Theory and Practice of Inference in Regression Discontinuity: A Fixed-Bandwidth Asymptotics Approach.*

Otávio Bartalotti†

Tulane University

October 28, 2013

Abstract

In regression discontinuity design (RD), researchers use bandwidths around the discontinuity. For a given bandwidth, one can estimate asymptotic variance based on the assumption that the bandwidth shrinks to zero as sample size increases (the traditional approach) or, alternatively, that the bandwidth is fixed. The main theoretical results for RD rely on the former, while most applications in the literature treat the estimates as parametric. This paper develops the "fixed-bandwidth" alternative asymptotic theory for local polynomial estimators, bridging the gap between theorists and practitioners and shedding light on implicit assumptions on both approaches. The fixed-bandwidth approach provides alternative formulas, i.e. alternative approximations, for the bias and variance of RD estimators. Simulations indicate that fixed-bandwidth approximations are usually better than traditional approximations, and improvements are nontrivial when there is heteroskedasticity. When there is no heteroskedasticity, both approximations are shown to be equivalent under some additional mild conditions. Feasible estimators of fixed-bandwidth standard errors are easy to implement and improve coverage of confidence intervals compared to the traditional approach, especially in the presence of heteroskedasticity. Fixed-bandwidth approximations are akin to treating RD estimators as *locally* parametric, providing theoretical justification for the common empirical practice of using heteroskedasticity-robust standard errors in RD settings.

1   Introduction

Regression discontinuity (RD) designs have been propelled to the spotlight of economic analysis in recent years[1], especially in policy and treatment evaluation literatures, as a form of estimating treatment effects in a

---

[1]Lee and Lemieux (2009), in a broad review of the RD literature, compile a list of more than 60 papers that apply RD design to many different contexts. The overwhelming majority of the papers have been published in the last decade.

non-experimental setting. In this design, treatment is assigned to observations based on values of an observed covariate, with the probability of receiving treatment jumping discontinuously at a known threshold. RD's appeal comes from the relative weak assumptions necessary for non-parametric identification of treatment effects and inference, which rely on RD's "quasi-experimental" characteristics.

Nonparametric local polynomial estimators have been widely used in recent RD literature, being the most common choice in empirical and theoretical work due to its ease of implementation and good asymptotic properties. These kernel-based estimators rely on fitting a polynomial function to a range of the data, which size is determined by a bandwidth, h, just above and below the cutoff. The implementation of such estimators depends on the choice of the bandwidth, which can be aided by several bandwidth selectors available in the literature. Nevertheless, bandwidths selected "typically lead to bandwidth choices that are too large for the usual distributional assumptions to be valid," as pointed out by Calonico et al. (2012).

The conventional "small-h" approach to obtain the asymptotic properties of RD estimates relies on a nonparametric approximation based on the idea that the bandwidth shrinks towards zero as the sample size grows. This guarantees nonparametric identification of the parameter of interest under mild conditions. In practice, to obtain an estimate of the treatment effect and perform inference, the empiricist is required to use a particular value of the bandwidth which is necessarily greater than zero. In some cases researchers may use very large bandwidths, for example when the running variable is discrete, so that the bandwidth must be discretely positive; or even with a continuous running variable, sample sizes often are small enough that precision concerns impel researchers to increase its effective sample size by using a relatively large bandwidth. Hence, even though asymptotic theory requires that $h \to 0$, in practice h is fixed.

From that fact arises a disconnect between theory and practice for inference in RD since most practitioners use the usual parametric inference methods, even though asymptotic theory for RD is based on nonparametric, small-h asymptotic approximations.

I provide a set of conditions under which the use of parametric tools of inference would be locally valid and develop an alternative asymptotic theory for the RD treatment effects estimator in which the bandwidth is fixed. This "fixed-h" approximation to the estimator's asymptotic distribution better incorporates the bandwidth size used by the researcher, refines the conventional approximation and leads naturally to standard error formulas that improve test coverage in simulations. This approach is akin to treat the model as parametric in the neighborhood of the cutoff.[2]

The intuition fits nicely with current practice in applied work, which has focused on the usual Huber-Eicker-White heteroskedasticity robust standard errors, essentially treating the estimates as locally parametric. Hence, this paper provides a theoretical framework that justifies such choice, bridging the gap between theory and practice regarding inference in RD. It also provides evidence that treating these nonparametric estimates as locally parametric can improve inference and successfully adjust standard errors to reflect the bandwidth used by researchers.

---

[2]I am thankful to Matias Cattaneo for helpful comments on this point.

Comparing fixed-h and small-h approximations provides additional clarity about assumptions involved in each approximation. First, when $h \to 0$, the fixed-h asymptotic approximation for the distribution of the parameter of interest reduces to the conventional small-h approximation, indicating that for small bandwidths approximations are similar and any differences in inference should be small. Second, usual small-h standard errors implicitly impose homoskedasticity and constant probability density of the running variable around the discontinuity. The intuition is similar to the one put forth by Fan and Gijbels (1996) and Lee and Card (2008) for the case of discrete running variables. When a bandwidth is used, estimating the conditional expectation of the outcome at the threshold is akin to estimate a parametric polynomial model inside the bandwidth. For larger bandwidths these parametric assumptions become more restrictive as we impose functional form and homoskedasticity to a larger support of the data and, can significantly affect inference performance.

Natural estimators for the asymptotic variance based on fixed-h results are presented and Monte Carlo simulations provide evidence that feasible inference incorporates improvements predicted by the theory, suggesting that theoretical gains in robustness can be translated to practical benefits in applied work. Those variance estimators are analogous to the Huber-Eicker-White heteroskedasticity robust standard errors. The use of such variance estimators in RD designs have been suggested based on intuitive arguments that did not follow directly from the usual nonparametric asymptotic approximations, see for example Lee and Lemieux (2009) and Calonico et al. (2012) among others. By treating the bandwidths as fixed this paper provides theoretical justification for those estimators in the nonparametric context, a practice common in empirical applications. To be clear, the results presented below are not about how to choose a bandwidth, but about how to perform inference appropriately for any bandwidth. When we treat the estimators as locally parametric our variance estimators become "robust to the choice of bandwidths".

The goal here is to obtain a good approximation to the behavior of estimators in the finite sample sizes practitioners actually use and to obtain clarity about the implicit and explicit assumptions being made when performing estimation and inference. The fixed-h asymptotic approximation delivers a better approximation to the finite-sample behavior of the local polynomial estimator in RD designs than the usual small-h approximation.[3] These results validate current practice in applied work and bridge the gap between theory and practice, clarifying the assumptions necessary for its validity.

This paper contributes to the emerging literature on inference for treatment effects in the context of RD designs. Hahn, Todd and van der Klaauw (1999, 2001) and Lee (2008), presented the conditions for identification of the average treatment effect of interest and its estimation in RD designs. Porter (2003) provided results on the asymptotic properties of the estimators for the treatment effect of interest, obtaining limiting distributions for estimators based on local polynomial regression and partially linear estimation. Calonico et al. (2012) studies alternative asymptotic approximations for the bias-corrected local polynomial RD estimator. Other studies about estimation, inference and bandwidth choice in RD designs include Imbens

---

[3] Hence a previous title to this paper: When the practitioner saved the theorist.

and Kalyanaraman (2012), Cattaneo et al. (2012) among others.

McCrary (2008) studies specification testing. Also, Imbens and Lemieux (2008) and Lee and Lemieux (2009) offer a broad review of the theoretical and applied literature with emphasis on the identification of the parameter of interest and its potential interpretations.

The proposed inferential framework using fixed bandwidths follows a growing literature that recognizes its potential for improvements in inference procedures in nonparametric methods. Notably, Neave (1970), later extended by Hashimzade and Vogelsang (2008), in the context of spectral density estimation, obtains more accurate approximations to the variance of nonparametric spectral estimates by acknowledging that, with a finite sample, the bandwidth used is fixed. The author argues that the assumption equivalent to the bandwidth converging to zero: "(...) is a convenient assumption mathematically in that, in particular, it ensures consistency of the estimates, but it is unrealistic when such results are used as approximations to the finite case(...)"(Neave 1970, p.70). Also, Fan (1998) provides an alternative approximation for goodness-of-fit tests for density function estimates in which the bandwidth used in the test is fixed, obtaining improved approximations to the asymptotic behavior of the test and more appropriate critical values for inference.

The same can be said in the regression discontinuity design. Even though $h \to 0$ is a convenient assumption that guarantees consistency of estimates for the average treatment effect, it will be unrealistic and potentially misleading. It is of theoretical and practical interest to obtain more accurate asymptotic distributions by treating h as fixed so that the theory used for inference is more accurate and aligned with the practice of applied economists.

Monte Carlo simulations indicate that asymptotic approximations based on fixed-h better characterize the behavior of the estimators and provide improved inference about the treatment effect both in theory (section 5.1) and in practice (section 5.2). These improvements are more important when the bandwidth is larger and local heteroskedasticity is present. Feasible tests based on the fixed-h approach obtain better coverage, outperforming small-h as predicted by the theory outlined in section 3, even at relatively small bandwidths.

In the case of the widely used local linear estimator with homoskedastic errors the variance estimators based on small-h asymptotics suggested in the literature produce relatively well behaved tests, performing better than the small-h asymptotic theory would suggest. This emphasizes the importance of the implicit homoskedasticity assumption in the suitability of small-h as a valid approximation in RD designs.

Finally, section 6 provides an empirical application to Lee (2008), exemplifying with actual data the improvements obtained by the proposed fixed-h standard error estimators relative to those based on the usual small-h asymptotics.

2   Model and Estimator

The interest lies in estimating the average treatment effect, τ, of a certain treatment or policy that affects part of a population of interest. As discussed in Porter (2003), Imbens and Lemieux (2008) and Lee and Lemieux (2009), RD designs are closely associated with the treatment effect literature.[4] There are two types of RD designs, sharp and fuzzy, and they differ as to how treatment is assigned to a certain observation and the impact of the discontinuity in its assignment. The main body of the paper focuses on sharp RD. The discussion and extensions for the fuzzy design are presented in the appendix.

2.1   Sharp Regression Discontinuity Design

In the sharp design, the treatment status, D, is a deterministic function of a so called "running" or "forcing" variable, x, such that,

$$d_i = \left\{ \begin{array}{l} 1 \text{ if } x_i \geq \overline{x} \\ 0 \text{ if } x_i < \overline{x} \end{array} \right\}$$

where $\overline{x}$ is the *known* cut-off point. Then, let Y_1 and Y_0 be the potential outcomes corresponding to the two possible treatment assignments. As usual, we cannot observe both potential outcomes, having access only to $Y = dY_1 + (1-d)Y_0$. As described by Hahn, Todd and van der Klaauw (2001) and Porter (2003), under a smoothness assumption that $E\left[Y_j \mid X = x\right]$ is continuous at $\overline{x}$ for $j = 0, 1$, the average treatment effect can be estimated by comparing points just above and just below the discontinuity. The discontinuity in treatment assignment at $\overline{x}$ provides the opportunity for identifying the average treatment effect at the cutoff without any additional parametric functional form restrictions on the conditional expectations of the outcome variable. The average causal effect of the treatment at the discontinuity is (Imbens and Lemieux, 2008)

$$\begin{aligned} \tau_S &\equiv& E\left[Y_1 - Y_0 \mid X = \overline{x}\right] \\ &=& \lim_{x \downarrow \overline{x}} E\left[Y \mid X = x\right] - \lim_{x \uparrow \overline{x}} E\left[Y \mid X = x\right] \end{aligned}$$

where the second equality holds under some smoothness assumptions regarding the conditional expectations (discussed below). The sharp regression discontinuity design uses the discontinuity in the conditional expectation of Y given X to uncover the average treatment effect. If the treatment effect is deemed constant across individuals, τ_S is the effect of treatment for each individual in the population. If we allow the treatment effect to differ among individuals, τ_S is the average treatment effect for individuals at the cutoff. Interestingly, Lee and Lemieux (2009) show that the so-called RD gap obtained by the comparison of observations just above and just below the cutoff can be interpreted as a weighted average treatment effect across all individuals, not only the individuals around the cutoff. In this case each individual would have weights directly proportional to the ex ante likelihood that an individual's realization of X will be close to

---

[4]Angrist and Pischke (2009) provide a simple introduction to the intuition of regression discontinuity.

the threshold. For a comprehensive review of RD designs and their applications and interpretation, see Lee and Lemieux (2009).

2.2 Local Polynomial Estimator

I analyze estimates for the parameter of interest, τ_S or τ_F, obtained by local polynomial estimator, which is the most common in applications. This is partially due to their easy implementation, nice properties and by the fact that local linear estimators have been the focus of several papers that disseminated the technique (Hahn, Todd and van der Klaauw 1999 and 2001, Imbens and Lemieux 2008 and Lee and Lemieux 2009). Fan and Gijbels (1996) also point out that from a theoretical point of view local polynomial estimators are attractive for estimation in the regression discontinuity setting given its nice boundary behavior and by the fact that it relies on the least squares principle which allows one to tap in the large statistical knowledge available.

The order p local polynomial estimator is defined as follows. In the sharp design case, given data $(y_i, x_i)_{i=1,2,\ldots,n}$, let $d_i = 1[x_i \geq \overline{x}]$, $k(\cdot)$ be a kernel function, h denote a bandwidth that controls size of the local neighborhood to be averaged over. Also, define the $p+1 \times 1$ vector $Z(x) = \left(1, \left(\frac{x-\overline{x}}{h}\right), \left(\frac{x-\overline{x}}{h}\right)^2, \ldots, \left(\frac{x-\overline{x}}{h}\right)^p\right)'$ and let $\left(\widehat{\alpha}_{p+}, \widehat{\beta}_{p+}\right)'$ be the solution to the minimization problem:[5]

$$\min_{a,b_1,\ldots,b_p} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left[y_i - a - b_1\left(\frac{x_i - \overline{x}}{h}\right) - \ldots - b_p\left(\frac{x_i - \overline{x}}{h}\right)^p\right]^2$$

and similarly $\left(\widehat{\alpha}_{p-}, \widehat{\beta}_{p-}\right)$ minimizes the same objective function but with $1 - d_i$ replacing d_i. The estimator of the parameter of interest is given by

$$\widehat{\tau}_S \equiv \widehat{\alpha}_p = \widehat{\alpha}_{p+} - \widehat{\alpha}_{p-}$$

3 Asymptotic Distributions

To derive the asymptotic distribution of the estimator for τ, the following assumptions are sufficient.

Assumption 1 $k(\cdot)$ *is a symmetric, bounded, Lipschitz function, zero outside a bounded set;* $\int k(u)du = 1$.

Assumption 1 allows for higher order kernels and a bounded support set for the kernel avoids the use of a trimming function.

Let f_o denote the marginal density of x and $m(x)$ denote the conditional expectation of y given x minus the discontinuity, i.e., $m(x) = E[y \mid x] - \alpha 1[x \geq \overline{x}]$, where $\overline{x}$ is the value of the running variable in which the discontinuity occurs.

---

[5]Note that in the sharp RD design, d_i will be identical to the treatment assignment variable D_i since the probability of being treated is zero below the threshold and one above it.

Assumption 2 *Suppose the data* $(y_i, x_i)_{i=1,2,\ldots,n}$ *is i.i.d. and* α *is defined by*

$$\alpha = \lim_{x \downarrow \overline{x}} E\left[y \mid X = x\right] - \lim_{x \uparrow \overline{x}} E\left[y \mid X = x\right]$$

For some compact interval $\aleph$ *of* x *with* $\overline{x} \in int(\aleph)$, f_o *is* l_f *times continuously differentiable and bounded away from zero;* $m(x)$ *is* l_m *times continuously differentiable for* $x \in \aleph \setminus \{\overline{x}\}$, *and* m *is continuous at* $\overline{x}$ *with finite right and left-hand derivatives to order* l_m.

In the sharp RD design $\tau_S = \alpha$ and the average treatment effect is obtained directly from the discontinuity in the conditional expectation of Y. In the following, I discuss the estimation of α and interpret it as the estimate for the average treatment effect of interest.

Assumption 2 guarantees smoothness of the density of x and the conditional expectation of y on both sides of the discontinuity while allowing for different right and left-side derivatives of m at $\overline{x}$. Also, bounding the density of x on the neighborhood around $\overline{x}$ guarantees there is density ("data") around the discontinuity to estimate the jump size.

Assumption 3 describes the behavior of the moments of the outcome variable around the discontinuity. Define, $\varepsilon = y - E\left[y \mid X = x\right] = y - m(x) - \alpha 1[x \geq \overline{x}]$.

Assumption 3 *(a)* $\sigma^2(x) = E\left[\varepsilon^2 \mid X = x\right]$ *is continuous for* $x \neq \overline{x}$, $x \in \aleph$, *and right and left-hand limits at* $\overline{x}$ *exist.*

(b) For some $\zeta > 0$, $E\left[|\varepsilon|^{2+\zeta} \mid X = x\right]$ *is uniformly bounded on* $\aleph$.

Assumption 3(a) allows the conditional variance of the outcome variable to be a function of the running variable and assures it is well behaved around the cutoff. Part (b) bounds the moments so that a central limit theorem can be applied.

Note that the fixed-h asymptotic distributions described in section 3 do not require additional assumptions over what is used in the standard, small-h literature, *e.g.*, Hahn, Todd and van der Klaauw (2001), Porter (2003) etc.

The asymptotic distribution for the local polynomial estimator of the average treatment effect for a fixed bandwidth, h is given by the following theorem.

Theorem 1 *Suppose assumptions 1 (a) and 3 hold. Suppose assumption 2 (a) holds with* $l_m \geq p+1$ *and* l_f *any nonnegative integer. If* h *is fixed and positive, as* $n \to \infty$, *then*

$$\sqrt{nh}(\widehat{\alpha}_p - \alpha_p^*) \xrightarrow{d} N\left(0, V_{fixed-h}\right) \tag{1}$$

where

$$V_{fixed-h} = e_1' \left[\left(\Gamma_+^*\right)^{-1} \Delta_+^* \left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1} \Delta_-^* \left(\Gamma_-^*\right)^{-1}\right] e_1 \tag{2}$$

$$\alpha_p^* = \alpha + B_{fixed-h}$$

$$B_{fixed-h} = e_1' \left\{ \begin{array}{l} \left(\Gamma_+^*\right)^{-1} \left[\int_0^\infty k(u) Z(\overline{x} + uh) m(\overline{x} + uh) f_o(\overline{x} + uh) du\right] - \\ -\left(\Gamma_-^*\right)^{-1} \left[\int_0^\infty k(u) Z(\overline{x} - uh) m(\overline{x} - uh) f_o(\overline{x} - uh) du\right] \end{array} \right\} \tag{3}$$

7

and

$$\Gamma^*_{+(-)} = \begin{bmatrix} \gamma_0^{+(-)} & \cdots & \gamma_p^{+(-)} \\ \vdots & \ddots & \vdots \\ \gamma_p^{+(-)} & \cdots & \gamma_{2p}^{+(-)} \end{bmatrix}, \ \Delta^*_{+(-)} = \begin{bmatrix} \delta_0^{+(-)} & \cdots & \delta_p^{+(-)} \\ \vdots & \ddots & \vdots \\ \delta_p^{+(-)} & \cdots & \delta_{2p}^{+(-)} \end{bmatrix},$$

$$e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}', \ \gamma_j^+ = \int_0^\infty k(u) u^j f_o(\overline{x} + uh) du,$$

$$\gamma_j^- = (-1)^j \int_0^\infty k(u) u^j f_o(\overline{x} - uh) du,$$

$$\delta_j^+ = \int_0^\infty k^2(u) u^j \sigma^2(\overline{x} + uh) f_o(\overline{x} + uh) du,$$

$$\delta_j^- = (-1)^j \int_0^\infty k^2(u) u^j \sigma^2(\overline{x} - uh) f_o(\overline{x} - uh) du$$

The proof is given in the appendix.

Theorem 1 provides the asymptotic distribution for the local polynomial estimator of the parameter of interest for any bandwidth value.

The fixed-h approach used in theorem 1 explicitly takes into consideration the bandwidth used, without assuming $h \to 0$, and captures the impact of h on the asymptotic variance, $V_{fixed-h}$. The asymptotic variance formulas are functions of known data and can be calculated for given functions $f_o(x)$ and $\sigma^2(x)$ or estimated in a dataset (see section 4).

As is well known, unless the true specification of the population model is known, the local polynomial estimator is biased in finite samples. Even though the treatment effect is nonparametrically identifiable, in practice the local polynomial estimator of the RD design will likely provide a biased estimate of the ATE. The usual asymptotic approximation, by forcing the bandwidth to converge to zero at a "fast enough" rate, can claim that the bias vanishes asymptotically and, hence, can be ignored. However, this approach is known to generate biased asymptotic approximations for the distribution of the parameter of interest that, if left uncorrected, will provide confidence intervals with the wrong coverage, over-rejecting the null hypotheses as pointed out by Calonico et al. (2012). That paper provides bias corrected approximations that incorporate the bias variability to the inference procedure, however it does not explicitly incorporate the potential gains of the improved asymptotic variance approximation in theorem 1.[6]

In most applications the researchers seem to ignore the presence of the bias, by relying on the standard small-h approximations. The fixed-h approach explicitly accounts for it, acknowledging the bias converges to $B_{fixed-h}$ which will depend directly on the bandwidth used, as should be expected.

The bias is the difference of the (scaled) linear projection for $m(x)$ on Z evaluated at $x = \overline{x}$ (i.e., the difference in intercepts) inside the bandwidth above and below the cutoff. Intuitively, the bias in $\widehat{\alpha}$ is a difference between the conditional expectation of the outcome above and below the cutoff that would have arisen in the absence of treatment, i.e., the difference that would have happened nevertheless and are erroneously attributed to the treatment or policy being analyzed. The fixed-h approach tackles the bias

---

[6] It seems that a natural extension of the of the results above would be to incorporate Calonico et al. (2012) bias-correction to the framework presented here.

problem "head on", making explicit the impact of the bandwidth on the bias of the estimate obtained.[7]

From a practical standpoint and to clarify the working assumption practitioners impose when obtaining estimates in RD, it is interesting to draw a parallel of the results in theorem 1 with the issue of model misspecification in parametric models (White 1982, 1996). The problem of estimating the ATE at the cutoff can be seen as one of correctly estimating $E[Y \mid X]$ on both sides of the cutoff. In this sense, the local polynomial estimator is a polynomial approximation to the unknown conditional expectation inside the bandwidth on each side, not different from standard parametric methods. By using a relatively small bandwidth we are fitting the conditional expectation on a restricted support and, hence, expect a polynomial of order p to produce a better fit than if we were trying to fit $E[Y \mid X]$ globally, this is the benefit associated with a nonparametric approach, since it allows the conditional expectation to be unrestricted *outside* the bandwidth. As emphasized by Fan and Gijbels (1996) one of the great advantages to use local polynomial approximations is to be able to rely on least square principles and have access to the statistical knowledge and generalizations connected to least square regression. By fixing the bandwidth when analyzing the asymptotic properties of the treatment effect estimator is akin to treating the local polynomial estimator in RD designs as a parametric estimator within the bandwidth.

Hence, if one assumes that the conditional expectation is parametric and the polynomial of order p above correctly specifies the model in a certain window around the cutoff, the estimator will be asymptotically unbiased. This is the implicit assumption on the functional form being imposed by practitioners when using RD. In applications the asymptotic bias will persist unless the researcher is willing to assume that the model for the conditional expectation is correctly specified in a range around the cutoff.[8]

Even in the absence of bias, the asymptotic variance described in theorem 1 incorporates the choice of bandwidth directly, accounting for potential heteroskedasticity inside the bandwidth which would be ignored by the conventional approximation as it will be shown in corollary 2.

The fixed-h asymptotic approximation in theorem 1 bears a close connection to the small-h approximation in Porter (2003). Porter's relevant approximation for the local polynomial estimator is restated below so that the connection between the approximations can be analyzed[9].

Theorem 2 *(Porter, 2003, theorem 3(a)) Suppose Assumptions 1 (a) and 3 hold. If Assumption 2 (a) holds with $l_m \geq p+1$ and l_f any nonnegative integer, $nh \to \infty$, $h^{p+1}\sqrt{nh} \to C_a$, where $0 \leq C_a < \infty$ then*

$$\sqrt{nh}(\widehat{\alpha}_p - \alpha^*_{small-h}) \xrightarrow{d} N(0, V_{small-h})$$

---

[7] The local polynomial approximation method mitigates the bias problem if it is able to approximate $m(x)$ appropriately, since it partially captures changes in $m(x)$ above and below the cutoff that would exist even in the absence of treatment by using the higher order polynomials.

[8] I am thankful to Matias Cattaneo for discussions on this point.

[9] Note that to recover exactly the same notation as used in Porter (2003), one needs to multiply $B_{small-h}$ by the scaling term $\sqrt{nh}$. Then, the first term in $B_{small-h}$ converges to $\frac{C_a}{(p+1)!}$, as in Porter.

where

$$V_{small-h} = \frac{\sigma^{2+}(\overline{x}) + \sigma^{2-}(\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1 \tag{4}$$

$$\alpha_{small-h}^* = \alpha + B_{small-h}$$

$$B_{small-h} = \frac{h^{p+1}}{(p+1)!} \left[m^{(p+1)+}(\overline{x}) - (-1)^{p+1} m^{(p+1)-}(\overline{x}) \right] e_1' \Gamma^{-1} \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix} \tag{5}$$

and

$$\Gamma = \begin{bmatrix} \gamma_0 & \cdots & \gamma_p \\ \vdots & \ddots & \vdots \\ \gamma_p & \cdots & \gamma_{2p} \end{bmatrix}, \Delta = \begin{bmatrix} \delta_0 & \cdots & \delta_p \\ \vdots & \ddots & \vdots \\ \delta_p & \cdots & \delta_{2p} \end{bmatrix},$$

$e_1 = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}'$, $\gamma_j = \int_0^\infty k(u) u^j \, du$, $\delta_j = \int_0^\infty k^2(u) u^j \, du$ *and* $m^{(l)+(-)}(x)$ *is the lth right (left)-hand derivative of $m(x)$ at point x.*

It is worth noting that both fixed-h and small-h asymptotic approximations are based on the **same estimator** for $\widehat{\alpha}$. For a given bandwidth the bias present in the estimate is given. Small-h asymptotics takes advantage of a first order approximation of the asymptotic bias term to sustain the argument that the bias vanishes as the bandwidth shrinks, however, it is important to have in mind that in practice bias will still be present, even when the bandwidth was chosen to "undersmooth". For a discussion about bias correction and inference in the small-h framework, see Calonico et al. (2012).

There are two noteworthy cases in which the formulas for the fixed-h asymptotic variance and bias simplify to the small-h approximation. First, as expected, when $h \to 0$, the fixed-h formulas for the asymptotic variance and bias of $\widehat{\alpha}$ in theorem 1 approach the asymptotic variance and bias of the small-h approximation in theorem 2.

Corollary 1

$$\lim_{h \to 0} V_{fixed-h} = V_{small-h}$$

$$\lim_{h \to 0} B_{fixed-h} = B_{small-h}$$

Hence, if h is small, fixed-h and small-h provide similar approximations to the asymptotic behavior of $\widehat{\alpha}$.

Secondly, if $f_o(x)$ and $\sigma^2(x)$ are constant around the cutoff and $m(x)$ can be exactly approximated by a polynomial of order $p+1$, the fixed-h asymptotic variance and bias approximations simplify to the small-h asymptotic formulas.

Corollary 2 *If, in the bandwidth around the cutoff, $f_o(x)$ and $\sigma^2(x)$ are constant and $m(x)$ can be exactly approximated by an expansion of order $p+1$ then, the asymptotic variance and bias of $\sqrt{nh}(\widehat{\alpha}_p - \alpha)$ obtained*

by fixed-h (theorem 1) and small-h (Porter 2003) are the same.

$$V_{fixed-h} = V_{small-h}$$

$$B_{fixed-h} = B_{small-h}$$

Refinements obtained by fixed-h are due to incorporating the behavior of $f_o(x)$ and $\sigma^2(x)$ in the ranges around the cutoff, while small-h considers only its values at the cutoff, $f_o(\overline{x})$ and $\sigma^2(\overline{x})$. Hence, heteroskedasticity inside the bandwidth could lead to poor performance by the small-h variance approximation relative to fixed-h.[10]

4   Variance Estimators

To be able to perform inference about α using the information in a given sample, appropriate estimates for the unknown terms in the asymptotic variance formulas from theorem 1 are necessary. Note that the components of the asymptotic variance of $\sqrt{nh}(\widehat{\alpha}_p - \alpha_p^*)$ can be written as,

$$\gamma_j^+ = \int_0^\infty k\left(u\right)u^j f_o(\overline{x}+uh)du = E\left[h^{-1}k\left(\frac{\overline{x}-x}{h}\right)\left(\frac{\overline{x}-x}{h}\right)^j d\right]$$

$$\delta_j^+ = \int_0^\infty k^2\left(u\right)u^j \sigma^2(\overline{x}+uh)f_o(\overline{x}+uh)du = E\left[h^{-1}k\left(\frac{\overline{x}-x}{h}\right)^2\left(\frac{\overline{x}-x}{h}\right)^j d\varepsilon^2\right]$$

and similarly for γ_j^- and δ_j^- . Then, the natural estimator for the asymptotic variance is given by the sample analogue of these quantities,

$$\widehat{\gamma}_j^+ = (nh)^{-1}\sum_{i=1}^n k\left(\frac{\overline{x}-x_i}{h}\right)\left(\frac{\overline{x}-x_i}{h}\right)^j d_i$$

$$\widehat{\delta}_j^+ = (nh)^{-1}\sum_{i=1}^n k\left(\frac{\overline{x}-x_i}{h}\right)^2\left(\frac{\overline{x}-x_i}{h}\right)^j d_i\widehat{\varepsilon}_i^2$$

which are consistent by standard arguments based on the Law of Large Numbers.[11]

Even though these estimators require the calculation of $4(2p+1)$ terms to obtain the plug-in estimator of the fixed-h variance-covariance matrix,

$$\left[\left(\widehat{\Gamma}_+^*\right)^{-1}\widehat{\Delta}_+^*\left(\widehat{\Gamma}_+^*\right)^{-1} + \left(\widehat{\Gamma}_-^*\right)^{-1}\widehat{\Delta}_-^*\left(\widehat{\Gamma}_-^*\right)^{-1}\right] \tag{6}$$

these are simple averages of the data and kernel weights and have the familiar "sandwich form". This estimator is analogue to the Huber-Eicker-White heteroskedasticity robust standard errors in a general

---

[10] If the conditions in corollary 2 hold, the fixed-h and small-h approximations are the same, but they could still differ in practice because they suggest different formulas for standard errors as discussed in section 4.

[11] The residuals used in these estimators will depend on the order of the local polynomial used to estimate the ATE of interest and are given by $\widehat{\varepsilon}_i = y_i - d_i\left(\widehat{\alpha}_{p+} + \widehat{\beta}_{1,p+}(x_i - \overline{x}) + ... + \widehat{\beta}_{p,p+}(x_i - \overline{x})^p\right) - (1 - d_i)\left(\widehat{\alpha}_{p-} + \widehat{\beta}_{1,p-}(x_i - \overline{x}) + ... + \widehat{\beta}_{p,p-}(x_i - \overline{x})^p\right).$

weighted least squares framework, and it comes naturally from the fixed-h approximation. The use of such variance estimators in RD designs have been suggested based in intuitive arguments that did not follow directly from the usual nonparametric asymptotic approximations, see for example Imbens and Lemieux (2008), Lee and Lemieux (2009) and Calonico (2012) among others. The fixed-h results above provide a theoretical framework that justifies the use of such estimators by practitioners. When we treat the estimators as locally parametric our variance estimators become "robust to the choice of bandwidths." This approach directly takes in consideration the impacts of higher order polynomials on the estimator's variance and is flexible regarding the conditional variance and density of X around the cutoff. Note that,

$$\left[\left(\widehat{\Gamma}_+^*\right)^{-1}\widehat{\Delta}_+^*\left(\widehat{\Gamma}_+^*\right)^{-1} + \left(\widehat{\Gamma}_-^*\right)^{-1}\widehat{\Delta}_-^*\left(\widehat{\Gamma}_-^*\right)^{-1}\right] \xrightarrow{p} \left[\left(\Gamma_+^*\right)^{-1}\Delta_+^*\left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1}\Delta_-^*\left(\Gamma_-^*\right)^{-1}\right] = V_{fixed-h}$$

by standard asymptotic arguments, since each term is just the sample analogue of a population expectation and the law of large numbers holds.

In the rectangular kernel case, the variance estimator in equation (6) simplifies to the usual heteroskedastic robust variance estimator using the data just above and below the cutoff. This reinforces the intuition that, for a given bandwidth used, adequate inference can be obtained by dealing with the problem as locally parametric This intuition fits nicely with the findings in Lee and Card (2008) that parametric assumptions are needed when discreteness in the running variable is present.

Case 1 *When the* rectangular kernel *is used, the fixed-h estimator for the asymptotic variance is given by,*

$$
\begin{aligned}
\left(\widehat{\Gamma}_+^*\right)^{-1}\widehat{\Delta}_+^*\left(\widehat{\Gamma}_+^*\right)^{-1} &= \left((2nh)^{-1}\sum_{i=i}^{n} d_i Z_i Z_i'\right)^{-1}\left((4nh)^{-1}\sum_{i=i}^{n} d_i Z_i Z_i'\widehat{\varepsilon}_i^2\right)\left((2nh)^{-1}\sum_{i=i}^{n} d_i Z_i Z_i'\right)^{-1} \\
&= nh\left(\sum_{i=i}^{n} d_i Z_i Z_i'\right)^{-1}\left(\sum_{i=i}^{n} d_i Z_i Z_i'\widehat{\varepsilon}_i^2\right)\left(\sum_{i=i}^{n} d_i Z_i Z_i'\right)^{-1}
\end{aligned}
$$

which is the heteroskedastic robust variance estimator on the OLS *regression of y_i on d_i and the polynomial of $(x_i - \overline{x})$ with order p using the data inside the bandwidth above the cutoff. Similarly for $\left(\widehat{\Gamma}_-^*\right)^{-1}\widehat{\Delta}_-^*\left(\widehat{\Gamma}_-^*\right)^{-1}$ with $(1 - d_i)$ replacing d_i.*

Variance estimators based on small-h asymptotics as proposed by Porter (2003) and Lee and Lemieux (2009), are not fully robust to local heteroskedasticity as suggested by corollary 2. For example, Porter (2003) suggests an estimator for the variance of $\widehat{\alpha}$ using the small-h approximation in corollary 1 which requires only the estimation of the conditional variance of the errors *at* the cutoff approaching both from

right and left and the density of x at the cutoff.[12]

$$\widehat{\sigma}^{2+}(\overline{x}) = \frac{(nh)^{-1} \sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right) d_i \widehat{\varepsilon}_i^2}{\frac{1}{2}\widehat{f}_o(\overline{x})}, \tag{7}$$

$$\widehat{\sigma}^{2-}(\overline{x}) = \frac{(nh)^{-1} \sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right) (1-d_i) \widehat{\varepsilon}_i^2}{\frac{1}{2}\widehat{f}_o(\overline{x})}, \tag{8}$$

$$\widehat{f}_o(\overline{x}) = (nh)^{-1} \sum_{i=1}^{n} k\left(\frac{\overline{x}-x_i}{h}\right), \tag{9}$$

then,

$$\frac{\widehat{\sigma}^{2+}(\overline{x}) + \widehat{\sigma}^{2-}(\overline{x})}{\widehat{f}_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1 \tag{10}$$

is the estimator for the asymptotic variance matrix.

The matrix $\Gamma^{-1}\Delta\Gamma^{-1}$ can be calculated directly because it is a deterministic function of the kernel. An additional drawback of the variance estimator in formula (10) is the need to estimate $f_o(\overline{x})$, which is not necessary if one uses the fixed-h variance estimator in formula (6). To obtain $\widehat{f}_o(\overline{x})$ we need to choose a kernel and a bandwidth for the density estimator, increasing the number of tuning parameters to be chosen. A natural choice would be both the kernel and bandwidth used in the estimation of the parameter of interest. In section 5, I present evidence that using the same bandwidth not only saves one the trouble of choosing another bandwidth, but also provides more reliable inference.

Imbens and Lemieux (2008) propose a plug-in estimator for $\frac{\sigma^{2+}(\overline{x})+\sigma^{2-}(\overline{x})}{f_o(\overline{x})}$ and obtain their estimate for the asymptotic variance of the local linear estimator by scaling it by $e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1$. If higher polynomial orders are used in the estimator, the only change in the formula for their variance estimator is the scaling term. This estimator suffers from the same drawbacks as the one proposed by Porter (2003).[13]

5 Simulations

This section presents simulation evidence displaying the empirical coverage of a standard t-statistic used to perform inference about the treatment effect of interest. All simulations are based on a Sharp RD design. The objective of the simulations is to evaluate the relative performance of tests based on the asymptotic variances obtained by fixed-h and small-h approximations. As shown in the previous results, it is expected that both approaches should yield similar test performance when the bandwidths are small, and differences in empirical coverage should be of greater importance when local heteroskedasticity is present around the cutoff.

---

[12] The estimator presented in formula (10) is not exactly the one presented in Porter (2003). He never suggested a specific estimator $\widehat{f}_o(\overline{x})$, so I chose the standard Rosenblatt-Parzen kernel estimator for $f_o(\overline{x})$ presented in Pagan and Ullah (1999).

[13] In fact, both small-h (Porter, 2003 and Imbens and Lemieux, 2008) variance estimators are based on an estimate for $\frac{\sigma^{2+}(\overline{x})+\sigma^{2-}(\overline{x})}{f_o(\overline{x})}$ and a scaling parameter that depends on the order of the polynomial and kernel used on the estimation of the parameter of interest.

To evaluate the relative performance of tests based on the fixed-h and small-h asymptotic variance approximations and their respective estimators, the focus in this paper is restricted to data generating processes for which the local linear estimator will have no or mild asymptotic bias. Obviously, if the bias in the local linear estimator is important, the inference on both approaches would suffer equally, since they use the same estimator, and would not allow an adequate comparison of the validity of both variance approximations. The fixed-h asymptotic bias approximation, even though more descriptive of the potential bias term is not feasible; for a bias-correction procedure that relies on the small-h first order approximation of the bias term, see Calonico et al. (2012). The adaptation of such bias correction mechanism to the fixed-h approach seems to be a natural way forward that I plan to address in future, ongoing, work.

Evidence from simulations presented below indicates that inference using the fixed-h theoretical approximation has better size behavior than the small-h approach, especially for larger bandwidths. Simulations using feasible estimators for the asymptotic variance indicate that tests based on fixed-h approach can improve over tests based on small-h, especially for larger bandwidths and when local heteroskedasticity are present. As discussed above, the fixed-h asymptotic approximation is akin to treating the estimates as locally parametric and applying the usual heteroskedasticity robust standard errors, hence providing a theoretical justification for what has been observed in applied work.

In the simulations I conducted 2,000 replications with sample size, n, equal to 750 observations, but the actual number of observations included can be significantly smaller, depending on the bandwidth used.[14] For models 1 through 4 the running variable, X, is drawn from a $Normal(50, 100)$, with cutoff set arbitrarily at $\overline{x} = 55$.[15] The error term, u, is also drawn from a normal distribution with mean 0 and standard error equal to 10 at the homoskedastic case but that varies for different scenarios of heteroskedasticity. To exemplify the distortions heteroskedasticity can create and how well the fixed-h asymptotic approximation can capture it, two heteroskedastic cases are considered, with the standard error for u defined as $\sigma(x) = \left(\frac{x}{17.4}\right)^2$ and $\sigma(x) = 10 + 0.25\,(x - \overline{x})^2$, for cases 1 and 2, respectively[16]. Finally, Model 5 is based in a empirical RD problem, and corresponds to the regression function fitted to Lee (2008)'s data above and below the cutoff using a polynomial of order 5. This approach follows Calonico (2012) and can provide some insight about the improvements to test coverage provided by the refined approximations developed in this paper in a more realistic setting. Then, for model 5, $X \sim 2Beta(2, 4) - 1$ and $\sigma(x) = 0.1295$ for the homoskedastic case, and $\sigma(x) = 0.1295 + (5x)^2$ for the heteroskedastic case Since this model introduces some relatively important

---

[14] Even with the relatively large number of observations in the total sample it is still the case that, for some samples, there are no observations inside the bandwidth for all possible choices of bandwidth. In this case, that sample is dropped to guarantee that an estimator can be obtained. this would favor the small-h approximation by making sure that at least some data very close to the cutoff is available.

[15] This choice has the advantages of being in a part of the support for the running variable in which its density is relatively high, increasing the chances that data will be available even for smaller bandwidths, which would favor tests based on the small-h approximation.

[16] These examples aim to highlight the behavior of the fixed-h and small-h approximations in different heteroskedastic contexts. Note that all cases have the same $\sigma(\overline{x})$.

bias for some bandwidths, it will allow us to compare the coverages obtained by both approaches in the presence of substantial bias.

The bandwidths used range from 0.2 to 20, or from $\frac{1}{50}$ to 2 standard deviations of the running variable, which is well within the ranges used in most applications.

The empirical coverages presented are the fraction of rejections in the 2,000 repetitions for a test of size 5% (two-sided). The models that describe how the outcome variable is generated are given by:

- Model 1: $y_i = \mu + \beta_1 x_i + \alpha d_i + u_i$

- Model 2: $y_i = \mu + \beta_1 x_i + \beta_2 x_i^2 + \alpha d_i + u_i$

- Model 3: $y_i = \mu + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \alpha d_i + u_i$

- Model 4: $y_i = \exp\left(\frac{x}{20}\right) + \alpha d_i + u_i$

- Model 5: $y_i = 0.48 + 1.27x_i + 7.18x_i^2 + 20.21x_i^3 + 21.54x_i^4 + 7.33x_i^5 + u_i \qquad$ if $x < 0$

  $\qquad\qquad 0.52 + 0.84x_i - 3.00x_i^2 + 7.99x_i^3 - 9.01x_i^4 + 3.56x_i^5 + u_i \qquad$ if $x \geq 0$

The true value of the parameters in models 1 through 4 are $\mu = 3$, $\alpha = 10$, $\beta_1 = 0.5$, $\beta_2 = -0.005$, $\beta_3 = 0.00002$.

The simulations presented use the local linear estimator $(p = 1)$, since it is the preferred choice in most applied work.[17]

The next subsection compares the test coverages obtained by the theoretical fixed-h and small-h asymptotic distributions derived in theorem 1 and corollary 1. Subsection 5.2 compares the empirical coverages obtained with (feasible) estimated standard errors.

5.1   Simulations for Infeasible Inference

This section presents simulations for infeasible tests that compare test coverages based on the theoretical fixed-h and small-h asymptotic variance formulas in equation 2 derived in theorem 1 and equation 4 derived by Porter (2003). The results obtained are infeasible since they depend on knowledge about $f_o(x)$ and $\sigma^2(x)$ around the cutoff. Nevertheless, they demonstrate the theoretical improvements that fixed-h's approximation provides over small-h asymptotics.

It is worth noting that the objective of these comparisons is to illustrate how the conventional small-h inference, even though valid for small bandwidths around the cutoff, becomes unreliable even in the absence of bias, as we move away from the cutoff. That is natural, since the small-h asymptotic approximation is

---

[17]Even though the results apply to any choice of kernel, the rectangular kernel is the main focus, so that the estimation procedure simplifies to the application of OLS on the data just above and below the cutoff, emphasizing the relationship of fixed-h inference and the standard heteroskedastic-robust standard errors.Similar results were obtained when using $p = 0, 2, 3$; Monte Carlo experiments using sample sizes equal to 500 and 1,000; and the Bartlett and truncated gaussian kernel. They are available from the author upon request.

derived based on the boundary variance and density and, hence, it should not be expected to adequately describe the estimator's behavior away from the threshold. Feasible variance estimators based in such approximation will be able to *partially* correct for these issues as it will be discussed in subsection 5.2.

The empirical coverage for model 1 is presented on figure 1, for which no bias is expected in the estimates, since the local linear estimator correctly specifies the relationship between Y and X inside any of the bandwidths used.

For smaller bandwidths, the use of small-h or fixed-h asymptotic variance generates similar empirical coverages, as expected, but there is a significant decrease in the small-h coverage as the bandwidth increases, with the fixed-h inference increasingly outperforming the standard approximation as the bandwidths increase.

For models 2 through 4 (X has a quadratic, cubic or exponential relationship to Y) both the asymptotic bias and variance approximations can be relevant.[18] The bias, even though mitigated by the use of the local linear function, is an important component in determining the test coverage and affects negatively both approaches somewhat equally. The simulations here focus on cases in which the bias is mild, so that the relative performance of tests based on each variance approximation can be more easily seen. Figure 2 compares the test coverages using fixed-h versus small-h standard error approximations while ignoring the bias, when data is described by model 4 respectively[19]. It is clear that the general pattern observed remains, with fixed-h outperforming small-h, specially for larger bandwidths. As expected, the impact of the potential bias on the empirical coverages becomes more important if the error term has smaller variance. In a set of simulations not reported here for brevity, in which u had a standard error equal to the unit the results were qualitatively similar but with coverages declining faster for both approximations as the bandwidth size increases, in that case the fixed-h unfeasible approximation still significantly outperforms small-h's for bandwidths larger than one standard error of the running variable ($h = 10$).

Finally, model 5, based on Lee (2008)'s data presents a qualitatively similar results as can be seem in figure 3. Note that the coverage varies severely depending on the bandwidth used due to the bias present in the estimation for each choice. Nevertheless, the bias is small enough in this case not to overwhelm the tests completely and it is clear that tests based on fixed-h asymptotic variance produce better coverage than small-h's even in the presence of bias and that the improvements are increasing in bandwidth sizes as predicted by the theory.

As described in section 3, the refinements obtained by fixed-h are due to considering the behavior of $f_o(x)$ and $\sigma^2(x)$ inside the bandwidth. Hence, in the presence of heteroskedasticity, the improvements of the fixed-h approximation are even more important.

For errors that are locally heteroskedastic, the (infeasible) tests based on the fixed-h asymptotic approximation behave very well on both heteroskedasticity cases described in the last section, highlighting its robustness. In the first case, in which the heteroskedasticity is increasing in x around the cutoff, the small-h

---

[18] As discussed in section 3 the local polynomial estimator could be analyzed under a parametric framework as the problem of estimating a (potentially) misspecified model for the relationship between y and x. (White 1982, 1996)

[19] Models 2 and 3 provide qualitatively similar results and are omitted.

asymptotic approximation presents a similar pattern of decreasing coverages as the bandwidths increase as can be seem for models 1 and 4, presented at figures 4 and 5. Nevertheless, the gap between fixed-h and small-h is more pronounced as expected due to the effect of local heteroskedasticity. In contrast, for the local heteroskedasticity described by case 2, the small-h based test has a steep decline[20] in coverage as the bandwidth increases, since it is not able to properly capture the effect of the heteroskedasticity in its asymptotic variance for larger bandwidths as presented in figures 6, 7 and 8 for models 1, 4 and 5, respectively.

The difference of the small-h performance in the two cases can provide useful intuition to when its weaknesses can prove most relevant. The second case was designed to be a "worst case scenario" heteroskedasticity for small-h asymptotics since the conditional variance of the error at the cutoff, $\sigma^2(\overline{x})$ is at the extreme of the range of values assumed by $\sigma^2(x)$ in any given bandwidth. As can be seen from formula (4), the small-h and fixed-h asymptotic variances will be more similar the closer $\sigma^2(\overline{x})$ is from the "weighted average" of $\sigma^2(x)$ inside the bandwidths. In the first case, since $\sigma^2(\overline{x})$ is at the "middle" of the range for the conditional variance, the distortion produced by the heteroskedasticity is less marked than in the second case.

Some points are worth emphasizing. First, the general pattern is that the empirical coverages obtained using the fixed-h results from theorem 1 outperform the small-h approximations, especially for larger bandwidths.

Second, the asymptotic variance refined calculations improve the precision of inference relative to the standard approach. For smaller bandwidths small-h asymptotics provide similar coverages to the fixed-h approach, making clear that the core difference is due to the suitability of the restrictions imposed on $f_o(x)$ and $\sigma^2(x)$ as the bandwidth increases (corollary 2). Naturally, those restrictions tend to be less realistic for larger bandwidths.

Third, in the presence of heteroskedasticity, the small-h asymptotic approximation can have very poor performance, while the fixed-h approach still provides a reliable asymptotic approximation for the estimator's behavior.

5.2   Simulations for Feasible Inference

The simulations in the previous subsection provided evidence that fixed-h asymptotic distribution approximations based on theorem 1 improve over the small-h approximations in the literature. In obtaining those results I used knowledge about the true DGP that is unavailable to the practitioner when implementing such estimators. Also, the coverages obtained by tests based on the unfeasible asymptotic variance formulas for the small-h approximation, can be "unfair" to its performance in practice, since its objective is to approximate the behavior of the treatment effect estimator *at the boundary*, and is not well suited to describe it when the bandwidth becomes larger, as it was patent from the simulations in the last section.

As described in section 4 natural estimators for the asymptotic variance of the parameters of interest are readily available and can be easily calculated for a given sample. This section presents simulations for

---

[20] Note the change in the scale of the y-axis, which now emcompasses the interval from 0 to 1.

the empirical coverage of the tests using different estimated standard errors. The first one is based on the fixed-h asymptotic distribution and is given by formula (6), which is akin to treating the estimates as locally parametric as discussed above. The second is proposed by Porter (2003) and described by formula (10).

For locally homoskedastic errors, the fixed-h standard errors' estimator incorporates the gains of improved inference described in the theory and shown in the infeasible simulations even for large bandwidths, as can be seen in figures 9 through 11 (models 1, 4 and 5, respectively). Tests based on both approximations overreject for very small bandwidths, even in the absence of significant bias, due to the relatively small amount of data available on these cases.

When errors are locally homoskedastic, tests obtained using small-h standard error estimators behave very similarly to fixed-h ones even for larger bandwidths for which the results in section 5.1 would lead one to expect a significantly smaller coverage based on the small-h approach. Essentially, the small-h variance estimators benefit from the fact that, in practice, the estimator for the standard errors partially captures the behavior of $f_o(x)$ and $\sigma^2(x)$ in the range around the cutoff that the small-h asymptotic approximation ignores by forcing $h \to 0$.

To see this point, note that the researcher is not able to exactly estimate $f_o(\overline{x})$ and $\sigma^2(\overline{x})$ from a given dataset as it would have been suggested by the theoretical small-$h's$ asymptotic variance formula. By being "forced" to estimate the variance and the density within the bandwidth, the small-h variance estimator is able to partially capture the local behavior of those terms.

As discussed in section 5.1 the presence of heteroskedasticity can generate substantial problems for the size of tests using small-h approximations. Figures 12 through 16 show simulations for the coverage of feasible tests using the fixed-h and small-h asymptotic variance estimators for the heteroskedastic cases described in section 5.1.

Differently from the homoskedastic case, the fixed-h variance estimator produces tests with better size than those based on the small-h approach. In the first case (figures 12 and 13), the fixed-h variance estimator in formula (6) produces tests with better empirical size even for relatively small bandwidths.[21] Both approaches tend to overreject for smaller bandwidths due to constrained data availability, and a researcher would be ill advised to use too small of a bandwidth.

In figures 14, 15 and 16, where heteroskedasticity is more severe, the fixed-h variance estimator produces tests with coverage very close to the test's nominal size, while for the small-h the coverage rapidly increases towards 1, as the bandwidth increases. Hence, there is evidence that heteroskedasticity can be accurately captured by tests based on fixed-h asymptotic approximations but small-h estimators can produce tests which perform substantially worse. The comparison of heteroskedastic cases 1 and 2 corroborates the theoretical result in theorem 1, that the distortions caused by local heteroskedasticity on small-h's tests are more

---

[21] When using the Bartlett kernel, the coverage for tests based on the small-h variance estimators showed slightly better coverage than those based on fixed-h for small bandwidths only, but in that case both approaches also showed a higher degree of overrejection than in the rectangular kernel case shown here.

important for patterns of heteroskedasticity in which the "weighted average" of the conditional variance on both sides of the cutoff does not approximate $\sigma^2(\overline{x})$, as it is specially evident on case 2 presented here.

Even though the empirical coverages obtained are similar using any of the asymptotic variance estimators when local homoskedasticity holds, it seems the fixed-h standard error estimator is a "safer choice" for practitioners since it is based on a asymptotic approximation that is more "robust to bandwidth choice" and its computation is very easy once a kernel and bandwidth are chosen. Using standard error estimates based on small-h asymptotics can lead to serious size distortions for larger bandwidths, especially in the presence of heteroskedasticity, even in the absence of bias.

Furthermore, the fixed-h variance estimator has the advantage of not requiring the estimation of $f_o(\overline{x})$. This entails the choice of (potentially different) kernel and bandwidth for $\widehat{f}_o(\overline{x})$. The additional choice of these two tuning parameters might significantly alter the empirical size of the tests performed about $\widehat{\tau}$ and depends on the discretion of the researcher.

To exemplify this issue, figure 17 shows the simulated empirical coverages obtained by using the small-h variance estimator for DGP 1[22] for five different scenarios, with homoskedastic errors. Each scenario differs by the choice of the bandwidth, h_f, used in formula (9) to obtain $\widehat{f}_o(\overline{x})$. The first reproduces the small-h result described above by choosing the same bandwidth used to estimate $\widehat{\tau}$, i.e., $h_f = h$, the other lines are the empirical coverages obtained by using bandwidth of 1, 5, 10 and 20[23] for $\widehat{f}_o(\overline{x})$ independent of the bandwidth used for $\widehat{\tau}$.

The choice of bandwidth on the estimation of $\widehat{f}_o(\overline{x})$ can have a relevant impact on the test coverages. Choosing the same bandwidth as used in estimating the parameter of interest provides more stable empirical coverages for a wide range of h relative to the cases in which the bandwidths are different. The cautious practitioner using the small-h variance estimator would be well advised to choose the same bandwidth for both estimators.

To the empirical researcher, a useful conclusion that can be drawn from the simulations presented is that by performing inference using the fixed-h approach, which is akin to treating the estimates as locally parametric and simplifies to the standard heteroskedastic robust standard errors in the rectangular kernel case, one can feel relatively confident about validity of the tests and its sizes for almost any bandwidth used.

The key issue facing the researcher in the choice of bandwidth is how to deal with the bias in practice. The bias is a main contributor for the divergence between the empirical and nominal sizes of tests. The results presented suggest that one should realize that for a given dataset, taking advantage of RD to estimate a treatment effect of interest is the exercise of estimating the conditional expectation of the outcome variable above and below the cutoff *inside* the bandwidth, and that can be seen under the parametric framework. Naturally, for larger bandwidths one would expect that the potential to misspecification increases, requiring higher order polynomials of X or the inclusion of covariates to guarantee the validity of estimates, even if

---

[22] The graphs for GDPs 2, 3 and 4 are available from the author upon request.

[23] $\frac{1}{20}, \frac{1}{4}, \frac{1}{2}$ and 1 standard deviations of the running variable, respectively.

the nonparametric thought experiment would suggest that those are not needed. The widespread practice in applied research to check the behavior of the estimates for different bandwidths seem like a sound practice and the standard errors should incorporate adequately the changes in bandwidth as suggested by the theoretical results and simulations above.[24]

6 Empirical Example

This section uses data from Lee's (2008) study of the electoral advantage of incumbency in the United States to exemplify potential differences on the small-h and fixed-h approximations discussed above.

As pointed out in Lee (2008) the U.S. Congressional electoral has a "built-in" RD since being the incumbent party in congressional district is a deterministic function of the candidate-party's vote share in the district during the last electoral cycle. This feature can be described in the following model

$$
\begin{aligned}
v_{i2} &= \alpha w_{i1} + \beta v_{i1} + \gamma d_{i2} + e_{i2} \\
d_{i2} &= 1\left[v_{i1} \geq \frac{1}{2}\right]
\end{aligned}
$$

where v_{it} is the democratic candidate's vote share in Congressional district i in election year t; w_{it} is a vector of characteristics or agents' choices (potentially unobserved) as of election day on period t, and d_{it} is the indicator function that signals if the Democratic party is the incumbent in district i at election period t. Also, we assume that $f_{i1}(v \mid w)$, the density of v_{i1} conditional on w_{i1}, is continuous in v. The main issue in the analysis, as discussed in detail by Lee (2008), is that w_{it} is potentially unobserved and would likely be correlated with being incumbent in a certain district. For example, w_{i1} would include party resources, demographic characteristics and political leaning of districts that would affect both the vote share in period 1 and 2, thus biasing the estimates for the causal effect of incumbency.

The thought experiment we would like to perform to get to the causal effect of incumbency would be to randomly allocate incumbency in district to Democrats and Republicans while keeping all other characteristics constant. This clearly cannot be done, but by looking at closely contested elections and given the inherent uncertainty regarding its outcome due to the presence of a random chance element we can consider that whether Democrats or Republicans are incumbents in those districts is decided randomly.[25] This can

---

[24]Adequate estimation of the bias based on the results on theorem 1 would require observing or estimating the counterfactual conditional expectation of y around the cutoff in the absence of treatment, which is not available for most cases where the RD design is relevant. The small-h asymptotic bias approximation (Porter 2003) lends itself for estimation, since estimates for $m^{(p+1)}$ above and below the cutoff can be obtained. However, the results in section 5.1 indicate that this would be a relatively poor approximation. Furthermore, to a large degree, bias reduction could be obtained by increasing the order of the local polynomial fitted above and below the cutoff, reducing the "misspecification" in the model (see section 3). For a detailed discussion on bias correction in RDD and the inference adjustments needed when performing such correction, see Calonico et al. (2012).

[25]For a discussion on the assumptions necessary for the validity of RDD in this example, see Lee (2008).

provide good estimates of the causal effect of incumbency in closely contested elections.

The data used is the same as in Lee (2008) and includes U.S. Congressional election returns from 1946 to 1998, excluding the years ending in '0' and '2' due to the decennial redistricting which characterizes the U.S. congressional electoral system. Also, even though most elections have Democrats and Republicans as the two strongest parties, third parties do obtain some share of the votes. In that case the threshold for victory in the election is not 50% of the votes. To solve this issue the vote share variable is defined as the difference in vote share between the Democrat candidate and the strongest opponent. Hence, the Democrat wins the election when this variable crosses the 0 threshold.[26]

Table 1 shows the local polynomial estimates for advantage of incumbency and the estimated fixed-h and small-h standard errors given by formulas 6 and 10, respectively.

Table 1: Incumbency Effects and Estimated Standard Errors - Lee (2008)

Dependent Variable: Democrat Vote Share - Election $t + 1$

Panel A: Nadaraya-Watson Estimator ($p = 0$)			
	All	$|Margin| \leq 0.5$	$|Margin| \leq 0.05$
Estimated Effect	0.351	0.257	0.096
(Fixed-h Standard Errors)	(0.0041)	(0.0038)	(0.0090)
[Small-h Standard Errors]	[0.0041]	[0.0038]	[0.0090]
Difference (%)	1.6%	0.5%	0.1%
Panel B: Local Linear Estimator ($p = 1$)			
Estimated Effect	0.118	0.090	0.048
(Fixed-h Standard Errors)	(0.0056)	(0.0062)	(0.0159)
[Small-h Standard Errors]	[0.0068]	[0.0071]	[0.0180]
Difference (%)	21.4%	14.5%	13.2%
Panel C: Local Polynomial Estimator ($p = 4$)			
Estimated Effect	0.077	0.066	0.105
(Fixed-h Standard Errors)	(0.0114)	(0.0144)	(0.0312)
[Small-h Standard Errors]	[0.0167]	[0.0179]	[0.0447]
Difference (%)	46.5%	24.3%	42.4%
Observations	6558	4900	610

Panel A presents the estimates using the Nadaraya-Watson estimator ($p = 0$) for different bandwidths. Column 1 uses all the data available, column 2 looks only at elections for which the margin of victory in period $t - 1$ was within 50% of the total votes; and column 3 uses only elections with margins lower than 5%.

---

[26] The data is available at http://economics.mit.edu/faculty/angrist/data1/mhe, for further discussion on the data and its issues see Lee (2008).

Panel B presents similar estimates and standard errors obtained by a local linear estimator ($p = 1$) which is the preferred specification in several RD applications in the literature and is expected to significantly reduce bias in the estimates of the incumbency advantage effect.

Panel C, presents the results that Lee (2008) called the "parametric fit" in the first column, which uses a polynomial of order 4 to fit the whole data, the other two columns use smaller bandwidths to emphasize how the order of the polynomial chosen to fit the data can impact significantly estimates and standard errors especially in small samples, overfitting the data.

The point estimates are exactly the same estimates presented by Lee (2008) for Panel A and the first column in Panel C. The remainder estimates are new and reinforce the intuition presented by Lee and the validity of the RD in this problem..

The results indicate that there is a significant incumbent advantage in U.S. Congressional races, even when we compare districts that had close elections during the previous electoral cycle, for which the determination of incumbent status can be considered "as good as randomized." See Table 2 at the appendix for similar results that include the pre-determined variables Lee (2008) uses and the comparisons for fixed-h and small-h standard errors. The observed differences in pre-determined variables between incumbents and challengers vanish as we compare districts that had competitive races previously, lending credibility to the RD as identification strategy for the incumbency effect.[27] More relevant to this paper are the differences between the competing standard error estimates.

The estimated standard errors, shown in Table 1 and Table 2 in the appendix, differ significantly, with the fixed bandwidth based standard errors being smaller in most of the cases as one would expect given the simulations on section 5. Also as expected, using smaller bandwidths comes at a large cost in terms of precision of the estimates, due to the smaller amount of data available which affect negatively both standard error estimators. The two last columns in all three panels show the increase of the estimated standard errors when the data is restricted to districts that had victory margins smaller than 5% in the previous election.

More interestingly, it is usual that the gap between the two standard errors estimates, as measured by their relative gap, declines as the bandwidth shrinks as predicted in section 3. For panel A, B and the first two columns of panel C, this pattern is confirmed as we compare the percent difference between the standard errors within panels.[28] For the last entry on panel C, note that both standard errors become larger compared with the larger bandwidth in column 2, but the relative gap increases. That can be due to the fact that the fixed-h standard error estimate requires the calculation of $4(2p + 1)$ terms (see equation 6) and, hence, be more susceptible to the combination of large polynomial order and small sample sizes induced by the smaller bandwidth. Nevertheless, it still provide tighter confidence intervals than the small-h estimated standard error.

Finally, by comparing panels along a column, note that the differences between fixed-h and small-h

---

[27] Note that fitting a polynomial of higher order, especially for small bandwidths, can reduce precision on the estimates and overfit the data.

[28] A similar pattern emerges from table 2 in the appendix.

estimated standard errors increase as the order of the polynomial used to fit the data increases for a same bandwidth. This is due to the fact that the small-h standard error estimator is based on a scaling term that is fixed for a given kernel and polynomial order choice, $e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1$ as discussed on page 13. Intuitively, as the polynomial order increases, the greater the distortion between using this approximation and the more refined approximation implied by the fixed-h approach, $\left(\widehat{\Gamma}_+^*\right)^{-1} \widehat{\Delta}_+^* \left(\widehat{\Gamma}_+^*\right)^{-1}$ is likely to be.

7   Conclusion

The use of regression discontinuity designs to obtain estimates of treatment effect, τ, has been widely used in recent years by researchers in economics. Special attention has been given to the use of local polynomial estimators to obtain the ATE of interest.

The standard literature on RD designs (Hahn, Todd and van der Klaauw 2001; Porter 2003; Imbens and Lemieux 2008) analyses the asymptotic behavior of the ATE's estimator by assuming that the bandwidth around the discontinuity, h, shrinks fast enough towards zero, $h \rightarrow 0$, to guarantee identification of the parameter of interest (small-h asymptotics). However, in practice, to obtain an estimate of the treatment effect and perform inference the empiricist is required to use a particular value of the bandwidth which is necessarily greater than zero.

Hence, a disconnect arises between theory and practice for inference in RD since most practitioners use the usual parametric inference methods, even though the asymptotic theory for RD is based on nonparametric, small-h asymptotic approximations.

This paper bridges the gap between theory and practice by providing a set of conditions under which the use of the usual parametric tools of inference would be locally valid and develops an alternative asymptotic theory for the RD treatment effects estimator by treating h as fixed. This fixed-h asymptotics explicitly acknowledges the fact that a researcher has to choose a bandwidth to implement the estimator and is usually bounded in their ability to reduce the bandwidth size by data availability constraints.

Additionally, the fixed-h asymptotic approximation for the behavior of the estimators of interest provides refined inference relative to asymptotic distributions based on the conventional approach. As it is shown in section 3, this approach is akin to treat the estimator as parametric in the neighborhood of the cutoff. This intuition fits nicely with current practice in applied work, which has focused on the usual Huber-Eicker-White heteroskedasticity robust standard errors, essentially treating the estimates as locally parametric. Hence, this paper provides a theoretical framework that justifies such choice, and provides evidence that treating these nonparametric estimates as locally parametric can improve inference and successfully adjust standard errors to reflect the choice of bandwidth by the researcher.

Also, conditions under which both fixed-h and small-h would provide the same approximations are provided. Fixed-h asymptotic bias and variance converge to small-h's as $h \rightarrow 0$ (corollary 1), indicating the approximations should be very similar for small bandwidths. Additionally, fixed-h and small-h variance

approximations are the same if one assumes that the errors are locally homoskedastic and the density of the running variable is locally constant in the bandwidth around the cutoff (corollary 2). Hence, the improvements obtained by the fixed-h approximations (partially) derive from taking in account the local heteroskedasticity which is implicitly ruled out by the small-h approximation.

Simulations provide evidence that fixed-h asymptotic distributions produce tests with better empirical coverage than the usual small-h approximations used in the literature, specially when larger bandwidths are used.

Simple feasible estimators for the refined, fixed-h, standard errors are provided and shown to incorporate the theoretical gains of the improved approximations in simulations. These estimators are simple to implement, and are similar to the Huber-Eicker-White heteroskedasticity robust standard errors in a weighted least squares framework commonly used in applications. Also, they have the advantage of not requiring the estimation of the density of the running variable at the discontinuity. In line with the theoretical findings, the fixed-h variance estimators can improve markedly over small-h estimators in the presence of heteroskedasticity and should be generally preferred. Simulations using heteroskedastic errors have provided evidence that feasible tests based on the fixed-h approach obtain better coverage, outperforming small-h even at relatively small bandwidths. As expected from the theory, tests based on the small-h asymptotic approximations have similar size performance to tests based on fixed-h when the errors are locally homoskedastic.

The results indicate that the fixed-h standard error estimator is a "safer choice" for practitioners since it is based on a more robust asymptotic approximation, specially if heteroskedasticity is suspected, and its computation is very easy once a kernel and bandwidth are chosen. If the popular rectangular kernel is used, the fixed-h standard errors simplify to the usual heteroskedasticity robust standard errors available in most statistical packages, which are valid and outperform tests that use small-h approximations. A simple application to Lee (2008) analysis of the electoral advantages of incumbency was presented and confirms that the improvements in terms of precision of the estimates are relevant in practice.

This provides tests with adequate size for any choice of bandwidth in the absence of bias. Regarding the presence of bias, the analysis above sheds light on the fact that, in practice, estimates in the RD context will be asymptotically biased unless one is willing to assume that the conditional expectation is parametric and the polynomial used correctly specifies the model in a certain window around the cutoff. This is the implicit assumption on the functional form being imposed by practitioners when using RD.[29]

Taking advantage of RD to estimate the treatment effect of interest is the exercise of estimating the conditional expectation of the outcome variable above and below the cutoff *inside* the bandwidth, and that can be seen under the "parametric framework". Naturally, for larger bandwidths one would expect that the potential to misspecification increases, requiring higher order polynomials of X or the inclusion of covariates

---

[29] For an interesting approach in bias correction and inference within the small-h framework, see Calonico et al. (2012). The adaptation of such bias correction mechanism to the fixed-h approach seems to be a natural way forward that I plan to address in future, ongoing, work.

to guarantee the validity of estimates. Using fixed-h standard errors (or the usual heteroskedastic robust for rectangular kernels) produces tests that incorporate adequately the changes in bandwidth and checking the behavior of estimates for different bandwidths seem like a sound practice for applied researchers exploiting RD.

References

[1] Angrist, J. and J. Pischke, 2009, Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

[2] Calonico, S.; M. Cattaneo and R. Titiunik, 2012, Robust Nonparametric Bias-Corrected Inference in Regression Discontinuity Design. Working Paper, Department of Economics, University of Michigan.

[3] Cattaneo, M.; B. Frandsen and R. Titiunik, 2012, Randomization Inference in the Regression Discontinuity Design. Working Paper, Department of Economics, University of Michigan.

[4] Fan, J., and I. Gijbels, 1996, Local Polynomial Modelling and its Applications. Chapman & Hall, New York.

[5] Fan, Y., 1998, Goodness-of-Fit Tests Based on Kernel Density Estimators with Fixed Smoothing Parameters. Econometric Theory 14, 604-621.

[6] Hahn, J.; P. Todd and W. van der Klaauw, 1999, Evaluating the Effect of an Antidiscrimination Law Using a Regression-Discontinuity Design. NBER Working Paper Series, n. 7131.

[7] Hahn, J.; P. Todd and W. van der Klaauw, 2001, Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. Econometrica 69, 201–209.

[8] Hashimzade, N. and T. Vogelsang, 2008, Fixed-b asymptotic approximation of the sampling behaviour of nonparametric spectral density estimators. Journal of Time Series Analysis 29 (1), 142–162.

[9] Imbens, G. and J. Angrist, 1994, Identification and Estimation of Local Average Treatment Effects. Econometrica 61, 467–476.

[10] Imbens, G. and T. Lemieux, 2008, Regression Discontinuity Designs: A Guide to Practice. Journal of Econometrics 142 (2), 615–635.

[11] Imbens, G. and K. Kalyanaraman, 2012, Optimal Bandwidth Choice for the Regression Discontinuity Estimator. Review of Economic Studies 79(3), 933-959.

[12] James, B., 2004, Probabilidade: um curso em nível intermediário. 3 ed., Rio de Janeiro, IMPA.

[13] Lee, D., 2008, Randomized Experiments from Non-random Selection in U.S. House Elections. Journal of Econometrics 142 (2), 675–697.

[14] Lee, D. and D. Card, 2008, Regression Discontinuity Inference with Specification Error. Journal of Econometrics 142 (2), 655–674.

[15] Lee, D.; T. Lemieux, 2009, Regression Discontinuity Designs in Economics. NBER Working Paper Series, n. 14723.

[16] McCrary, J., 2008, Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. Journal of Econometrics 142 (2), 698–714.

[17] Neave, H., 1970, An Improved Formula for the Asymptotic Variance of Spectrum Estimates. The Annals of Mathematical Statistics 41, 70-77.

[18] Pagan, A and A. Ullah, 1999, Nonparametric Econometrics. Cambridge University Press.

[19] Porter, J., 2003, Estimation in the Regression Discontinuity Model. Unpublished Working Paper, Department of Economics, University of Wisconsin. *Manuscript date*: May, 7, 2003.

[20] White, H., 1982, Maximum Likelihood Estimation of Misspecified Models. Econometrica 50, 1–24.

[21] White, H., 1996, Estimation Inference and Specification Analysis. Cambridge University Press.

Figure 1: Simulation for Infeasible Inference - Linear GDP - Homoskedastic Errors

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Linear Model – Homoskedastic Errors – Kernel: Rectangular
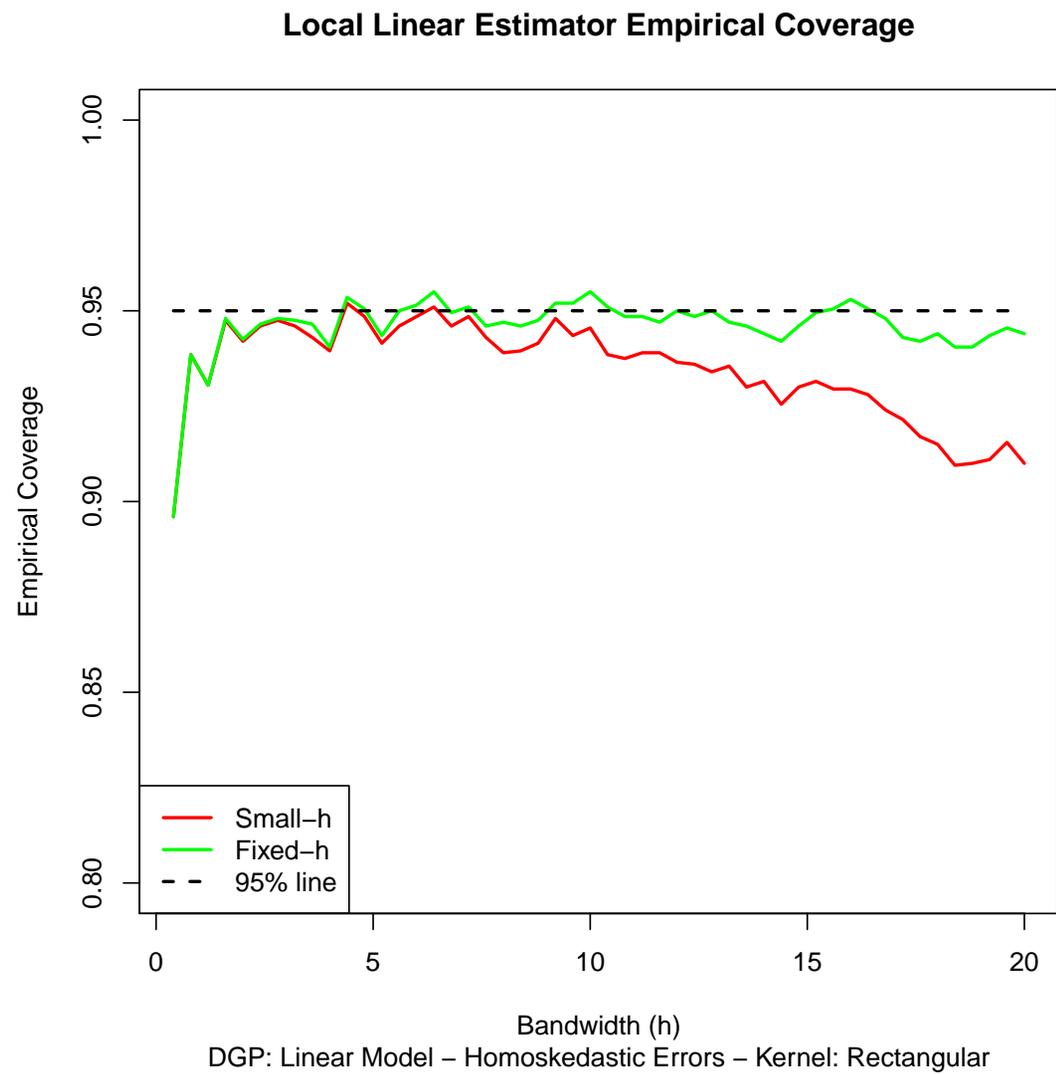
Figure 2: Simulation for Infeasible Inference - Exponential GDP - Homoskedastic Errors

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Exponential Model – Homoskedastic Errors – Kernel: Rectangular
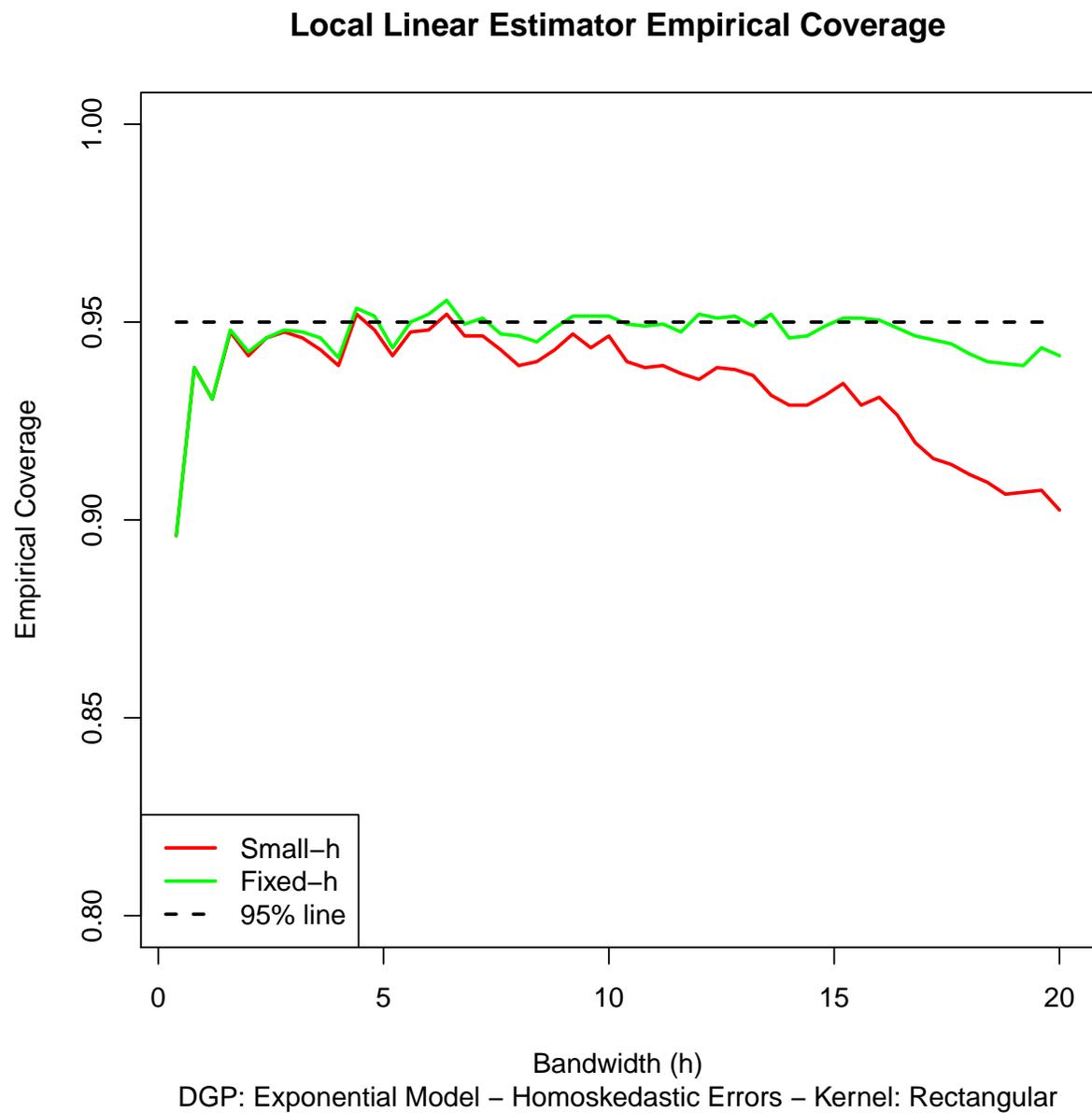
Figure 3: Simulation for Infeasible Inference - Lee (2008) GDP - Homoskedastic Errors

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Lee (2008) – Homoskedastic Errors – Kernel: Rectangular
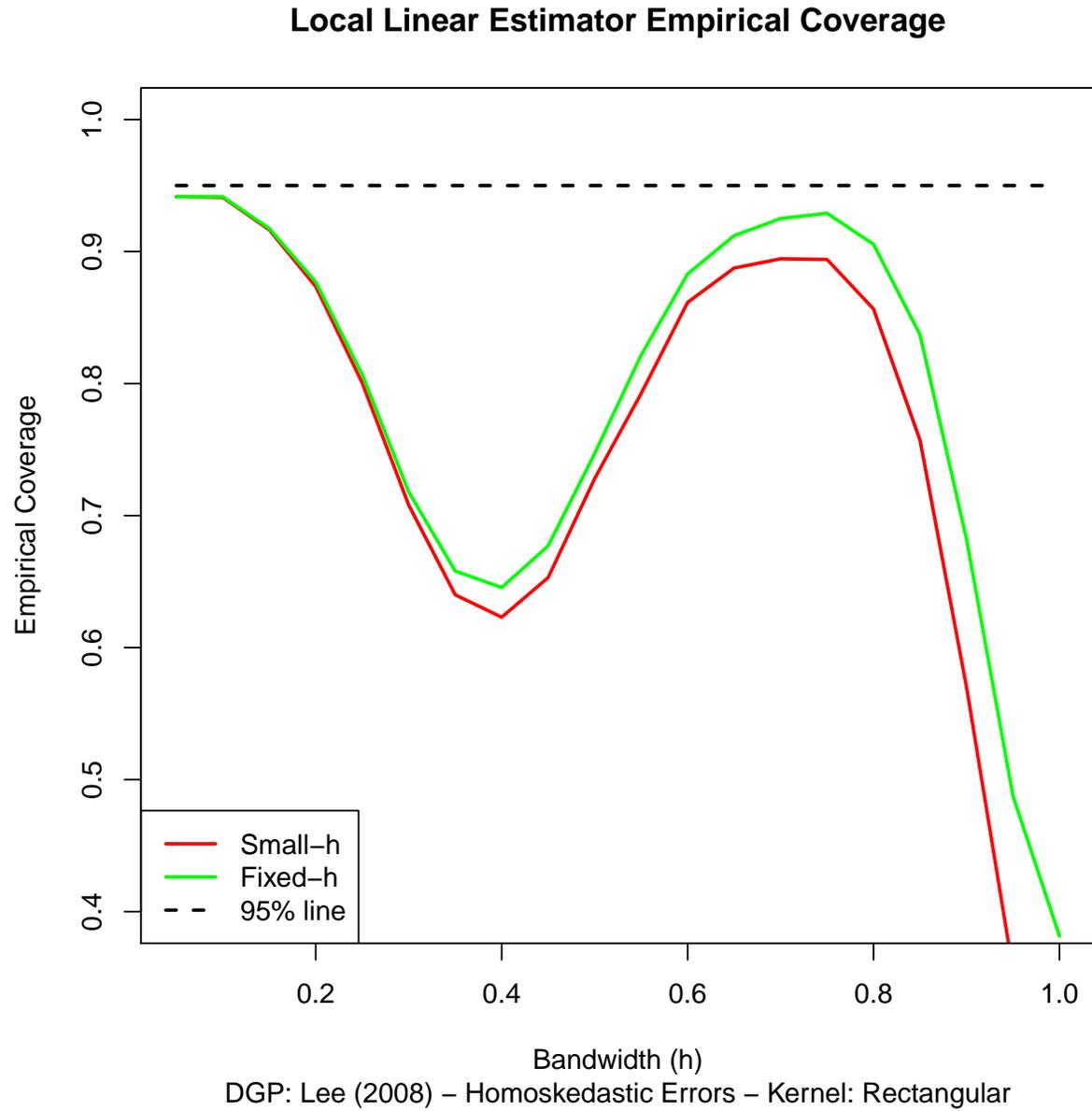
Figure 4: Simulation for Infeasible Inference - Linear GDP - Heteroskedastic Errors (Case 1)

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Linear Model – Heteroskedastic Errors (Case 1) – Kernel: Rectangular
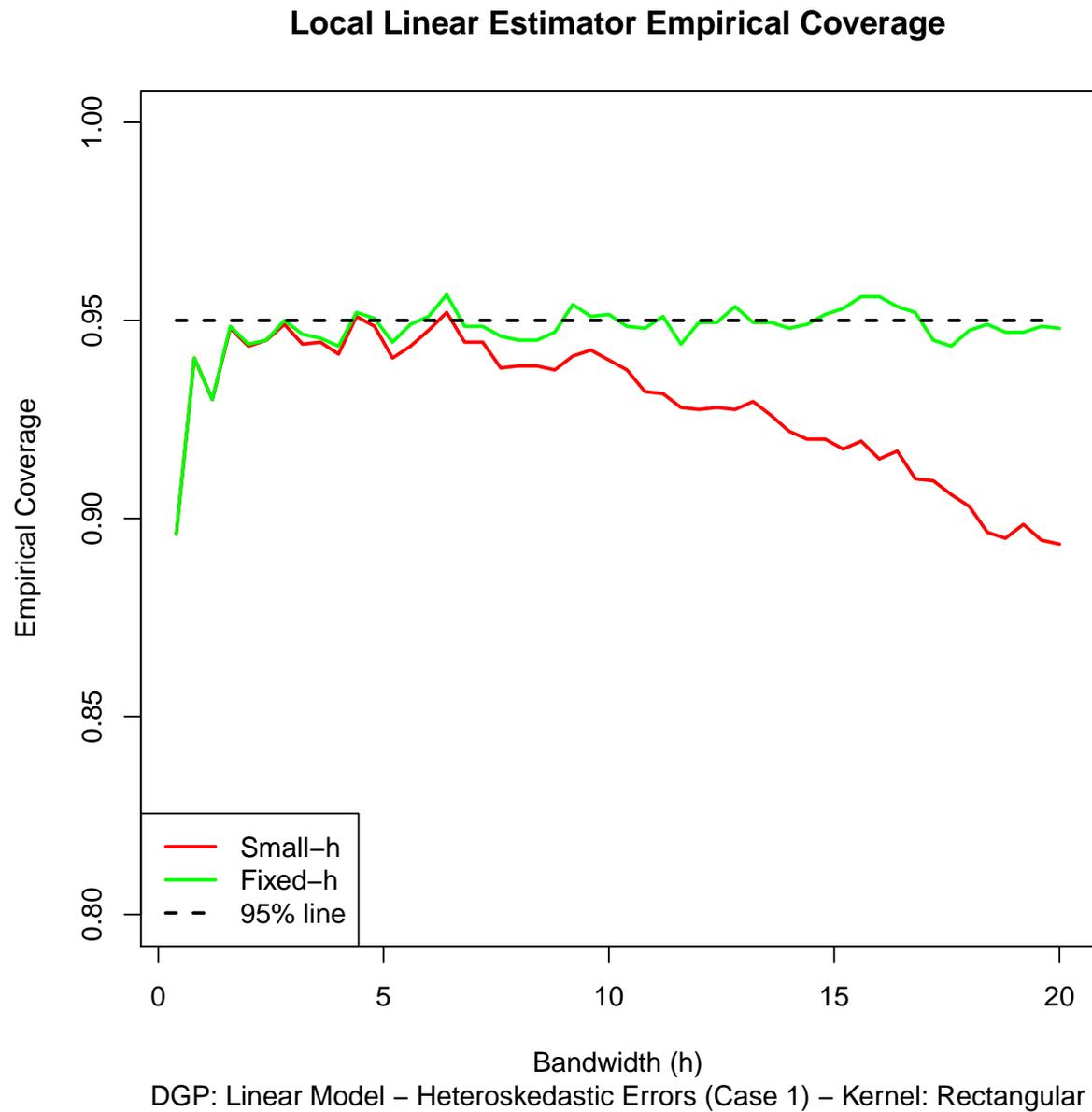
Figure 5: Simulation for Infeasible Inference - Exponential GDP - Heteroskedastic Errors (Case 1)

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Exponential Model – Heteroskedastic Errors (Case 1) – Kernel: Rectangular
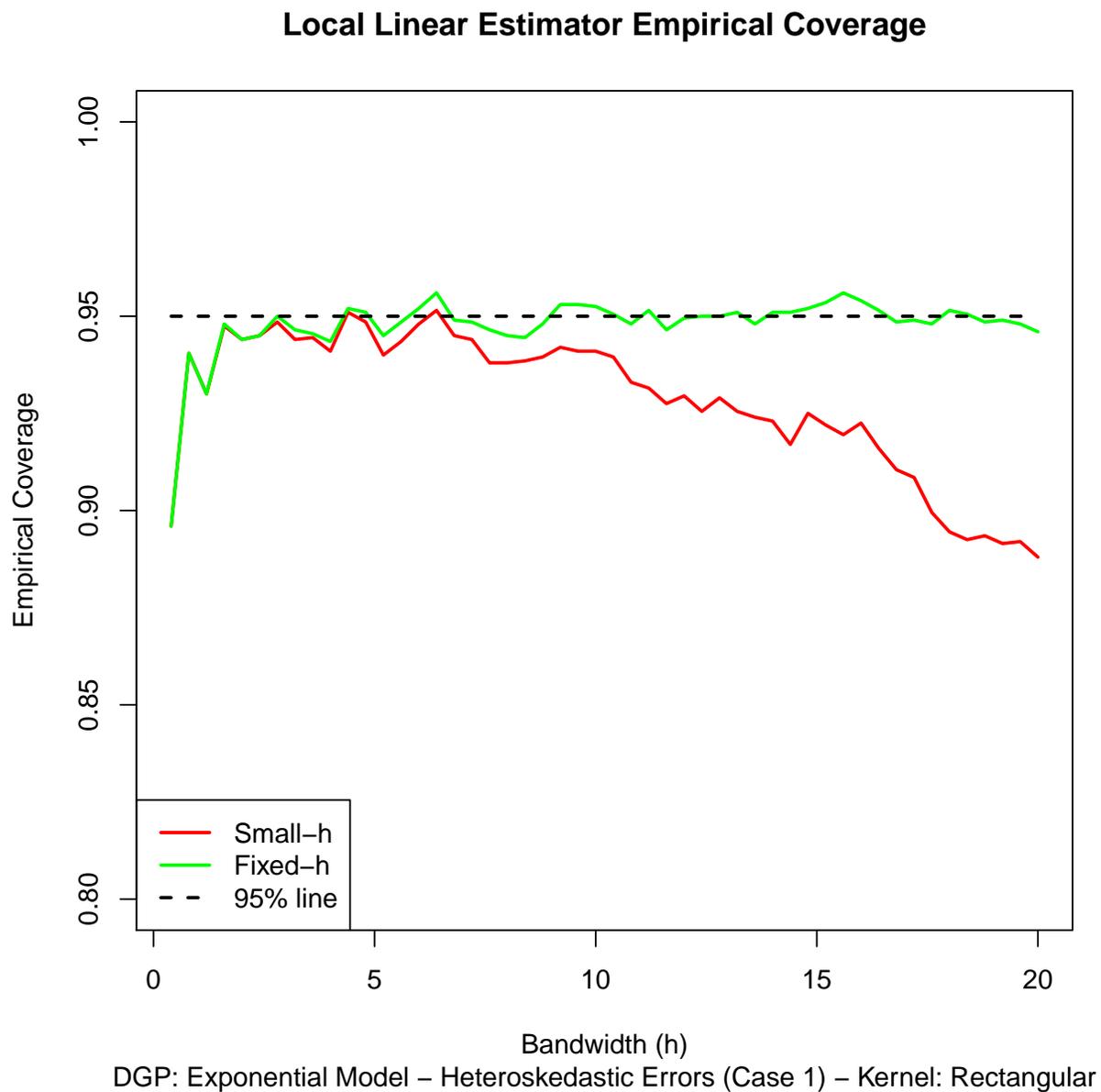
Figure 6: Simulation for Infeasible Inference - Linear GDP - Heteroskedastic Errors (Case 2)

Local Linear Estimator Empirical Coverage

DGP: Linear Model – Heteroskedastic Errors (Case 2) – Kernel: Rectangular
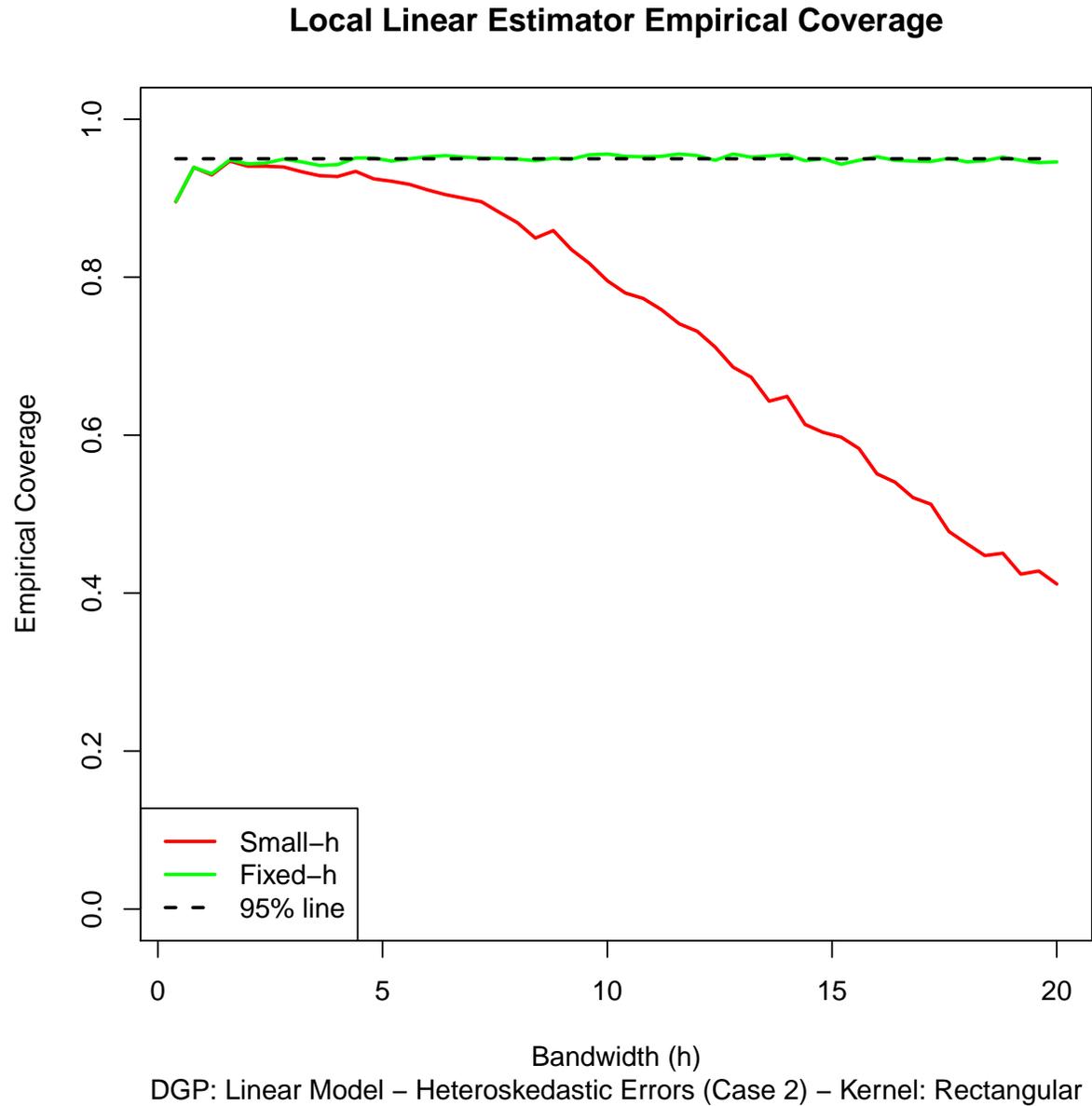
Figure 7: Simulation for Infeasible Inference - Exponential GDP - Heteroskedastic Errors (Case 2)

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Exponential Model – Heteroskedastic Errors (Case 2) – Kernel: Rectangular
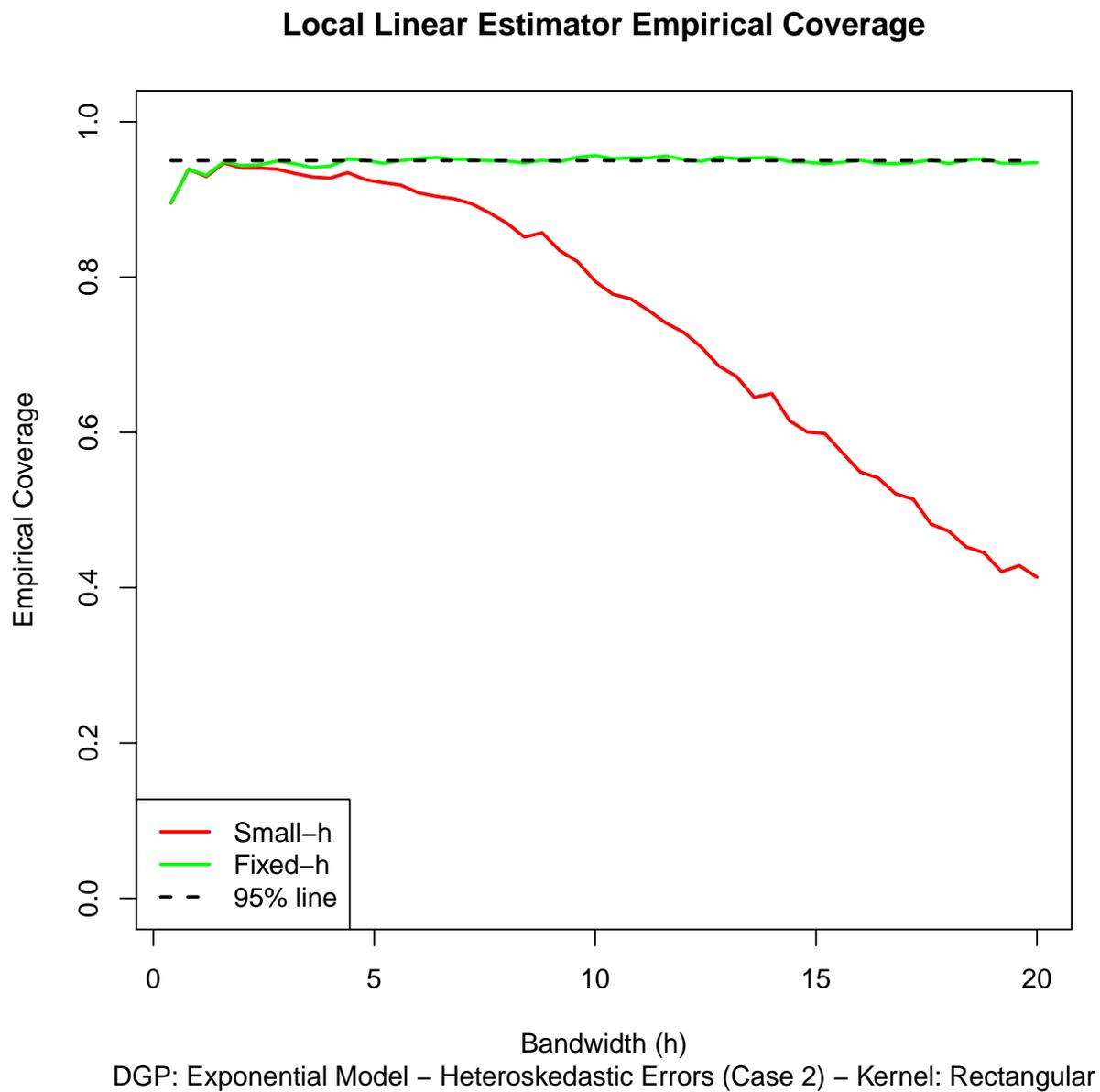
Figure 8: Simulation for Infeasible Inference - Lee (2008) GDP - Heteroskedastic Errors

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Lee (2008) – Heteroskedastic Errors – Kernel: Rectangular
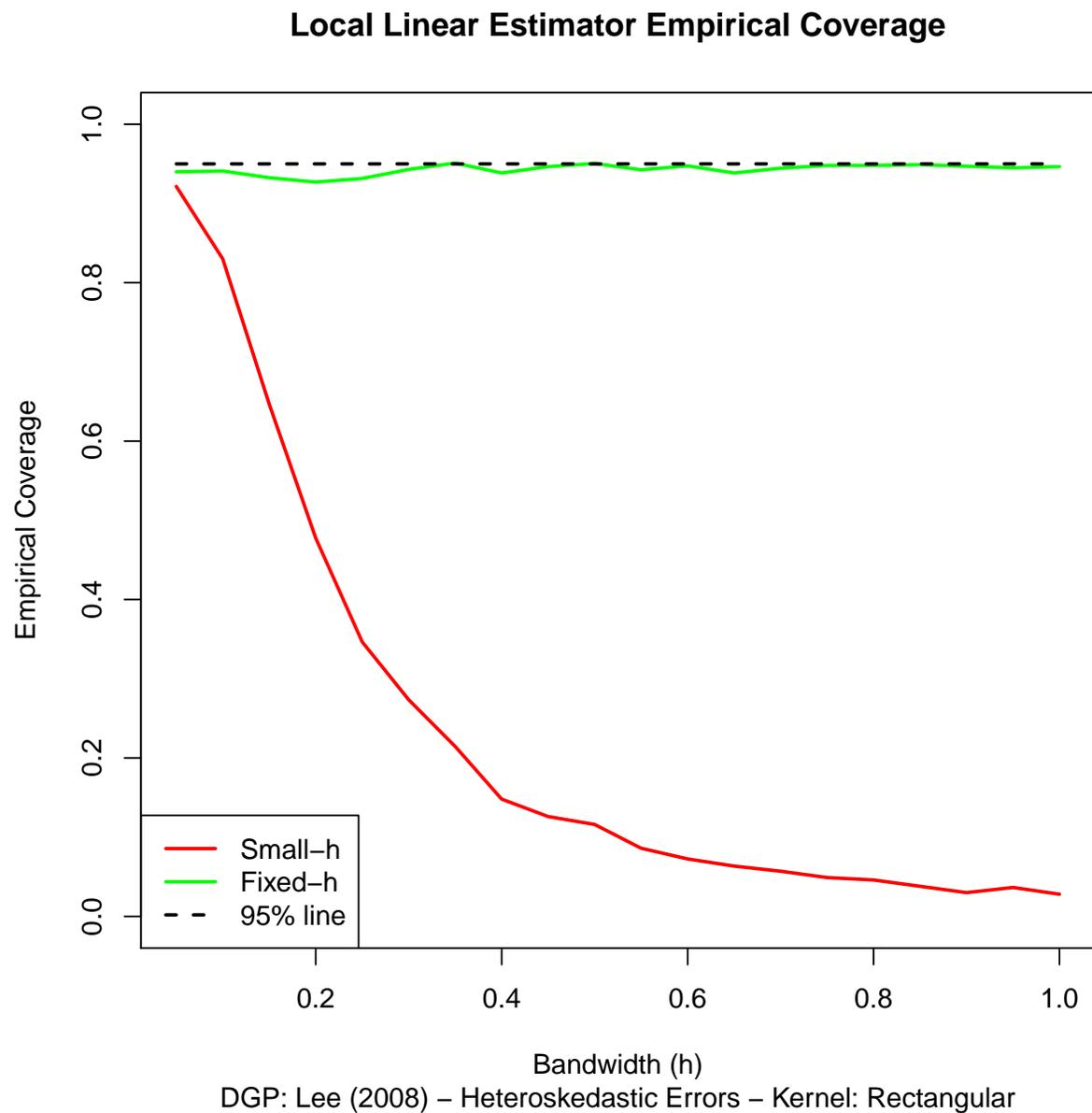
Figure 9: Simulation for Feasible Inference - Linear GDP - Homoskedastic Errors

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Linear Model – Homoskedastic Errors – Kernel: Rectangular

Local Linear Estimator Empirical Coverage

Local Linear Estimator Empirical Coverage

DGP: Lee (2008) – Homoskedastic Errors – Kernel: Rectangular

Figure 12: Simulation for Feasible Inference - Linear GDP - Heteroskedastic Errors (Case 1)

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Linear Model – Heteroskedastic Errors (Case 1) – Kernel: Rectangular
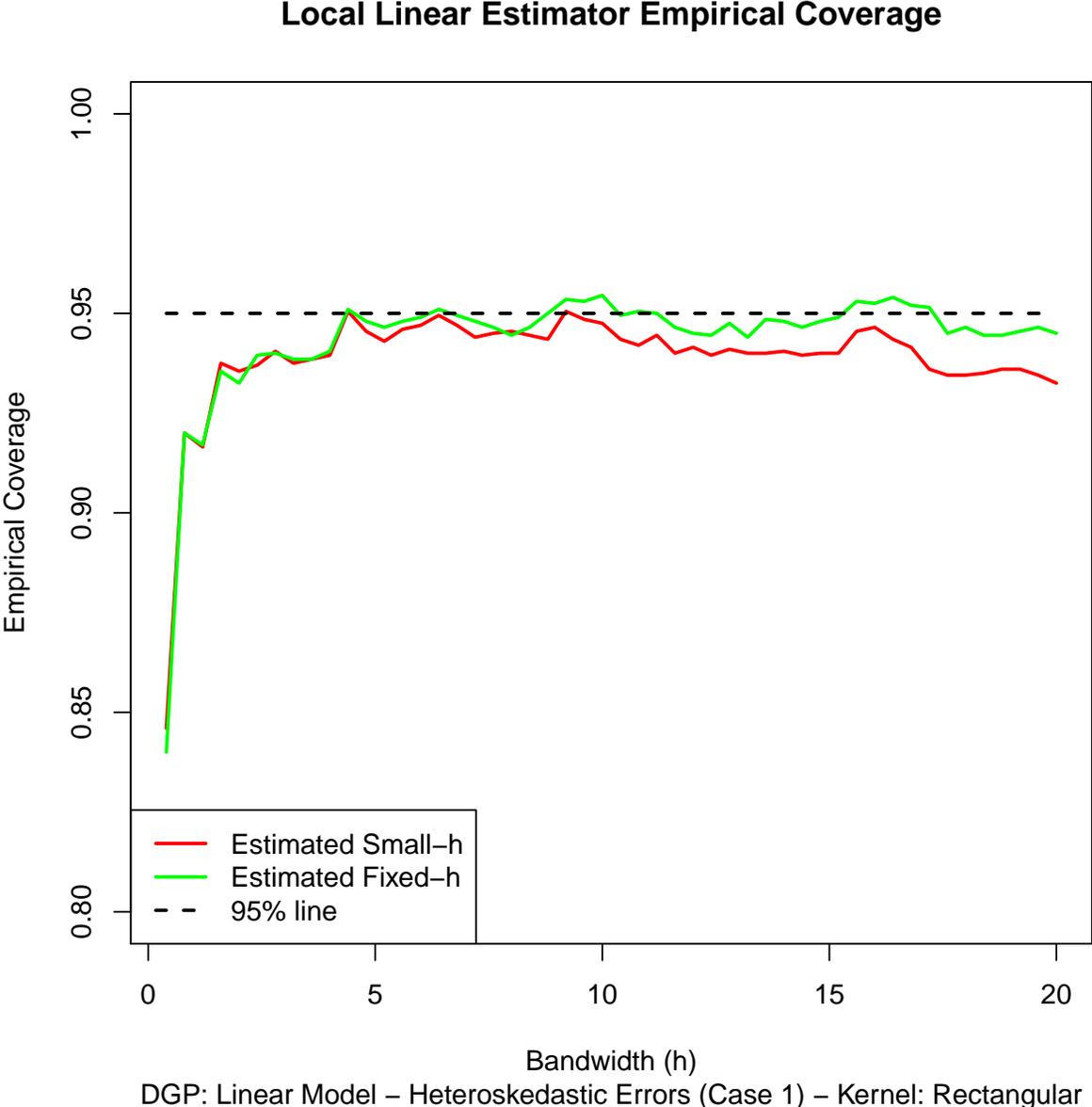
Figure 13: Simulation for Feasible Inference - Exponential GDP - Heteroskedastic Errors (Case 1)

Local Linear Estimator Empirical Coverage

DGP: Exponential Model – Heteroskedastic Errors (Case 1) – Kernel: Rectangular
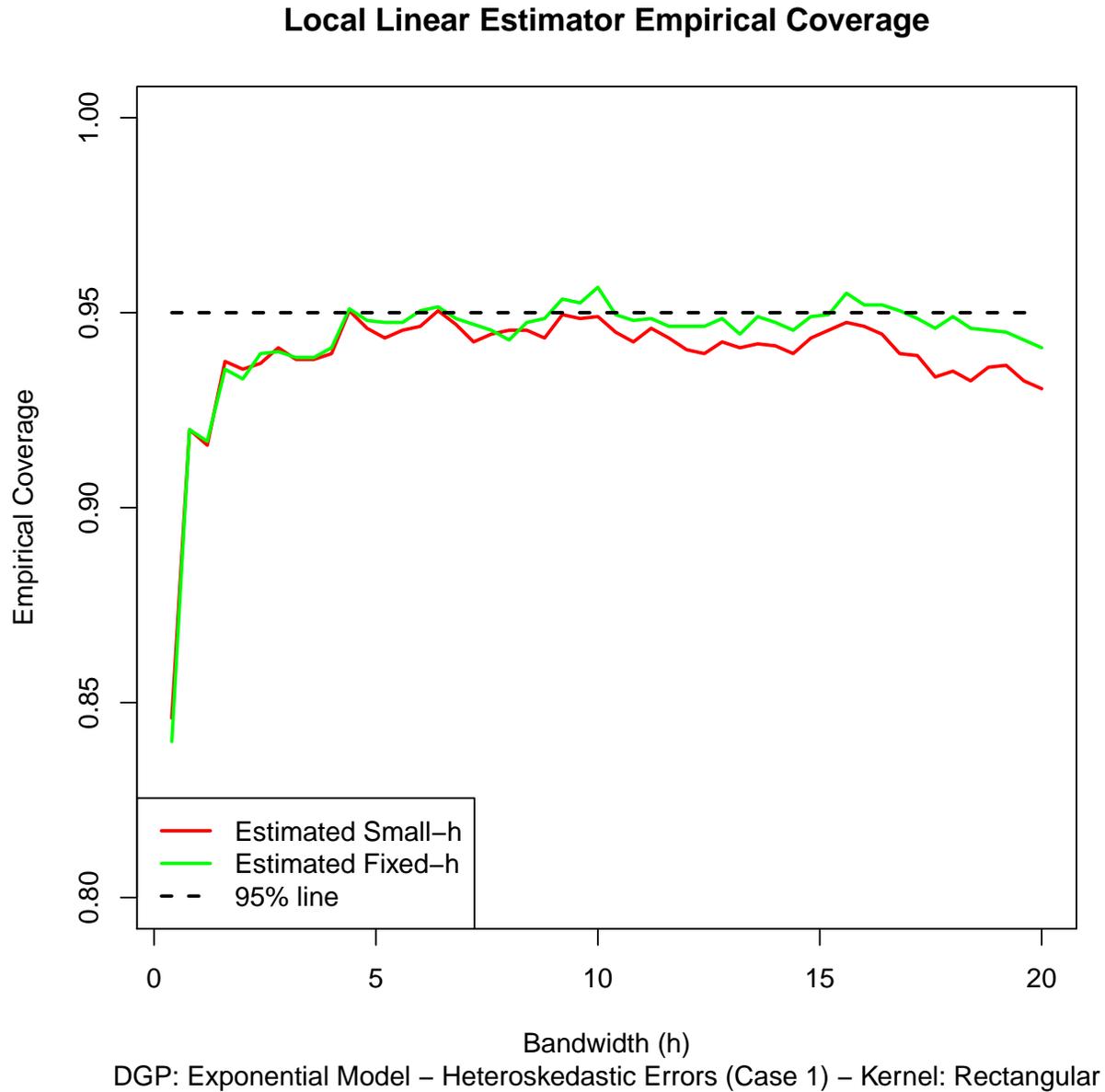
Figure 14: Simulation for Feasible Inference - Linear GDP - Heteroskedastic Errors (Case 2)

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Linear Model – Heteroskedastic Errors (Case 2) – Kernel: Rectangular
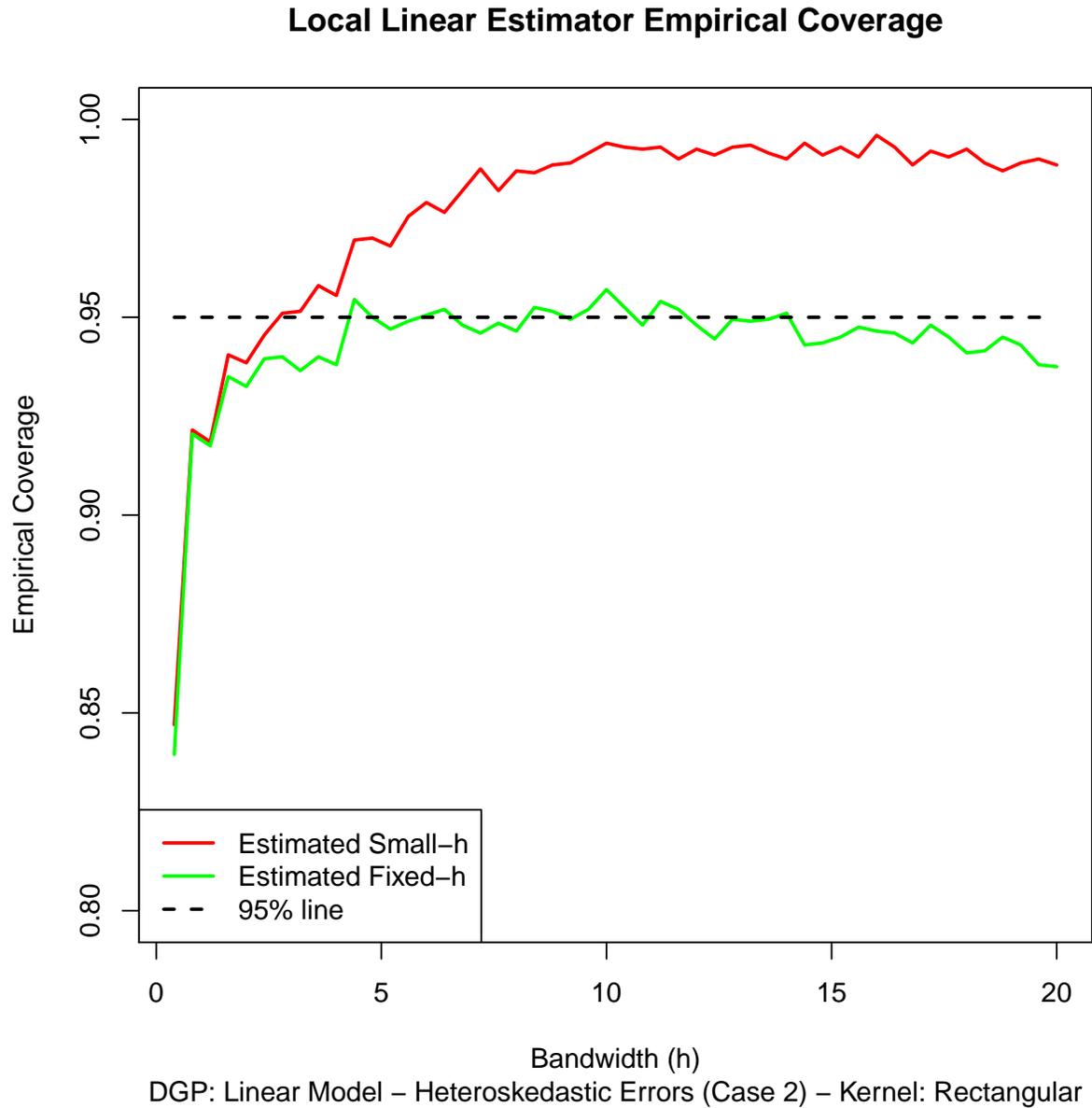
Figure 15: Simulation for Feasible Inference - Exponential GDP - Heteroskedastic Errors (Case 2)

Local Linear Estimator Empirical Coverage

DGP: Exponential Model – Heteroskedastic Errors (Case 2) – Kernel: Rectangular

Local Linear Estimator Empirical Coverage

Bandwidth (h)

DGP: Lee (2008) – Heteroskedastic Errors – Kernel: Rectangular

Figure 17: Simulation for Feasible Inference - Bandwidth Choice for $\hat{f}_o(\bar{x})$

Small−h Sensitivity to Density Bandwidth (hf)

Bandwidth (h)
Local Linear Estimator – DGP: Linear Model – Homoskedastic Errors
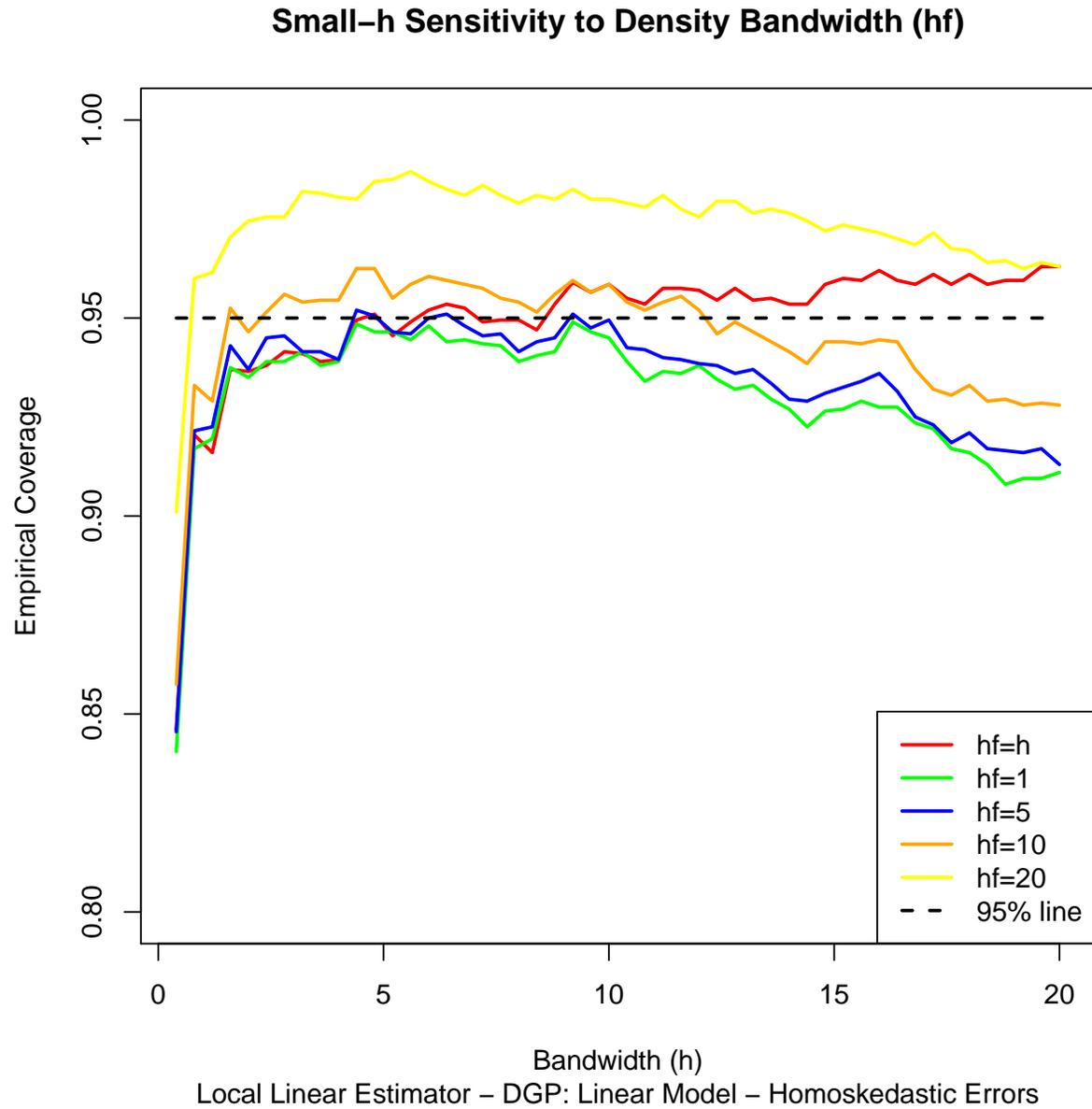
8 Appendix

Table 2: Electoral outcomes and pre-determined election characteristics: 1948-1996 - Lee (2008)

Variable	Nadaraya-Watson Estimator ($p=0$)			Local Linear Estimator ($p=1$)			Local Polynomial Estimator ($p=4$)		
	All	$Margin\leq 0.5$	$Margin\leq 0.05$	All	$Margin\leq 0.5$	$Margin\leq 0.05$	All	$Margin\leq 0.5$	$Margin\leq 0.05$
Democrat vote share $t+1$	0.351	0.257	0.096	0.118	0.090	0.048	0.077	0.066	0.105
(fixed-h std. error)	(0.0041)	(0.0038)	(0.0090)	(0.0056)	(0.0062)	(0.0159)	(0.0114)	(0.0144)	(0.0312)
[small-h std. error]	[0.0041]	[0.0038]	[0.0090]	[0.0068]	[0.0071]	[0.0180]	[0.0167]	[0.0179]	[0.0447]
Democrat vote share $t-1$	0.313	0.215	0.027	0.066	0.032	0.010	-0.004	0.002	0.104
(fixed-h std. error)	(0.0042)	(0.0040)	(0.0106)	(0.0059)	(0.0070)	(0.0214)	(0.0134)	(0.0177)	(0.0431)
[small-h std. error]	[0.0043]	[0.0040]	[0.0106]	[0.0070]	[0.0074]	[0.0211]	[0.0173]	[0.0184]	[0.0525]
Democrat win prob. $t-1$	0.780	0.724	0.137	0.562	0.335	0.160	-0.0116	0.004	0.545
(fixed-h std. error)	(0.0077)	(0.0097)	(0.0392)	(0.0159)	(0.0228)	(0.0802)	(0.0436)	(0.0609)	(0.1809)
[small-h std. error]	[0.0076]	[0.0097]	[0.0391]	[0.0146]	[0.0181]	[0.0782]	[0.0342]	[0.0440]	[0.1944]
Democrat Political Exper.	3.551	3.246	0.673	2.459	1.101	0.584	0.035	0.223	1.627
(fixed-h std. error)	(0.0657)	(0.0796)	(0.2070)	(0.1073)	(0.1412)	(0.4202)	(0.2537)	(0.3491)	(0.8309)
[small-h std. error]	[0.0739]	[0.0820]	[0.2102]	[0.1459]	[0.1576]	[0.4203]	[0.3570]	[0.3924]	[1.0448]
Opposition Political Exper.	-2.631	-2.458	-0.162	-1.676	-0.729	-0.417	0.131	0.145	-0.365
(fixed-h std. error)	(0.0563)	(0.0840)	(0.1648)	(0.0897)	(0.1153)	(0.3643)	(0.2072)	(0.2840)	(0.8260)
[small-h std. error]	[0.0492]	[0.0608]	[0.1651]	[0.0966]	[0.1171]	[0.3291]	[0.2363]	[0.2913]	[0.8189]
Democrat Electoral Exper.	3.480	3.200	0.673	2.405	1.083	0.552	0.0149	0.240	1.557
(fixed-h std. error)	(0.0673)	(0.0813)	(0.2118)	(0.1098)	(0.1446)	(0.4309)	(0.2593)	(0.3566)	(0.8755)
[small-h std. error]	[0.0750]	[0.0835]	[0.214]	[0.1485]	[0.1614]	[0.4289]	[0.3641]	[0.4021]	[1.0656]
Opposition Electoral Exper.	-2.607	-2.415	-0.154	-1.623	-0.713	-0.440	0.122	0.120	-0.054
(fixed-h std. error)	(0.0575)	(0.0638)	(0.1692)	(0.0915)	(0.1179)	(0.3753)	(0.2126)	(0.2917)	(0.8539)
[small-h std. error]	[0.0505]	[0.0622]	[0.1693]	[0.0993]	[0.1201]	[0.3372]	[0.2434]	[0.2989]	[0.8394]
Observations	6558	4900	610	6558	4900	610	6558	4900	610

8.1 Fuzzy Regression Discontinuity Design

In the fuzzy design the probability of receiving treatment still changes discontinuously at the threshold, but is not required to go from 0 to 1, allowing for a smaller jump in the probability of receiving treatment at the cutoff,

$$\lim_{x \downarrow \overline{x}} \Pr(d \mid X = x) \neq \lim_{x \uparrow \overline{x}} \Pr(d \mid X = x)$$

This framework allows for a greater range of applications since it includes cases in which the incentives to receive (or assign) treatment change discontinuously at the threshold, but are not strong enough to induce all individuals above it to be treated (and those below not to be treated). The average treatment effect at the cutoff can be identified by the ratio of the change in the conditional expectation for the outcome variable to the change in the conditional probability of receiving treatment (Imbens and Lemieux, 2008):

$$\tau_F \equiv \frac{\lim_{x \downarrow \overline{x}} E\left[Y \mid X = x\right] - \lim_{x \uparrow \overline{x}} E\left[Y \mid X = x\right]}{\lim_{x \downarrow \overline{x}} E\left[d \mid X = x\right] - \lim_{x \uparrow \overline{x}} E\left[d \mid X = x\right]}$$

This parameter's interpretation is closely linked to the instrumental variables approach. As emphasized by Hahn, Todd and van der Klaauw (2001), Imbens and Lemieux (2008) and Lee and Lemieux (2009), a causal interpretation of this ratio requires the same assumptions for local average treatment effects (LATE) presented in Imbens and Angrist (1994). For that assume monotonicity, i.e., that the treatment status is non-increasing in the cutoff value, or, as stated by Lee and Lemieux (2009, p. 23): "(...)X crossing the cutoff cannot simultaneously cause some units to take up and others to reject the treatment." Also, crossing the cutoff cannot affect the outcome other than by the receipt of treatment, otherwise we would erroneously attribute changes in the conditional expectation of Y due to changes in X to the treatment.

Under these additional assumptions, τ_F has an interpretation similar to the IV estimator, the average treatment effect for the individuals at the threshold (due to the RD design) and only for those whose participation on treatment was affected by the cutoff. Those individuals are described as compliers in the Average Treatment Effect literature[30]. Hence (Imbens and Lemieux, 2008),

$$\tau_F \equiv E\left[Y_1 - Y_0 \mid \text{individual is a complier and } X = \overline{x}\right]$$

Similarly as in the sharp RD design, Lee and Lemieux (2009) show that the fuzzy RD design estimator can be interpreted as a weighted LATE with an individual's weight directly proportional to the ex ante likelihood that an individual's realization of X will be close to the threshold.

In the Fuzzy RD design the estimator of the parameter of interest is given by the ratio

$$\widehat{\tau}_F = \frac{\widehat{\alpha}}{\widehat{\theta}}$$

where $\widehat{\alpha}$ is any of the estimators described in the section 2.2 and $\widehat{\theta}$ is the estimator for the change in the

---

[30] See Imbens and Angrist (1994), Hahn, Todd and van der Klaauw (2001), Imbens and Lemieux (2008) and Lee and Lemieux (2009).

probability of being in the treated group at the cutoff. Note that $\widehat{\theta}$ is obtained by using the estimators described in the main text with the treatment assignment variable, D_i, as the dependent variable.

To obtain the asymptotic distribution of the fuzzy RD estimator, the delta method can be used, similarly to the result in Porter (2003).

Theorem 3 *If*

$$
\begin{pmatrix} \sqrt{nh}(\widehat{\alpha} - \alpha^*) \\ \sqrt{nh}(\widehat{\theta} - \theta^*) \end{pmatrix} \xrightarrow{d} N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} V_\alpha & C_{\alpha\theta} \\ C_{\alpha\theta} & V_\theta \end{bmatrix} \right)
$$

then

$$
\sqrt{nh} \left(\frac{\widehat{\alpha}}{\widehat{\theta}} - \frac{\alpha^*}{\theta^*} \right) \xrightarrow{d} N \left(0, \frac{1}{\theta^{*2}} V_\alpha - 2\frac{\alpha^*}{\theta^{*3}} C_{\alpha\theta} + \frac{\alpha^{*2}}{\theta^{*4}} V_\theta \right)
$$

where $\alpha^* = \alpha + B_\alpha$, $\theta^* = \theta + B_\theta$ and B_α and B_θ are the bias terms for the estimators as defined in theorem 1 for local polynomial estimators.

The proof of the proposition follows directly from the Delta Method and is omitted. The condition of multivariate normality required in this proposition follows from usual multivariate central limit theorem using a Cramer-Wold device (James, 2004 and Pagan and Ullah, 1999). Note that,

$$
\begin{aligned}
\frac{\alpha^*}{\theta^*} &= \frac{\alpha + B_\alpha}{\theta + B_\theta} \\
&= \frac{\alpha + B_\alpha}{\theta} \frac{\theta}{\theta + B_\theta} \\
&= \frac{\alpha}{\theta} \frac{\theta}{\theta + B_\theta} + \frac{B_\alpha}{\theta} \frac{\theta}{\theta + B_\theta}
\end{aligned}
\tag{11}
$$

for given values of α and θ, if $|\theta| < |\theta + B_\theta|$ then $0 \leq \left| \frac{\theta}{\theta + B_\theta} \right| \leq 1$. Clearly, if there is no bias in the estimate for α or θ, i.e., $B_\alpha = 0$ and $B_\theta = 0$, the fuzzy design RD estimator will be consistent for the true treatment effect. If $B_\alpha = 0$ and $B_\theta \neq 0$, the estimator will suffer an attenuation bias and tests for the null hypotheses that the treatment is unimportant will be conservative. If $B_\alpha \neq 0$ and $B_\theta = 0$, the estimator's bias is similar to the one seem for the sharp RD design, only being scaled by $\frac{1}{\theta}$. Finally, if $B_\alpha \neq 0$ and $B_\theta \neq 0$, any increase in B_α increases the bias in the ATE estimator but there will be a trade-off regarding the size of B_θ since its impact in the first and second terms will be in opposite directions.

All the terms that appear in the asymptotic distribution above, except for $C_{\alpha\theta}$, can be obtained from theorem 1 by using local polynomial estimators discussed in section 2.2.

Theorem 4 *Suppose $\sigma_{\varepsilon\eta} = E\left[\varepsilon\eta \mid X = x\right]$ is continuous for $x \neq \overline{x}$, $x \in \aleph$ and the left and right-hand limits at $\overline{x}$ exist. If $\widehat{\alpha}$ and $\widehat{\theta}$ are the local polynomial estimators and the conditions of theorem 1 hold for both estimators, then*

$$
C_{\alpha\theta} = e_1' \left[\left(\Gamma_+^*\right)^{-1} \Delta_+^\rho \left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1} \Delta_-^\rho \left(\Gamma_-^*\right)^{-1} \right] e_1
$$

where $\Delta_{+(-)}^\rho = \begin{bmatrix} \rho_0^{+(-)} & \cdots & \rho_p^{+(-)} \\ \vdots & \ddots & \vdots \\ \rho_p^{+(-)} & \cdots & \rho_{2p}^{+(-)} \end{bmatrix}$,

$\rho_j^+ = \int_0^\infty k^2\left(u\right)u^j \sigma_{\varepsilon\eta}(\overline{x} + uh)f_o(\overline{x} + uh)du,$

$\rho_j^- = (-1)^j \int_0^\infty k^2\left(u\right)u^j \sigma_{\varepsilon\eta}(\overline{x} - uh)f_o(\overline{x} - uh)du,$ Γ_+^* and Γ_-^* are defined as in previous results.

Similarly to the result in corollary 1, as $h \to 0$, the fixed-h covariance formulas converges to the small-h asymptotic covariance.

Corollary 3 *If* $h \longrightarrow 0$*, then the asymptotic covariance,* $C_{\alpha\theta}$*, obtained by fixed-h (theorem 4) and small-h (Porter 2003) are the same:*

$$\lim_{h \longrightarrow 0} C_{\alpha\theta} = \frac{\sigma_{\varepsilon\eta}^+(\overline{x}) + \sigma_{\varepsilon\eta}^-(\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1$$

Also, a result similar to the corollary 2 is readily available.

Corollary 4 *If in the bandwidth around the cutoff,* $f_o(x)$ *and* $\sigma_{\varepsilon\eta}(x)$ *are constant, then the asymptotic covariance,* $C_{\alpha\theta}$*, obtained by fixed-h (theorem 4) and small-h (Porter 2003) are the same.*

$$C_{\alpha\theta} = \frac{\sigma_{\varepsilon\eta}^+(\overline{x}) + \sigma_{\varepsilon\eta}^-(\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1$$

Similarly to the sharp discontinuity case the improvements obtained by fixed-h arise from incorporating the behavior of $\sigma_{\varepsilon\eta}(x)$ and $f_o(x)$ in the range around the cutoff while small-h does not.

8.2 Proofs

Proof of Theorem 1. The local polynomial estimator is given by

$$\widehat{\alpha}_p = \widehat{\alpha}_{p+} - \widehat{\alpha}_{p-}$$

note that,

$$\begin{aligned}
\widehat{\alpha}_{p+} &= e_1' \left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i Z_i'\right]^{-1}\left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i y_i\right] \\
&= e_1' D_{n+}\left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i y_i\right]
\end{aligned}$$

where $D_{n+} = \left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i Z_i'\right]^{-1}$. Similarly, for $D_{n-} = \left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)(1 - d_i)Z_i Z_i'\right]^{-1}$

$$\widehat{\alpha}_{p-} = e_1' D_{n-}\left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)(1 - d_i)Z_i y_i\right]$$

Then,

$$\begin{aligned}
\widehat{\alpha}_{p+} &= e_1' D_{n+}\left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i\left[m(x_i) + \alpha d_i + \varepsilon_i\right]\right] \\
&= e_1' D_{n+}\left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i m(x_i) + \frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i \alpha + \frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i \varepsilon_i\right] \\
&= e_1' D_{n+}\left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i m(x_i)\right] + \alpha e_1' D_{n+}\left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i\right] + \\
&\quad + e_1' D_{n+}\left[\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right)d_i Z_i \varepsilon_i\right]
\end{aligned}$$

note that $Z_i = Z_i Z_i' e_1$, $e_1' e_1 = 1$ then

$$e_1' D_{n+} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i \right] = e_1' D_{n+} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i Z_i' \right] e_1 = 1$$

and

$$\widehat{\alpha}_{p+} - \alpha = e_1' D_{n+} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i m(x_i) \right] + e_1' D_{n+} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i \varepsilon_i \right]$$

similarly

$$\widehat{\alpha}_{p-} = e_1' D_{n-} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) (1 - d_i) Z_i m(x_i) \right] + e_1' D_{n-} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) (1 - d_i) Z_i \varepsilon_i \right]$$

Then

$$
\begin{aligned}
\sqrt{nh}(\widehat{\alpha}_p - \alpha) &= \sqrt{nh}(\widehat{\alpha}_{p+} - \alpha - \widehat{\alpha}_{p-}) \\
&= e_1' D_{n+} \left\{ \sqrt{nh} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i m(x_i) \right] + \left[\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i \varepsilon_i \right] \right\} - \\
&\quad - e_1' D_{n-} \left\{ \begin{array}{l} \sqrt{nh} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) (1 - d_i) Z_i m(x_i) \right] + \\ + \left[\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) (1 - d_i) Z_i \varepsilon_i \right] \end{array} \right\}
\end{aligned}
$$

For the denominator terms D_{n+} and D_{n-},

$$D_{n+}^{-1} = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i Z_i'$$

and each element of this matrix is given by

$$\left[D_{n+}^{-1} \right]_{l,j} = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2}$$

which has asymptotic variance converging to zero since

$$
\begin{aligned}
Var\left(\left[D_{n+}^{-1} \right]_{j,l} \right) &= \frac{1}{(nh)^2} Var\left(\sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2} \right) \\
&= \frac{1}{nh^2} Var\left(k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2} \right) \\
&\leq \frac{1}{nh} E\left[\frac{1}{h} k^2\left(\frac{x - \overline{x}}{h}\right) d \left(\frac{x - \overline{x}}{h}\right)^{2(j+l-2)} \right] \\
&= \frac{1}{nh} \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h} k^2\left(\frac{x - \overline{x}}{h}\right) \left(\frac{x - \overline{x}}{h}\right)^{2(j+l-2)} f_o(x) dx
\end{aligned}
$$

Note that the terms in the integral and the integral itself are $O(1)$ and $\frac{1}{nh} = o(1)$. Hence, $Var\left(\left[D_{n+}^{-1} \right]_{l,j} \right) \to$

0. Now,

$$
\begin{aligned}
\left[D_{n+}^{-1}\right]_{l,j} &= E\left\{\left[D_{n+}^{-1}\right]_{l,j}\right\} + o_p(1) \\
&= \frac{1}{nh}E\left[\sum_{i=1}^{n} k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2}\right] + o_p(1) \\
&= E\left[\frac{1}{h}k\left(\frac{x_i - \overline{x}}{h}\right) d_i \left(\frac{x_i - \overline{x}}{h}\right)^{j+l-2}\right] + o_p(1) \\
&= \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h}k\left(\frac{x - \overline{x}}{h}\right) \left(\frac{x - \overline{x}}{h}\right)^{j+l-2} f_o(x)dx + o_p(1) \\
&= \int_0^{\infty} k\left(u\right) u^{j+l-2} f_o(\overline{x} + uh)du + o_p(1)
\end{aligned}
$$

Let, $\gamma_j^+ = \int_0^\infty k\left(u\right) u^j f_o(\overline{x} + uh)du$ and Γ_+^* is the $(p+1) \times (p+1)$ matrix with (j,l) element γ_{j+l-2}^+ for $j,l = 1, ..., p+1$. Then,

$$
D_{n+} \xrightarrow{p} \left(\Gamma_+^*\right)^{-1}
$$

Similarly,

$$
D_{n-} \xrightarrow{p} \left(\Gamma_-^*\right)^{-1}
$$

where Γ_-^* is the $(p+1) \times (p+1)$ matrix with (j,l) element γ_{j+l-2}^- for $j,l = 1, ..., p+1$, and $\gamma_j^- = (-1)^j \int_0^\infty k\left(u\right) u^j f_o(\overline{x} - uh)du$.

Now we will derive the asymptotic distribution of $\frac{1}{\sqrt{nh}}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right) d_i Z_i \varepsilon_i$. Following Porter (2003) I use the Cramer-Wold device to derive the asymptotic distribution. Let λ be a nonzero, finite vector. Then,

$$
\begin{aligned}
&E\left[\left|\frac{1}{\sqrt{nh}}\sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h}\right) d_i \lambda' Z_i \varepsilon_i\right|^{2+\varsigma}\right] \\
&= \sum_{i=1}^n \left(\frac{1}{nh}\right)^{\frac{\varsigma}{2}} \frac{1}{nh} E\left[\left|k\left(\frac{x_i - \overline{x}}{h}\right)\right|^{2+\varsigma} d_i \left|\lambda' Z_i\right|^{2+\varsigma} \left|\varepsilon_i\right|^{2+\varsigma}\right] \\
&= \left(\frac{1}{nh}\right)^{\frac{\varsigma}{2}} \frac{1}{h} E\left[\left|k\left(\frac{x - \overline{x}}{h}\right)\right|^{2+\varsigma} d \left|\sum_{l=1}^p \lambda_l \left(\frac{x - \overline{x}}{h}\right)^l\right|^{2+\varsigma} E\left[\left|\varepsilon\right|^{2+\varsigma} \mid X = x\right]\right] \\
&\leq \left(\frac{1}{nh}\right)^{\frac{\varsigma}{2}} \frac{1}{h} E\left[\left|k\left(\frac{x - \overline{x}}{h}\right)\right|^{2+\varsigma} d \left(\sum_{l=1}^p \left|\lambda_l \left(\frac{x - \overline{x}}{h}\right)^l\right|\right)^{2+\varsigma} E\left[\left|\varepsilon\right|^{2+\varsigma} \mid X = x\right]\right] \\
&\leq \left(\frac{1}{nh}\right)^{\frac{\varsigma}{2}} \frac{1}{h} \sup_{x \in \aleph}\left\{E\left[\left|\varepsilon\right|^{2+\varsigma} \mid X = x\right]\right\} E\left[\left|k\left(\frac{x - \overline{x}}{h}\right)\right|^{2+\varsigma} d \sum_{l=1}^p \left|\lambda_l \left(\frac{x - \overline{x}}{h}\right)^l\right|^{2+\varsigma}\right] \\
&= \left(\frac{1}{nh}\right)^{\frac{\varsigma}{2}} \frac{1}{h} \sup_{x \in \aleph}\left\{E\left[\left|\varepsilon\right|^{2+\varsigma} \mid X = x\right]\right\} \int_{\overline{x}}^{\overline{x}+h} \left|k\left(\frac{x - \overline{x}}{h}\right)\right|^{2+\varsigma} \sum_{l=1}^p \left|\lambda_l \left(\frac{x - \overline{x}}{h}\right)^l\right|^{2+\varsigma} f_o(x)dx \\
&= \left(\frac{1}{nh}\right)^{\frac{\varsigma}{2}} O(1)O(1) = O\left(\left(\frac{1}{nh}\right)^{\frac{\varsigma}{2}}\right) = o(1)
\end{aligned}
$$

then, $\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i \varepsilon_i$ follows Liapunov's CLT. Note that,

$$E\left[\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i \varepsilon_i\right]$$
$$= E\left[\frac{\sqrt{n}}{\sqrt{h}} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i E\left[\varepsilon \mid X = x\right]\right] = 0$$

and

$$Var\left[\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i \varepsilon_i\right]$$
$$= \frac{1}{h} Var\left[k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i \varepsilon_i\right]$$
$$= \frac{1}{h} E\left[k^2\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i Z_i' \varepsilon_i^2\right]$$
$$= \frac{1}{h} E\left[k^2\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i Z_i' E\left[\varepsilon_i^2 \mid X = x\right]\right]$$
$$= \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h} k^2\left(\frac{x-\overline{x}}{h}\right) Z Z' \sigma^2(x) f_o(x) dx$$

It helps to remember that $Z_i Z_i'$ is a function of the x,

$$Z_i Z_i' = \begin{bmatrix} 1 & \left(\frac{x_i-\overline{x}}{h}\right) & \cdots & \left(\frac{x_i-\overline{x}}{h}\right)^p \\ \left(\frac{x_i-\overline{x}}{h}\right) & \left(\frac{x_i-\overline{x}}{h}\right)^2 & \cdots & \left(\frac{x_i-\overline{x}}{h}\right)^{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \left(\frac{x_i-\overline{x}}{h}\right)^p & \left(\frac{x_i-\overline{x}}{h}\right)^{p+1} & \cdots & \left(\frac{x_i-\overline{x}}{h}\right)^{2p} \end{bmatrix}$$

Let $\delta_j^+ = \int_{\overline{x}}^{\overline{x}+h} \frac{1}{h} k^2\left(\frac{x-\overline{x}}{h}\right) \left(\frac{x-\overline{x}}{h}\right)^j \sigma^2(x) f_o(x) dx = \int_0^\infty k^2(u) u^j \sigma^2(\overline{x}+uh) f_o(\overline{x}+uh) du$ and Δ_+^* is the $(p+1) \times (p+1)$ matrix with (j,l) element δ_{j+l-2}^+ for $j,l = 1,...,p+1$. Then,

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i \varepsilon_i \xrightarrow{p} N(0, \Delta_+^*)$$

Similarly we can show that

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) (1-d_i) Z_i \varepsilon_i \xrightarrow{p} N(0, \Delta_-^*)$$

where Δ_-^* is the $(p+1) \times (p+1)$ matrix with (j,l) element δ_{j+l-2}^- for $j,l = 1,...,p+1$, and $\delta_j^- = \int_{\overline{x}-h}^{\overline{x}} \frac{1}{h} k^2\left(\frac{x-\overline{x}}{h}\right) \left(\frac{x-\overline{x}}{h}\right)^j \sigma^2(x) f_o(x) dx = (-1)^j \int_0^\infty k^2(u) u^j \sigma^2(\overline{x}-uh) f_o(\overline{x}-uh) du$.

The bias term is given by

$$\sqrt{nh} e_1' \left\{ D_{n+} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) d_i Z_i m(x_i)\right] - D_{n-} \left[\frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x_i-\overline{x}}{h}\right) (1-d_i) Z_i m(x_i)\right]\right\}$$

Notice that if the rectangular kernel is used this is nothing else than the difference between the intercepts estimated by the linear projection of $m(x)$ on Z, above and below the cutoff point using only the data inside the bandwidth.

Note that,

$$E\left[\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{x_i-\overline{x}}{h}\right)d_iZ_im(x_i)\right] = E\left[\frac{1}{h}k\left(\frac{x_i-\overline{x}}{h}\right)d_iZ_im(x_i)\right]$$

$$= \int_{\overline{x}}^{\overline{x}+h}\frac{1}{h}k\left(\frac{x-\overline{x}}{h}\right)Z(x)m(x)f(x)dx$$

$$= \int_{0}^{\infty}k\left(u\right)Z(\overline{x}+uh)m(\overline{x}+uh)f(\overline{x}+uh)du$$

and similarly,

$$\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i)Z_im(x_i) \xrightarrow{p} \int_{\overline{x}-h}^{\overline{x}}\frac{1}{h}k\left(\frac{x-\overline{x}}{h}\right)Z(x)m(x)dx$$

Hence, the bias term can be approximated by,

$$e_1'\left\{\begin{array}{l}\left(\Gamma_+^*\right)^{-1}\left[\int_0^\infty k\left(u\right)Z(\overline{x}+uh)m(\overline{x}+uh)f(\overline{x}+uh)du\right]-\\ -\left(\Gamma_-^*\right)^{-1}\left[\int_0^\infty k\left(u\right)Z(\overline{x}-uh)m(\overline{x}-uh)f(\overline{x}-uh)du\right]\end{array}\right\}$$

■

Proof of Corollary 1. First, note that, if $h\to 0$,

$$\gamma_j^+ = \lim_{h\to 0}\int_0^\infty k\left(u\right)u^jf_o(\overline{x}+uh)du$$

$$= f_o(\overline{x})\int_0^\infty k\left(u\right)u^jdu \qquad (12)$$

$$= f_o(\overline{x})\gamma_j$$

and

$$\delta_j^+ = \lim_{h\to 0}\int_0^\infty k^2\left(u\right)u^j\sigma^2(\overline{x}+uh)f_o(\overline{x}+uh)du$$

$$= \sigma^{2+}(\overline{x})f_o(\overline{x})\int_0^\infty k^2\left(u\right)u^jdu \qquad (13)$$

$$= \sigma^{2+}(\overline{x})f_o(\overline{x})\delta_j$$

and similarly for γ_j^- and δ_j^-. Then, for the variance,

$$\lim_{h\to 0}\left(\Gamma_+^*\right)^{-1}\Delta_+^*\left(\Gamma_+^*\right)^{-1}+\left(\Gamma_-^*\right)^{-1}\Delta_-^*\left(\Gamma_-^*\right)^{-1}$$

$$= (f_o(\overline{x})\Gamma)^{-1}\left[\sigma^{2+}(\overline{x})f_o(\overline{x})\Delta\right](f_o(\overline{x})\Gamma)^{-1}+(f_o(\overline{x})\Gamma)^{-1}\left[\sigma^{2-}(\overline{x})f_o(\overline{x})\Delta\right](f_o(\overline{x})\Gamma)^{-1}$$

$$= \frac{\sigma^{2+}(\overline{x})+\sigma^{2-}(\overline{x})}{f_o(\overline{x})}e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1$$

For the bias, if we approximate $m(\overline{x}+uh)=m(x)$ just above $m(\overline{x})$:

$$m(x) = m(\overline{x})+m'^+(\overline{x})(x-\overline{x})+...+\frac{1}{p!}m^{(p)+}(\overline{x})(x-\overline{x})^p+\frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})(x-\overline{x})^{p+1}+o\left(h^{p+1}\right)$$

$$= LP^+(m(x)\text{ on }Z(x))+\frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})(x-\overline{x})^{p+1}+o\left(h^{p+1}\right)$$

and similarly for approximating $m(x)$ just below the cutoff. When we evaluate $LP^+(m(x)\text{ on }Z(x))$ at $\overline{x}$, we get the intercept $m(\overline{x})$ and the "residual" as described above. A helpful fact is that, by the definition of

$Z(x),$

$$\int_0^\infty k(u) Z(\overline{x}+uh)u^{p+1}du \;=\; \int_0^\infty k(u) \begin{bmatrix} 1 \\ \vdots \\ u^p \end{bmatrix} u^{p+1}du = \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix} \tag{14}$$

$$\int_0^\infty k(u) Z(\overline{x}-uh)u^{p+1}du \;=\; \int_0^\infty k(u) \begin{bmatrix} 1 \\ \vdots \\ (-u)^p \end{bmatrix} u^{p+1}du = \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ (-1)^p\gamma_{2p+1} \end{bmatrix} \tag{15}$$

Note that $\Gamma^{-1}\begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix}$ is equal both above and below the cutoff. The bias formula in theorem 1 is given by

$$e_1' \left\{ \begin{array}{l} \left(\Gamma_+^*\right)^{-1}\left[\int_0^\infty k(u)Z(\overline{x}+uh)m(\overline{x}+uh)f_o(\overline{x}+uh)du\right] - \\ -\left(\Gamma_-^*\right)^{-1}\left[\int_0^\infty k(u)Z(\overline{x}-uh)m(\overline{x}-uh)f_o(\overline{x}-uh)du\right] \end{array} \right\} \tag{16}$$

as discussed in section 3 the main term is just the difference between the intercepts of the linear projections of $k(u)m(x)$ on $k(u)Z(x)$ in the bandwidth above below the cutoff, which is equal to the linear projections evaluated at $\overline{x}$. Hence, plugging the bias formula for the linear projection, formula (16) can be written as

$$e_1' \left\{ \begin{array}{l} \left[\begin{pmatrix} m(\overline{x}) \\ m_{p+} \end{pmatrix} + \left(\Gamma_+^*\right)^{-1}\int_0^\infty k(u)Z(\overline{x}+uh)\left(\frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})(uh)^{p+1} + o\left(h^{p+1}\right)\right)f_o(\overline{x}+uh)du\right] \\ -\left[\begin{pmatrix} m(\overline{x}) \\ m_{p-} \end{pmatrix} + \left(\Gamma_-^*\right)^{-1}\int_0^\infty k(u)Z(\overline{x}-uh)\left(\frac{(-1)^{p+1}}{(p+1)!}m^{(p+1)-}(\overline{x})(uh)^{p+1} + o\left(h^{p+1}\right)\right)f_o(\overline{x}-uh)du\right] \end{array} \right\}$$

$$= e_1' \left\{ \begin{array}{l} \left[\left(\Gamma_+^*\right)^{-1}\int_0^\infty k(u)Z(\overline{x}+uh)\left(\frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})(uh)^{p+1} + o\left(h^{p+1}\right)\right)f_o(\overline{x}+uh)du\right] \\ -\left[\left(\Gamma_-^*\right)^{-1}\int_0^\infty k(u)Z(\overline{x}-uh)\left(\frac{(-1)^{p+1}}{(p+1)!}m^{(p+1)-}(\overline{x})(uh)^{p+1} + o\left(h^{p+1}\right)\right)f_o(\overline{x}-uh)du\right] \end{array} \right\}$$

$$= \frac{h^{p+1}}{(p+1)!}e_1' \left\{ \left\{ \begin{array}{l} \left[\left(\Gamma_+^*\right)^{-1}\int_0^\infty k(u)Z(\overline{x}+uh)m^{(p+1)+}(\overline{x})u^{p+1}f_o(\overline{x}+uh)du\right] \\ -\left[\left(\Gamma_-^*\right)^{-1}\int_0^\infty k(u)Z(\overline{x}-uh)\left((-1)^{p+1}m^{(p+1)-}(\overline{x})u^{p+1}\right)f_o(\overline{x}-uh)du\right] \end{array} \right\} + o(1) \right\}$$

$$= \frac{h^{p+1}}{(p+1)!}e_1' \left\{ \left\{ \begin{array}{l} m^{(p+1)+}(\overline{x})\left[\left(\Gamma_+^*\right)^{-1}\int_0^\infty k(u)Z(\overline{x}+uh)u^{p+1}f_o(\overline{x}+uh)du\right] \\ -(-1)^{p+1}m^{(p+1)-}(\overline{x})\left[\left(\Gamma_-^*\right)^{-1}\int_0^\infty k(u)Z(\overline{x}-uh)u^{p+1}f_o(\overline{x}-uh)du\right] \end{array} \right\} + o(1) \right\}$$

where $\begin{pmatrix} m(\overline{x}) \\ m_{p+(-)} \end{pmatrix}$, is the vector of coefficients of the linear projection of $m(x)$ on $Z(x)$ is the bandwidth above (below) the cutoff. If $h \to 0$, using the equalities in equations (14), (15), (13) and (12),

$$= \frac{\lim_{h\to 0}\left(h^{p+1}\right)}{(p+1)!}\left[m^{(p+1)+}(\overline{x}) - (-1)^{p+1}m^{(p+1)-}(\overline{x})\right]e_1'\Gamma^{-1}\begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix}$$

which is the same limit of the bias term in the small-h approximation. ∎

Proof of Corollary 2. First, note that, if $h > 0$ and, in the bandwidth around the cutoff, $f_o(x) = f_o(\overline{x})$, $\sigma^2(x) = \sigma^2(\overline{x})$ and

$$m(x) = m(\overline{x}) + m'^+(\overline{x})(x - \overline{x}) + ... + \frac{1}{p!}m^{(p)+}(\overline{x})(x - \overline{x})^p + \frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})(x - \overline{x})^{p+1}$$

then,

$$
\begin{aligned}
\gamma_j^+ &= \int_0^\infty k(u)u^j f_o(\overline{x} + uh)du \\
&= f_o(\overline{x})\int_0^\infty k(u)u^j du \\
&= f_o(\overline{x})\gamma_j
\end{aligned}
\tag{17}
$$

and

$$
\begin{aligned}
\delta_j^+ &= \int_0^\infty k^2(u)u^j \sigma^2(\overline{x} + uh)f_o(\overline{x} + uh)du \\
&= \sigma^{2+}(\overline{x})f_o(\overline{x})\int_0^\infty k^2(u)u^j du \\
&= \sigma^{2+}(\overline{x})f_o(\overline{x})\delta_j
\end{aligned}
\tag{18}
$$

and similarly for γ_j^- and δ_j^-. Then, for the variance,

$$
\begin{aligned}
& \left(\Gamma_+^*\right)^{-1}\Delta_+^*\left(\Gamma_+^*\right)^{-1} + \left(\Gamma_-^*\right)^{-1}\Delta_-^*\left(\Gamma_-^*\right)^{-1} \\
&= (f_o(\overline{x})\Gamma)^{-1}\left[\sigma^{2+}(\overline{x})f_o(\overline{x})\Delta\right](f_o(\overline{x})\Gamma)^{-1} + (f_o(\overline{x})\Gamma)^{-1}\left[\sigma^{2-}(\overline{x})f_o(\overline{x})\Delta\right](f_o(\overline{x})\Gamma)^{-1} \\
&= \frac{\sigma^{2+}(\overline{x}) + \sigma^{2-}(\overline{x})}{f_o(\overline{x})}e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1
\end{aligned}
$$

For the bias, the strategy is basically the same as in the proof of corollary 1:

$$
\begin{aligned}
m(x) &= m(\overline{x}) + m'^+(\overline{x})(x - \overline{x}) + ... + \frac{1}{p!}m^{(p)+}(\overline{x})(x - \overline{x})^p + \frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})(x - \overline{x})^{p+1} \\
&= LP^+(m(x) \text{ on } Z(x)) + \frac{1}{(p+1)!}m^{(p+1)+}(\overline{x})(x - \overline{x})^{p+1}
\end{aligned}
$$

and similarly for approximating $m(x)$ just below the cutoff. When we evaluate $LP^+(m(x) \text{ on } Z(x))$ at $\overline{x}$, we get the intercept $m(\overline{x})$ and the "residual" as described above. Once again, using formulas (14) and (15), the bias formula in theorem 1 is given by

$$
e_1'\left\{
\begin{array}{l}
\left(\Gamma_+^*\right)^{-1}\left[\int_0^\infty k(u)Z(\overline{x} + uh)m(\overline{x} + uh)f_o(\overline{x} + uh)du\right] - \\
-\left(\Gamma_-^*\right)^{-1}\left[\int_0^\infty k(u)Z(\overline{x} - uh)m(\overline{x} - uh)f_o(\overline{x} - uh)du\right]
\end{array}
\right\}
$$

as discussed in section 3 the main term is just the difference between the intercepts of the linear projections of $k(u)m(x)$ on $k(u)Z(x)$ in the bandwidth above below the cutoff, which is equal to the linear projections evaluated at $\overline{x}$. Hence, plugging the bias formula for the linear projection:

54

$$e_1' \left\{ \begin{array}{l} \left[\begin{pmatrix} m(\overline{x}) \\ m_{p+} \end{pmatrix} + \left(\Gamma_+^*\right)^{-1} \int_0^\infty k\left(u\right) Z(\overline{x} + uh) \left(\frac{1}{(p+1)!} m^{(p+1)+}(\overline{x}) \left(uh\right)^{p+1} \right) f_o(\overline{x} + uh) du \right] \\ - \left[\begin{pmatrix} m(\overline{x}) \\ m_{p-} \end{pmatrix} + \left(\Gamma_-^*\right)^{-1} \int_0^\infty k\left(u\right) Z(\overline{x} - uh) \left(\frac{(-1)^{p+1}}{(p+1)!} m^{(p+1)-}(\overline{x}) \left(uh\right)^{p+1} \right) f_o(\overline{x} - uh) du \right] \end{array} \right\}$$

$$= \; e_1' \left\{ \begin{array}{l} \left[\left(\Gamma_+^*\right)^{-1} \int_0^\infty k\left(u\right) Z(\overline{x} + uh) \left(\frac{1}{(p+1)!} m^{(p+1)+}(\overline{x}) \left(uh\right)^{p+1} \right) f_o(\overline{x} + uh) du \right] \\ - \left[\left(\Gamma_-^*\right)^{-1} \int_0^\infty k\left(u\right) Z(\overline{x} - uh) \left(\frac{(-1)^{p+1}}{(p+1)!} m^{(p+1)-}(\overline{x}) \left(uh\right)^{p+1} \right) f_o(\overline{x} - uh) du \right] \end{array} \right\}$$

$$= \; \frac{h^{p+1}}{(p+1)!} e_1' \left\{ \begin{array}{l} \left[\left(\Gamma_+^*\right)^{-1} \int_0^\infty k\left(u\right) Z(\overline{x} + uh) m^{(p+1)+}(\overline{x}) u^{p+1} f_o(\overline{x} + uh) du \right] \\ - \left[\left(\Gamma_-^*\right)^{-1} \int_0^\infty k\left(u\right) Z(\overline{x} - uh) \left((-1)^{p+1} m^{(p+1)-}(\overline{x}) u^{p+1} \right) f_o(\overline{x} - uh) du \right] \end{array} \right\}$$

$$= \; \frac{h^{p+1}}{(p+1)!} e_1' \left\{ \begin{array}{l} m^{(p+1)+}(\overline{x}) \left[\left(\Gamma_+^*\right)^{-1} \int_0^\infty k\left(u\right) Z(\overline{x} + uh) u^{p+1} f_o(\overline{x} + uh) du \right] \\ - (-1)^{p+1} m^{(p+1)-}(\overline{x}) \left[\left(\Gamma_-^*\right)^{-1} \int_0^\infty k\left(u\right) Z(\overline{x} - uh) u^{p+1} f_o(\overline{x} - uh) du \right] \end{array} \right\}$$

where $\begin{pmatrix} m(\overline{x}) \\ m_{p+(-)} \end{pmatrix}$, is the vector of coefficients of the linear projection of $m(x)$ on $Z(x)$ is the bandwidth above (below) the cutoff. Using $f_o(x) = f_o(\overline{x})$ and the equalities in formulas (14), (15), (18) and (17),

$$= \frac{h^{p+1}}{(p+1)!} \left[m^{(p+1)+}(\overline{x}) - (-1)^{p+1} m^{(p+1)-}(\overline{x}) \right] e_1' \Gamma^{-1} \begin{bmatrix} \gamma_{p+1} \\ \vdots \\ \gamma_{2p+1} \end{bmatrix}$$

∎

Proof of Theorem 4. To obtain the Covariance term for the asymptotic variance of the Fuzzy Regression Discontinuity estimator, note that the covariance will be determined by the expectation of the product of the stochastic terms.

The covariance between the estimators for the outcome of interest and the treatment probability will be given by two independent terms, one for each side of the threshold. The upper side is given by

$$E \left\{ e_1' D_{n+} \left[\frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h} \right) d_i Z_i y_i \right] \left[\frac{1}{nh} \sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h} \right) d_i Z_i t_i \right] D_{n+} e_1 \right\}$$

Where t_i is the dummy variable indicating that the observation has received treatment. In obtaining the asymptotic covariance, the bias term of the estimator can be ignored, hence

$$E \left[e_1' D_{n+} \left(\frac{1}{\sqrt{nh}} \sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h} \right) d_i Z_i \varepsilon_i \right) \left(\frac{1}{\sqrt{nh}} \sum_{i=1}^n k\left(\frac{x_i - \overline{x}}{h} \right) d_i Z_i \eta_i \right)' D_{n+} e_1 \right]$$

$$= \; E \left[e_1' D_{n+} \left(\frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^n k\left(\frac{x_i - \overline{x}}{h} \right) k\left(\frac{x_j - \overline{x}}{h} \right) d_i d_j Z_i Z_j' \varepsilon_i \eta_j \right) D_{n+} e_1 \right]$$

$$= E\left[e_1' D_{n+}\left(\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i-\overline{x}}{h}\right)^2 d_i Z_i Z_i' E\left[\varepsilon_i \eta_i \mid X=x\right]\right) D_{n+}e_1\right]$$

$$= E\left[e_1' D_{n+}\left(\frac{1}{nh}\sum_{i=1}^n k\left(\frac{x_i-\overline{x}}{h}\right)^2 d_i Z_i Z_i' \sigma_{\varepsilon\eta}(x_i)\right) D_{n+}e_1\right]$$

$$= E\left[e_1' D_{n+}\left(\frac{1}{h} k\left(\frac{x_i-\overline{x}}{h}\right)^2 d_i Z_i Z_i' \sigma_{\varepsilon\eta}(x_i)\right) D_{n+}e_1\right]$$

$$= e_1' D_{n+}\left[\int_{\overline{x}}^{\overline{x}+h}\frac{1}{h} k\left(\frac{x-\overline{x}}{h}\right)^2 ZZ'\sigma_{\varepsilon\eta}(x)f_o(x)dx\right] D_{n+}e_1$$

where I used the assumption that $E\left[\varepsilon_i \eta_j \mid X=x\right]=0$ for $j \neq i$.

Similarly for the second term,

$$E\left[e_1' D_{n-}\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i)Z_i\varepsilon_i\right)\left(\frac{1}{\sqrt{nh}}\sum_{i=1}^n k\left(\frac{x_i-\overline{x}}{h}\right)(1-d_i)Z_i\eta_i\right)' D_{n-}e_1\right]$$

$$= e_1' D_{n-}\left[\int_{\overline{x}-h}^{\overline{x}}\frac{1}{h} k\left(\frac{x-\overline{x}}{h}\right)^2 ZZ'\sigma_{\varepsilon\eta}(x)f_o(x)dx\right] D_{n-}e_1$$

Let $\rho_j^+ = \int_{\overline{x}}^{\overline{x}+h}\frac{1}{h} k^2\left(\frac{x-\overline{x}}{h}\right)\left(\frac{x-\overline{x}}{h}\right)^j \sigma_{\varepsilon\eta}(x)f_o(x)dx = \int_0^\infty k^2(u)u^j\sigma_{\varepsilon\eta}(\overline{x}+uh)f_o(\overline{x}+uh)du$, Δ_+^ρ is the $(p+1)\times(p+1)$ matrix with (j,l) element ρ_{j+l-2}^+ for $j,l=1,...,p+1$, $\rho_j^- = \int_{\overline{x}-h}^{\overline{x}}\frac{1}{h} k^2\left(\frac{x-\overline{x}}{h}\right)\left(\frac{x-\overline{x}}{h}\right)^j \sigma_{\varepsilon\eta}(x)f_o(x)dx = (-1)^j\int_0^\infty k^2(u)u^j\sigma_{\varepsilon\eta}(\overline{x}-uh)f_o(\overline{x}-uh)du$ and Δ_-^ρ is the $(p+1)\times(p+1)$ matrix with (j,l) element ρ_{j+l-2}^- for $j,l=1,...,p+1$ Then the asymptotic covariance is given by

$$C_{\alpha\theta} = e_1'\left[\left(\Gamma_+^*\right)^{-1}\Delta_+^\rho\left(\Gamma_+^*\right)^{-1}+\left(\Gamma_-^*\right)^{-1}\Delta_-^\rho\left(\Gamma_-^*\right)^{-1}\right]e_1$$

■

Proof of Corollary 3. Using the results in equations 12 and noting that, if $h \to 0$

$$\rho_j^+ = \lim_{h\to 0}\int_0^\infty k^2(u)u^j\sigma_{\varepsilon\eta}(\overline{x}+uh)f_o(\overline{x}+uh)du$$

$$= \sigma_{\varepsilon\eta}^+(\overline{x})f_o(\overline{x})\delta_j$$

and similarly for ρ_j^-. Then,

$$\lim_{h\to 0}e_1'\left[\left(\Gamma_+^*\right)^{-1}\Delta_+^\rho\left(\Gamma_+^*\right)^{-1}+\left(\Gamma_-^*\right)^{-1}\Delta_-^\rho\left(\Gamma_-^*\right)^{-1}\right]e_1$$

$$= e_1'\left[(f_o(\overline{x})\Gamma)^{-1}\sigma_{\varepsilon\eta}^+(\overline{x})f_o(\overline{x})\Delta(f_o(\overline{x})\Gamma)^{-1}+(f_o(\overline{x})\Gamma)^{-1}\sigma_{\varepsilon\eta}^-(\overline{x})f_o(\overline{x})\Delta(f_o(\overline{x})\Gamma)^{-1}\right]e_1$$

$$= \frac{\sigma_{\varepsilon\eta}^+(\overline{x})+\sigma_{\varepsilon\eta}^-(\overline{x})}{f_o(\overline{x})}e_1'\Gamma^{-1}\Delta\Gamma^{-1}e_1$$

■

Proof of Corollary 4. The proof follows very closely corollary 3. Using the results in equations 12 and noting that, if $h>0$ and $f_o(x)=f_o(\overline{x})$ and $\sigma_{\varepsilon\eta}(x)=\sigma_{\varepsilon\eta}(\overline{x})$ for any x in the range around the cutoff

$$\rho_j^+ = \int_0^\infty k^2(u)u^j\sigma_{\varepsilon\eta}(\overline{x}+uh)f_o(\overline{x}+uh)du$$

$$= \sigma_{\varepsilon\eta}^+(\overline{x})f_o(\overline{x})\delta_j$$

and similarly for ρ_j^-. Then,

$$e_1' \left[\left(\Gamma_+^* \right)^{-1} \Delta_+^\rho \left(\Gamma_+^* \right)^{-1} + \left(\Gamma_-^* \right)^{-1} \Delta_-^\rho \left(\Gamma_-^* \right)^{-1} \right] e_1$$

$$= e_1' \left[\left(f_o(\overline{x}) \Gamma \right)^{-1} \sigma_{\varepsilon\eta}^+ (\overline{x}) f_o(\overline{x}) \Delta \left(f_o(\overline{x}) \Gamma \right)^{-1} + \left(f_o(\overline{x}) \Gamma \right)^{-1} \sigma_{\varepsilon\eta}^- (\overline{x}) f_o(\overline{x}) \Delta \left(f_o(\overline{x}) \Gamma \right)^{-1} \right] e_1$$

$$= \frac{\sigma_{\varepsilon\eta}^+ (\overline{x}) + \sigma_{\varepsilon\eta}^- (\overline{x})}{f_o(\overline{x})} e_1' \Gamma^{-1} \Delta \Gamma^{-1} e_1$$

■