1 - c)^j, \ldots, (X_{N_{2,+}} - c)^j\right)$, 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 $m_+^{(2)}(c)$. For $\hat{m}_-^{(2)}(c)$ 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

$$\hat{r}_+ = \frac{720 \cdot \hat{\sigma}^2(c)}{N_{2,+} \cdot h_{2,+}^4}, \qquad \text{and} \quad \hat{r}_- = \frac{720 \cdot \hat{\sigma}^2(c)}{N_{2,-} \cdot h_{2,-}^4}. \tag{4.12}$$

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

$$\hat{h}_{\text{opt}} = C_K \cdot \left(\frac{2\hat{\sigma}^2(c)}{\hat{f}(c) \cdot \left(\left(\hat{m}_+^{(2)}(c) - \hat{m}_-^{(2)}(c) \right)^2 + (\hat{r}_+ + \hat{r}_-) \right)} \right)^{1/5} \cdot N^{-1/5}. \tag{4.13}$$

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 $C_K \approx 3.4375$.

G8iven $\hat{h}_{\text{opt}}$, we estimate τ_{RD} as

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

where $\hat{m}_h(x)$ is 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*)

$$\hat{\tau}_{RD} - \tau_{RD} = O_p\left(N^{-12/35}\right), \tag{4.14}$$

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

$$\hat{\tau}_{RD} - \tau_{RD} = O_p\left(N^{-4/5}\right), \tag{4.15}$$

(*iii*)

$$\frac{\hat{h}_{\text{opt}} - h_{\text{opt}}}{h_{\text{opt}}} = o_p(1), \tag{4.16}$$

and (*iv*)*:*

$$\frac{\text{MSE}(\hat{h}_{\text{opt}}) - \text{MSE}(h_{\text{opt}})}{\text{MSE}(h_{\text{opt}})} = o_p(1). \tag{4.17}$$

Note that if Assumption 3.6 fails, the convergence rate for $\hat{\tau}_{\text{RD}}$ is slower. This is somewhat counterintuitive. The conventional intuition goes as follows. 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 becomes $N^{-1/7}$, and the optimal rate for the MSE becomes $N^{-6/7}$ and thus the optimal rate for $\hat{\tau}_{\text{RD}} - \tau_{\text{RD}}$ becomes $N^{-3/7}$, better than the $N^{-2/3}$ we have when Assumption 3.6 holds. The reason this does not show up in the theorem is that the optimal bandwidth does not adapt to the vanishing of the difference in second derivatives. If Assumption 3.6 fails, the bandwidth goes to zero as $N^{-4/35}$ (instead of the optimal $N^{-1/7}$), and so the MSE becomes $N^{-24/35}$, leading to $\hat{\tau}_{\text{RD}} - \tau_{\text{RD}} = O_p(N^{-12/35})$, slower than the optimal rate of $N^{-3/7}$, and even slower than the rate we achieve when Assumption 3.6 holds ($N^{-2/5}$). One could modify the regularization term to take account of this, but in practice it is unlikely to make a difference.

4.4 Ludwig-Miller Cross-validation

In this section we briefly describe the cross-validation 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 δ-th quantile of the X_i among the subsample of observations with $X_i < c$ and $X_i \geq c$ respectively, so that

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

and

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

Now the LM cross-validation criterion we use is of the form:

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

(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 cross-validation. By using a value for δ close to zero, we only use observations close to the threshold to evaluate the cross-valdiation criterion. The only concern is that by using too small value of δ, we may not get a precisely estimated cross-validation 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 δ is unlikely to lead to an optimal bandwidth in general. Moreover, the criterion focuses implicitly on minimizing a criterion more akin to $\mathbb{E}\left[(\hat{\mu}_+ - \mu_+)^2 - (\hat{\mu}_- - \mu_-)^2\right]$, (with the errors in estimating μ_- and μ_+ squared before adding them up, rather than rather than $\text{MSE}(h) = \mathbb{E}[((\hat{\mu}_+ - \mu_+) - (\hat{\mu}_- - \mu_-))^2]$ in (4.16), where the error in the difference $\mu_+ - \mu_-$ is squared. As a result t even letting $\delta \to 0$ with the sample size in the cross-validation procecure is unlikely to result in an optimal bandwidth.

5 Extensions

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. In an important theoretical paper HTV discussion identification in this setting. The focus is on the ratio

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

In principle 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. Thus, in principle there are four bandwidth choices to be made. However just as we argued for a single bandwidth in the SRD setting one might make the same argument here though with less force. We follow the suggestion however of Imbens and Lemieux (2008): i.e. use the algorithm above and estimate two optimal bandwidths, one for the outcome regression, say $\hat{h}^Y_{\text{opt}}$, and one for the treatment regression, $\hat{h}^W_{\text{opt}}$. It might be appealing conceptually to use the same bandwidth for both $\hat{h}_{\text{FRD,opt}}$, and one could simply take the optimal bandwidth for the outcome variable: $\hat{h}_{\text{FRD,opt}} = \hat{h}^Y_{\text{opt}}$, given that the discontinuity in the treatment regression is typically precisely estimated.

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. 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.

6 An Illustration and Some Simulations

6.1 Data

To illustrate the implementation of these methods we use data 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.

6.2 IK algorithm on 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 $\overline{Y}_{h_1,-} = 0.4219$ and a sample variance of $S^2_{Y,h_1,-} = 0.1047^2$, and $N_{h_1,+} = 862$ units with values for X_i in the interval $[0, h_1]$, with an average outcome of $\overline{Y}_{h_1,+} = 0.5643$ and a sample variance of $S^2_{Y,h_1,+} = 0.1202^2$. 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

$$\hat{\sigma}^2(0) = \frac{(N_{h_1,-} - 1) \cdot S^2_{Y,h_1,-} + (N_{h_1,+} - 1) \cdot S^2_{Y,h_1,+}}{N_{h_1,-} + N_{h_1,+}} = 0.1128^2.$$

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. We estimate this global cubic regression function by dropping observations with covariate values below the median of the covariate for observations with covariate values below the threshold, and dropping observations with covariate values above the median of the covariate for observations with covariate values above the threshold. For the 2740 (3818) observations with $X_i < 0$ ($X_i > 0$), the median of the forcing variable is -0.2485 (0.3523). Next, we estimate, using the data with $X_i \in [-0.2485, 0.3523]$, the polynomial regression function of order three, with a jump at the threshold:

$$Y_i = \gamma_0 + \gamma_1 \cdot X_i + \gamma_2 \cdot X_i^2 + \gamma_3 \cdot X_i^3 + \gamma_4 \cdot 1_{X_i \geq 0} + \varepsilon_i.$$

The least squares estimate for γ_3 is $\hat{\gamma}_3 = -0.9102$, and thus the third derivative at zero is estimated as $\hat{m}^{(3)}(0) = 6 \cdot \hat{\gamma}_3 = -5.4611$. This leads to the two bandwidths

$$h_{2,+} = 3.56 \cdot \left(\frac{\hat{\sigma}^2(0)}{\hat{f}(0) \cdot \max\left(\left(\hat{m}^{(3)}(0) \right)^2, 0.01 \right)} \right)^{1/7} \cdot N_+^{-1/7} = 0.3674, \qquad \text{and } \; h_{2,-} = 0.3852.$$

The two pilot bandwidths are used to fit two quadratics. The quadratic to the right of 0 is fitted on $[0, 0.3674]$, yielding $\hat{m}^{(2)}_+(0) = -0.5233$ and the quadratic to the left is fitted on [-0.3852, 0] yielding $\hat{m}^{(2)}_-(0) = 0.4904$.

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

$$\hat{r}_+ = \frac{720 \cdot \hat{\sigma}^2(0)}{N_{2,+} h^4_{2,+}} = \frac{720 \cdot 0.1128^2}{1983 \cdot 0.3674^4} = 0.2634 \qquad \text{and } \; \hat{r}_- = \frac{720 \cdot \hat{\sigma}^2(0)}{N_{2,-} h^4_{2,-}} = 0.3036.$$

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

$$\hat{h}_{\text{opt}} = C_K \left(\frac{2 \cdot \hat{\sigma}^2(0)}{\hat{f}(0) \cdot \left[\left(\hat{m}_+^{(2)}(0) - \hat{m}_-^{(2)}(0) \right)^2 + (\hat{r}_+ + \hat{r}_-) \right]} \right)^{1/5} N^{-1/5} = 0.2649.$$

Without the regularization the bandwidth would be $\tilde{h}_{\text{opt}} = 0.2892$.

6.3 Six Estimates of the Effect of Incumbency for the Lee Data

Here we calculate six estimates of the ultimate object of interest, the size of the discontinuity in $m(x)$ at zero. The first four are based on local linear estimation with the edge kernel, and the bandwidth choosen optimally ($\hat{h}_{\text{opt}}$), optimally without regularization ($\tilde{h}_{\text{opt}}$), or cross-validation ($\hat{h}_{\text{cv}}$, with $\delta = 0.5$). For comparison we report estimates on global least squares regression of polynomial regression functions on either side of the threshold, using a first, second and third order polynomial. The point estimates and standard errors are presented in Table 1. To investigate the sensitivity 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 0.5. The solid vertical line indicates the optimal bandwidth ($\hat{h}_{\text{opt}} = 0.2649$), and the dashed vertical line the LM cross-validation bandwidth $h_{\text{cv}} = 0.2231$, based on $\delta = 0.5$. For the LM cross-validation, Figure 4 shows the criterion function for $\delta = 0.5$.

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 two designs, Designs I, and II, and two sample sizes, $N = 100$, and $N = 500$. In all cases we use normal disturbances, with standard deviation equal to the standard deviation of the outcome in the Lee data set, $S_Y = 0.2411$. The density of the forcing variable is that of $2 \cdot Z_i - 1$, were Z_i has a Beta distribution with parameters $\alpha = 2$ and $\beta = 4$. The two designs differ in the population regression function. The first design is motivated by the configuration of the Lee data. 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

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

In the second design the regression function is quadratic:

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

implying the data generating process is close to the point where the bandwidth h_{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.

We report results for five bandwidth choices in Table 2. The first two are infeasible: the optimal bandwith h^*, which minimizes the expected squared error, and h_{opt}, which minimizes the asymptotic approximation to the expected squared error In addition we report the results based on the proposed bandwidth $\hat{h}_{\text{opt}}$, the non-regularized bandwidth $\tilde{h}_{\text{opt}}$, and the Ludwig-Miller cross-validation bandwidth $\hat{h}_{\text{cv}}$. In Table 2 we present for the two designs, for the two sample sizes and the five bandwidth choices 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 τ. In the design inspired by the Lee data the optimal bandwidth h^* is quite high. This bandwidth choice outperforms the feasible ones in terms of RMSE quite substantially. Among the feasible bandwidth choices the unregularized bandwidth choice performs slightly better in terms of RMSE than the regularized one: both are substantially better than cross-validation. The slight improvement of the unregularized bandwidth comes at the expense of substantially more variation in the bandwidth choice across simulations: the standard deviation of the bandwidth choice is higher by a factor four. This remains true even in the larger sample.

In the second design the regularized bandwidth choice substantially outperforms the other feasible bandwidth choices. It has lower RMSE and substantially less variation.

7 Conclusion

In this paper we propose a fully data-driven, asymptotically optimal bandwidth choice for regression discontinuity settings. This bandwidth choice can provide an objective starting point for assessing sensitivity to bandwidth choice in such 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 global measures, not tailored to the specific features of the regression discontinuity setting. We compare our proposed bandwidth selection procedure to the cross-validation 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. We find that our proposed method works well in

realistic settings 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 $\sigma^2(x) = \mathrm{Var}(Y_i|X_i = x)$ is bounded, and continuous at $x = 0$, (v), we have a kernel $K : \mathbb{R}^+ \mapsto \mathbb{R}$, with $K(u) = 0$ for $u \geq \overline{u}$, and $\int_0^{\overline{u}} K(u)du = 1$, and define $K_h(u) = K(u/h)/h$. Define $\mu = m(0)$, and

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

Then:

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

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

and

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

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

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

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

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

The variance of F_j can be bounded by

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

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

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

Hence

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

□

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,*

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

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

Proof of Lemma A.1: Define $R = [\iota \quad X]$, where ι 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 $B = \mathbb{E}[\hat{m}(0)|X_1,\ldots,X_N] - m(0)$. Note that $\mathbb{E}(\hat{m}(0)|X) = e_1'(R'WR)^{-1}R'WM$ where $M = (m(X_1),\ldots,m(X_N))'$. Let $m^{(k)}(x)$ denote the kth 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

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

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 ith 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

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

Thus

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

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).

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

where Σ 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. □

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

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

and

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

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

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

Conditional on $X_1,\ldots,X_N$ 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

$$\begin{aligned} \mathbb{E}\left[\,(\hat{\tau}_{\text{RD}} - \tau_{\text{RD}})^2\middle|\, X_1,\ldots,X_N\right] &= \mathbb{E}\left[\,(\hat{\mu}_+ - \hat{\mu}_- - (\hat{\mu}_+ - \hat{\mu}_-))^2\middle|\, X_1,\ldots,X_N\right] \\ &= \mathbb{E}\left[\,(\hat{\mu}_+ - \mu_+)^2\middle|\, X_1,\ldots,X_N\right] + \mathbb{E}\left[\,(\hat{\mu}_- - \mu_-)^2\middle|\, X_1,\ldots,X_N\right] \end{aligned}$$

$$-2 \cdot \mathbb{E}\left[\hat{\mu}_+ - \mu_+ \middle| X_1, \ldots, X_N\right] \cdot \mathbb{E}\left[\hat{\mu}_- - \mu_- \middle| X_1, \ldots, X_N\right]$$
$$= C_1 \cdot h^4 \cdot \left(m_+^{(2)}(c) - m_-^{(2)}(c)\right)^2 + \frac{C_2}{N \cdot h} \cdot \left(\frac{\sigma_+^2(c)}{f_+(c)} + \frac{\sigma_-^2(c)}{f_-(c)}\right). + o_p\left(h^4 + \frac{1}{N \cdot h}\right),$$

proving the first result in Lemma 3.1.
For the second part of Lemma 3.1, solve

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

which leads to

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

□

Motivation for the Bandwidth Choice in Equation (4.11) 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 ν-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 $K(t) = 1_{0 \le t \le 1}$, 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

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

where

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

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

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

and

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

First we show that the various estimates of the functionals in $\hat{C}_{opt}$, $\hat{\sigma}^2(c)$, $\hat{f}(c)$, $\hat{m}_+^{(2)}(c)$ and $\hat{m}_-^{(2)}(c)$ converge to their counterparts in C_{opt}, $\sigma^2(c)$, $f(c)$, $m_+^{(2)}(c)$ 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 $\hat{\sigma}^2(c)$. 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 $\sigma_-^2(c) = \sigma_+^2(c) = \sigma^2(c)$. If the two limiting variances are different, $\hat{\sigma}^2(c)$ is consistent for $(\sigma_-^2(c) \cdot f_-(c) + \sigma_+^2(c) \cdot f_+(c))/(f_+(c) + f_-(c))$.
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

$$h_{2,+} = o_p(1) \qquad \text{and} \quad \left(N h_{2,+}^5\right)^{-1} = o_p(1). \tag{A.4}$$

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_{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_{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

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

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

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

Therefore

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

and

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

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

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

Substituting in, we have

$$N^{4/5} \cdot \left(\text{AMSE}\left(\hat{h}_{\text{opt}}\right) - \text{AMSE}\left(h_{\text{opt}}\right) \right)$$

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

$$= o_p(1),$$

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 $\hat{C}_{\text{opt}} \to C_{\text{opt}}$, 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 $\hat{\tau}_{\text{RD}} - \tau_{\text{RD}} = O_p(N^{-2/5}$.
Finally, consider (i). If Assumption 3.6 holds, then $\hat{\tau}_{\text{RD}} - \tau_{\text{RD}} = O_p(N^{-2/5})$, 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 $\hat{\tau}_{\text{RD}} - \tau_{\text{RD}} = O_p(N^{-12/35})$. □

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.

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 Klaauw, W., (2001), "Regression discontinuity," *Econometrica*, 69(1), 201-209.

Imbens, G.W. 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, Cross-validation and Generalized Cross-validation: 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 Klaauw, W., (2008), "Regression–Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22(2): 219-245.

Table 1: Lee Data: RD estimates and bandwidths

Procedure	h	$\hat{\tau}_{\text{RD}}$	(s.e.)
$\hat{h}_{\text{opt}}$	0.2649	0.0782	0.0083
$\tilde{h}_{\text{opt}}$ (no regularization)	0.2892	0.0798	0.0079
$\hat{h}_{\text{cv}}$	0.2231	0.0754	0.0090
Linear	global	0.1182	0.0065
Quadratic	global	0.0519	0.0088
Cubic	global	0.1115	0.0136

Table 2: Simulation results

	N=100				N=500			
	$\hat{h}$		$\hat{\tau}_{\text{RD}}$		$\hat{h}$		$\hat{\tau}_{\text{RD}}$	
	MEAN	STD	BIAS	RMSE	MEAN	STD	BIAS	RMSE
Design I								
h^*	1.16	–	0.03	0.12	0.89	–	0.02	0.06
h_{opt}	0.28	–	0.03	0.20	0.21	–	0.03	0.10
$\hat{h}_{\text{opt}}$	0.43	0.12	0.04	0.18	0.44	0.12	0.04	0.08
$\tilde{h}_{\text{opt}}$	0.65	0.48	0.04	0.17	0.69	0.63	0.04	0.08
$\hat{h}_{\text{cv}}$	0.90	0.81	0.03	0.22	0.41	0.45	0.03	0.10
Design II								
h^*	0.57	–	0.04	0.15	0.49	–	0.01	0.07
h_{opt}	0.74	–	0.11	0.18	0.54	–	0.03	0.07
$\hat{h}_{\text{opt}}$	0.43	0.12	-0.00	0.18	0.42	0.10	-0.01	0.08
$\hat{h}_{\text{noreg}}$	0.65	0.55	0.06	0.21	0.62	0.45	0.05	0.14
$\hat{h}_{\text{cv}}$	0.23	0.09	-0.03	0.25	0.20	0.03	-0.04	0.11

Fig 1: Density for Forcing Variable

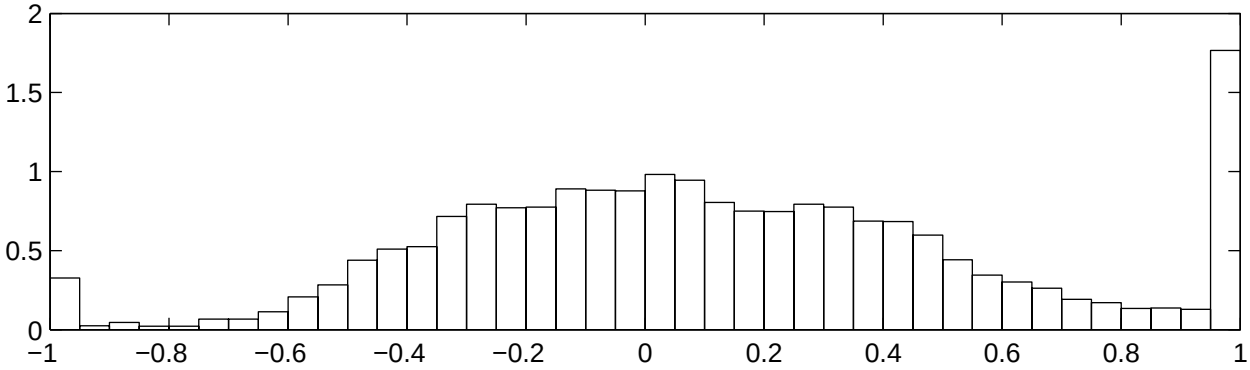

Fig 2: Regression Function for Margin

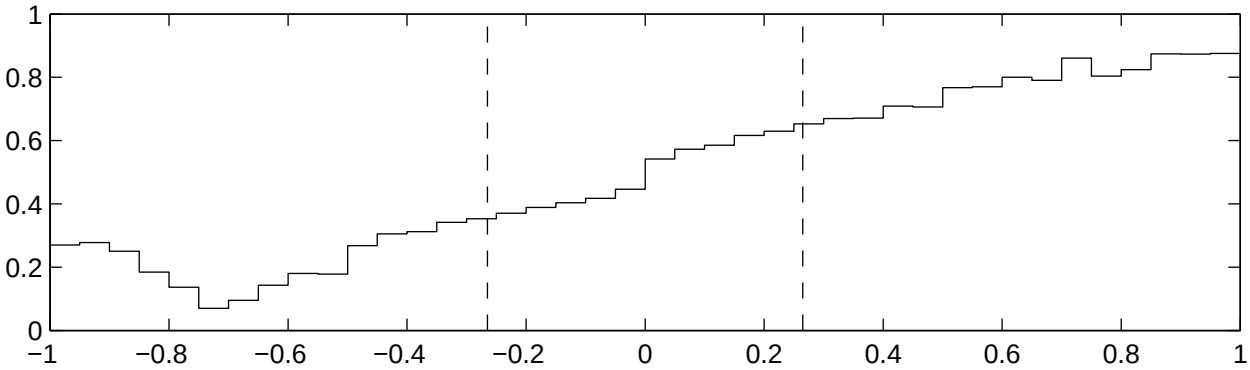

Fig 3: RD Estimates and Confidence Intervals by Bandwidth

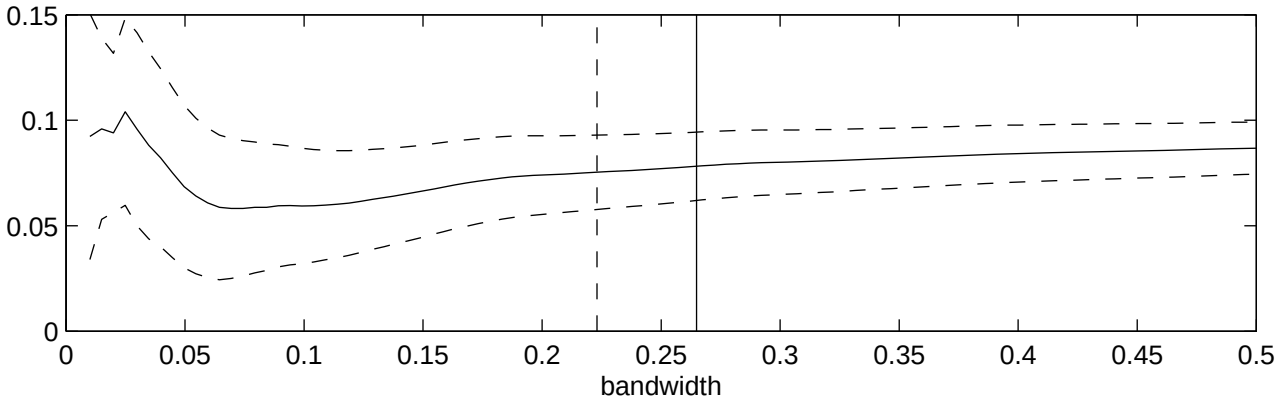

Fig 4: Ludwig−Miller Crossvalidation (delta=0.5)