Yale University
EliScholar – A Digital Platform for Scholarly Publishing at Yale

Cowles Foundation Discussion Papers | Cowles Foundation

12-1-2014

A Simple Adjustment for Bandwidth Snooping

Timothy B. Armstrong

Michal Kolesár

Follow this and additional works at: https://elischolar.library.yale.edu/cowles-discussion-paper-series

 Part of the Economics Commons

Recommended Citation

Armstrong, Timothy B. and Kolesár, Michal, "A Simple Adjustment for Bandwidth Snooping" (2014). *Cowles Foundation Discussion Papers*. 2372.
https://elischolar.library.yale.edu/cowles-discussion-paper-series/2372

This Discussion Paper is brought to you for free and open access by the Cowles Foundation at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Cowles Foundation Discussion Papers by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

A SIMPLE ADJUSTMENT FOR BANDWIDTH SNOOPING

By

Timothy B. Armstrong and Michal Kolesár

December 2014

COWLES FOUNDATION DISCUSSION PAPER NO. 1961



COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

A Simple Adjustment for Bandwidth Snooping*

Timothy B. Armstrong† Yale University

Michal Kolesár‡ Princeton University

November 30, 2014

Abstract

Kernel-based estimators are often evaluated at multiple bandwidths as a form of sensitivity analysis. However, if in the reported results, a researcher selects the bandwidth based on this analysis, the associated confidence intervals may not have correct coverage, even if the estimator is unbiased. This paper proposes a simple adjustment that gives correct coverage in such situations: replace the Normal quantile with a critical value that depends only on the kernel and ratio of the maximum and minimum bandwidths the researcher has entertained. We tabulate these critical values and quantify the loss in coverage for conventional confidence intervals. For a range of relevant cases, a conventional 95% confidence interval has coverage between 70% and 90%, and our adjustment amounts to replacing the conventional critical value 1.96 with a number between 2.2 and 2.8. A Monte Carlo study confirms that our approach gives accurate coverage in finite samples. We illustrate our approach with two empirical applications.

*We thank Joshua Angrist, Matias Cattaneo, Victor Chernozhukov, Kirill Evdokimov, Bo Honoré, Chris Sims, and numerous seminar participants for helpful comments and suggestions. We also thank Matias Cattaneo for sharing the Progresa dataset. All remaining errors are our own.

†email: timothy.armstrong@yale.edu

‡email: mkolesar@princeton.edu

1 Introduction

Kernel and local polynomial estimators of objects such as densities and conditional means involve a choice of bandwidth. To assess the sensitivity of the results to the choice of bandwidth, it is often recommended that researchers compute the estimates and confidence intervals for several bandwidths (Imbens and Lemieux, 2008), or plot them against a continuum of bandwidths (Lee and Lemieux, 2010; DiNardo and Lee, 2011). This recommendation is followed widely in applied work.[1]. However, such practice leads to a well-known problem that if the decision of which bandwidth to select is influenced by these results, the confidence interval at the selected bandwidth may undercover even if the estimator is unbiased.

This problem does not only arise when the selection rule is designed to make the results of the analysis look most favorable (for example by choosing a bandwidth that minimizes the p-value for some test). As we discuss in Section 4 below, undercoverage can also occur from honest attempts to report a confidence interval with good statistical properties. We use the term "bandwidth snooping" to refer to any situation where a researcher considers multiple values of the bandwidth in reporting confidence intervals.

This paper proposes a simple adjustment to confidence intervals that ensures correct coverage in these situations: replace a quantile of a Normal distribution with a critical value that depends only on the kernel, order of the local polynomial, and the ratio of the maximum and minimum bandwidths that the researcher has tried. We tabulate these adjusted critical values for a several popular choices of the kernel.

To explain the adjustment, consider a kernel estimator of a conditional mean based on an i.i.d. sample $\{(X_i, Y_i)\}_{i=1}^n$. Our approach applies more broadly to local polynomial estimators, and to other nonparametric quantities such as the regression discontinuity parameter (see the applications in Section 5), but we describe our approach in this context first for ease of exposition. We are interested in a conditional mean $E(Y_i \mid X_i = x)$ evaluated at a point x which we normalize

[1]For prominent examples of papers which report results for multiple, or a continuum of bandwidths in regression discontinuity designs, see, for instance, van Der Klaauw (2002), Lemieux and Milligan (2008), Ludwig and Miller (2007), or Card, Dobkin, and Maestas (2009)

to zero for notational convenience. The Nadaraya-Watson kernel estimator is given by

$$\hat{\theta}(h) = \frac{\sum_{i=1}^{n} Y_i k(X_i/h)}{\sum_{i=1}^{n} k(X_i/h)}$$

for some kernel function k. For a given h, $\hat{\theta}(h)$ is approximately unbiased for the pseudo-parameter

$$\theta(h) = \frac{EY_i k(X_i/h)}{Ek(X_i/h)}$$

and, if we take $h \to 0$ with the sample size, $\hat{\theta}(h)$ will converge to $\lim_{h\to 0} \theta(h) = E(Y_i \mid X_i = 0) =: \theta(0)$ under appropriate conditions on the smoothness of the conditional mean. Given an estimator $\hat{\sigma}(h)$ of the variance of $\sqrt{nh}(\hat{\theta}(h) - \theta(h))$, the t-statistic $\sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$ is approximately Normal with mean zero and variance one. Letting $z_{1-\alpha/2}$ denote the $1-\alpha/2$ quantile of the standard Normal distribution, the standard confidence interval $[\hat{\theta}(h) \pm z_{1-\alpha/2}\hat{\sigma}(h)/\sqrt{nh}]$, is therefore an approximate $100 \cdot (1-\alpha)\%$ confidence interval for $\theta(h)$. If $|\theta(h) - \theta(0)|$ is small enough relative to $\hat{\sigma}(h)/\sqrt{nh}$, the standard confidence interval is also approximate $100 \cdot (1-\alpha)\%$ confidence interval for $\theta(0) = E(Y_i|X_i = 0)$.

However, if a confidence interval is reported using some h^* that is chosen based on the results of examining $\hat{\theta}(h)$ over h in some interval $[\underline{h}, \overline{h}]$, the standard confidence interval around $\hat{\theta}(h^*)$, $[\hat{\theta}(h^*) \pm z_{1-\alpha/2}\hat{\sigma}(h^*)/\sqrt{nh^*}]$ may undercover even if $\theta(h^*) = \theta(0)$ (i.e., even if there is no bias). To address this problem, we propose confidence intervals that cover $\theta(h)$ simultaneously for all h in some specified interval $[\underline{h}, \overline{h}]$ with a prespecified probability. In particular, we derive a critical value $c_{1-\alpha}$ such that

$$P\left(\theta(h) \in [\hat{\theta}(h) \pm c_{1-\alpha}\hat{\sigma}(h)/\sqrt{nh}] \text{ for all } h \in [\underline{h}, \overline{h}]\right) \overset{n\to\infty}{\to} 1-\alpha. \tag{1}$$

In other words, our critical values allow for a uniform confidence band for $\theta(h)$. Thus, the confidence interval for the selected bandwidth, h^*, $[\hat{\theta}(h^*) \pm c_{1-\alpha}\hat{\sigma}(h^*)/\sqrt{nh^*}]$ will be robust to a bandwidth search over $[\underline{h}, \overline{h}]$ no matter what selection rule was used to pick h^*. Under additional conditions (such as undersmoothing or bias-correction), the selected confidence interval will have correct coverage for the parameter itself.

We state our results in terms of confidence intervals for $\theta(h)$, rather than $\theta(0)$ for three reasons. We discuss these here briefly and also refer the reader to Section 4 for more detailed descriptions of examples of situations where a confidence interval satisfying (1) is useful. We also note that, in certain settings, our methods can be used to obtain confidence intervals for $\theta(0)$ with certain optimality properties (see Example 4.2 in Section 4 below).

Our first reason for stating our results in terms of $\theta(h)$ is that it allows us to separate out the effect of multiple testing, which is the main focus of this paper, from the effect of bias on the coverage of the confidence interval. Methods for mitigating the bias through undersmoothing or bias correction, such as those proposed by Calonico, Cattaneo, and Titiunik (2014), can be incorporated into our procedure, and will lead to coverage of $\theta(0)$ under essentially the same conditions needed for pointwise coverage (i.e. if $\hat{\theta}(h)$ is "undersmoothed and/or bias corrected enough" that the pointwise CI has good pointwise coverage of $\theta(0)$ at each $h \in [\underline{h}, \overline{h}]$, our uniform CI will cover $\theta(0)$ uniformly over this set). We implement one of these approaches in our Monte Carlo study in Section 7.

We note, however, that any confidence interval promising coverage of $\theta(0)$ in the above setup (including those based on bias correction or undersmoothing) must require some form of smoothness or shape restrictions on the conditional mean (see Low, 1997, and the discussion at the end of Section 1.1 and Example 4.2 below). Since our confidence bands cover $\theta(h)$ under milder smoothness conditions than those needed for coverage of $\theta(0)$, they can be used to assess the sensitivity of an estimator to such assumptions. For example, if a particular method for bias correction or undersmoothing, applied to a particular data set, states that bias can be safely ignored for $h \leq 3$, and one finds that the confidence bands for, say, $h = 2$ and $h = 3$ do not overlap even after our correction, then one can conclude that the assumptions needed for this form of bias correction do not match the data.

This leads us to our second reason for stating our results in terms of $\theta(h)$: it allows the researcher to assess sensitivity to bandwidth choice, while taking into account the possibility that the chosen bandwidth h^* may be based on this sensitivity analysis. For sensitivity analysis, $\theta(h)$, rather than $\theta(0)$, is typically of interest. E.g., in the situation described immediately above, the researcher would make the following conclusion: "$\theta(2)$ and $\theta(3)$ must be very different relative to sampling error, since the confidence intervals (with the snooping adjustment) do not

overlap; since the bias correction method used in forming $\hat{\theta}(h)$ promised that $\theta(h)$ would be about the same (close to $\theta(0)$) for $h \leq 3$, the smoothness assumptions needed for this bias correction method to work must not do a good job of describing this data set." Our confidence bands can thus be used to formalize certain conclusions about confidence regions being "sensitive" to bandwidth choice that come out of the common practice of evaluating $\hat{\theta}(h)$ at multiple values of h.

Our third reason for stating the results in terms of $\theta(h)$ is that in many applications, $\theta(h)$, taken as a function indexed by the bandwidth, is a parameter of economic interest in its own right, in which case our confidence bands are simply confidence bands for this function. As we discuss in detail in Sections 4 and 5 below, this situation arises, for instance, in estimation of local average treatment effects for different sets of compliers, or in estimation of average treatment effects under unconfoundedness with limited overlap, where $\theta(h)$ corresponds to average treatment effects for different subpopulations that are indexed by h.

An advantage of our approach is that the critical value $c_{1-\alpha}$ depends only on the ratio $\overline{h}/\underline{h}$ and the kernel k (in the case of local polynomial estimators, $c_{1-\alpha}$ depends only on these objects, the order of the polynomial and whether the point is on the boundary of the support). In practice, researchers often report a point estimate $\hat{\theta}(h^*)$ and a standard error $\hat{\sigma}(h^*)/\sqrt{nh^*}$. As long as the kernel and order of the local polynomial are also reported, a reader can use the critical values tabulated in this paper to construct a confidence interval that takes into account a specification search over a range $[\underline{h}, \overline{h}]$ that the reader believes the original researcher used. The reader can then assess the sensitivity of the conclusions of the analysis to bandwidth specification search by, e.g., computing the largest value of $\overline{h}/\underline{h}$ for which the robust confidence interval does not include a particular value. As an example to give a sense of the magnitudes involved, we find that, with the uniform kernel and a local constant estimator, the critical value for a two sided uniform confidence band with $1 - \alpha = 0.95$ and $\overline{h}/\underline{h} = 3$ is about 2.6 (as opposed to 1.96 with no correction). If one instead uses the pointwise-in-h critical value of 1.96 and searches over $h \in [\underline{h}, \overline{h}]$ with $\overline{h}/\underline{h} = 3$, the true coverage (of $\theta(h)$) will be approximately 80%. The situation for the triangular kernel is more favorable, with a critical value of around 2.25 for the case with $\overline{h}/\underline{h} = 3$, and with the coverage of the pointwise-in-h procedure around 91%, although the situation for both gets worse as $\overline{h}/\underline{h}$ increases.

Our results also give analytic formulas for the critical values that are asymptotically valid in the case where $\overline{h}/\underline{h} \to \infty$. These results are based on extreme value limit theorems for $\sup_{\underline{h} \leq h \leq \overline{h}} \sqrt{nh}|\hat{\theta}(h) - \theta(h)|/\hat{\sigma}(h)$ that are valid in the case where $\overline{h}/\underline{h} \to \infty$ with the sample size, and may be of interest in their own right. Formally, these results show that $[\sup_{\underline{h} \leq h \leq \overline{h}} \sqrt{nh}|\hat{\theta}(h) - \theta(h)|/\hat{\sigma}(h)]/\sqrt{2 \log\log \overline{h}/\underline{h}}$ converges to a constant, and that a further scaling by $\sqrt{2 \log\log \overline{h}/\underline{h}}$ gives an extreme value limiting distribution. These results are related to a connection between our problem and the application of the law of the iterated logarithm to the sequential design of experiments, which we discuss briefly in Section 2. From a practical standpoint, these results suggest that one can examine a large range of bandwidths without paying too much of a penalty (since $\sqrt{2 \log\log(\overline{h}/\underline{h}))}$ increases very slowly relative to $\overline{h}/\underline{h}$). Indeed, when we examine how the critical values vary as a function of $\overline{h}/\underline{h}$ under the assumption that $\overline{h}/\underline{h}$ is fixed as $n \to \infty$, we find that, while using our correction is important for obtaining correct coverage, the critical values increase relatively slowly once $\overline{h}/\underline{h}$ is above 5.

We examine the finite sample properties of our procedure with a Monte Carlo study and find that uniform coverage of $\theta(h)$ is close to the nominal level, and that uniform coverage of $\theta(0)$ is good so long as the bias is small enough. We also illustrate our approach with two empirical applications. The first application is a regression discontinuity study based on Lee (2008). We find that, while the confidence regions are somewhat larger when one allows for examination of estimates at multiple bandwidths, the overall conclusions of that study are robust to a large amount of bandwidth snooping. The second application is a regression discontinuity setup of Calonico, Cattaneo, and Titiunik (2014). Here, in contrast, we find that the significance of the results is sensitive to bandwidth snooping.

The rest of the paper is organized as follows. Section 1.1 discusses related literature. Section 2 gives a nontechnical discussion of the derivation of our asymptotic distribution results in a simplified setup. Section 3 states our main asymptotic distribution result under general high-level conditions. Section 3.1 gives a step-by-step explanation of how to find the appropriate critical value in our tables and implement the procedure. Section 4 gives examples of situations where our approach of computing a uniform-in-h confidence band is useful. Section 5 presents some applications of our results and gives primitive conditions for the validity of our critical values in these applications. Section 6 presents an illustration of our approach in two empirical

applications. Section 7 presents the results of a Monte Carlo study. Section 8 concludes. Proofs and auxiliary results, as well as additional tables and figures, are given in the appendix and a supplemental appendix. Since Section 2 and the beginning of Section 3 are concerned primarily with theoretical aspects of our problem, readers who are primarily interested in implementation can skip Section 2 and the beginning of Section 3 up to Section 3.1.

1.1 Related literature

The idea of controlling for multiple inference by constructing a uniform confidence band has a long tradition in the statistics literature, and the number of papers treating this topic is too large to cover all of them here. Chapter 9 of Lehmann and Romano (2005) gives an overview of this problem and early contributions. White (2000), Romano and Wolf (2005) and Miller and Siegmund (1982) are examples of papers that use these ideas with a similar spirit to our application, but for different problems. The term "snooping," which we take from this literature, goes back even further (see, e.g., Selvin and Stuart, 1966). To our knowledge, the application to kernel estimators and the critical values derived in this paper are in general new, although certain cases involving the uniform kernel, constant conditional means and homoskedastic errors reduce to mild extensions of well known results (see the discussion below).

On a technical level, our results borrow from the literature on Gaussian approximation to empirical processes and extreme value approximations to Gaussian processes. We use an approximation of Sakhanenko (1985) (see also Shao, 1995) and a derivation that is similar in spirit to Bickel and Rosenblatt (1973) to obtain an approximation of the kernel estimator by a Gaussian process. We obtain extreme value limits for suprema of these processes using classical results (see Leadbetter, Lindgren, and Rootzen, 1983). In the general case, these extreme value limiting results appear to be new. In the case of conditional mean estimation with homoskedastic errors, a constant conditional mean, and a uniformly distributed covariate, this step of the derivation reduces to a theorem of Darling and Erdos (1956) (see also Einmahl and Mason, 1989), so our results can be considered an extension of this theorem. While our goal is to obtain critical values with a simple form using the structure of our problem, a general bootstrap approach to obtaining uniform confidence bands without obtaining an asymptotic distribution has been used recently by Chernozhukov, Chetverikov, and Kato (2013), and these results could be useful in extensions

of our results to other settings. Of course, our results also borrow from the broader literature on nonparametric kernel and local polynomial estimation. This literature is too large to give a full treatment here, but Fan and Gijbels (1996) and Pagan and Ullah (1999) are both useful textbook references, particularly for an econometric perspective.

Our interest in nonparametric estimators for a function at a point stems from several applications in empirical economics, which we treat in Section 5. The regression discontinuity setting uses nonparametric estimates of a conditional mean at a point, and has become increasingly popular in empirical work (see, among many others, Hahn, Todd, and Van der Klaauw, 2001; Sun, 2005; Imbens and Lemieux, 2008; Imbens and Kalyanaraman, 2012; Calonico, Cattaneo, and Titiunik, 2014). We treat this application in Section 5.1. Inference on a conditional mean at the boundary of the support of a covariate is relevant in econometric models that are "identified at infinity," (see, among others Chamberlain, 1986; Heckman, 1990; Andrews and Schafgans, 1998). Certain settings with heterogeneous treatment effects and instrumental variables, considered by (among others) Heckman and Vytlacil (2005), Heckman, Urzua, and Vytlacil (2006) and Imbens and Angrist (1994), are closely related to these ideas, and are considered in Section 5.2. The issue of "identification at infinity" also arises in the use of trimmed estimators for inference on average treatment effects under unconfoundedness (see, among others, Crump, Hotz, Imbens, and Mitnik 2009 and Khan and Tamer 2010). We cover this application in Section 5.3.

An important area of application of multiple tests involving tuning parameters is adaptive inference and testing (in our context, this amounts to constructing a confidence band for $\theta(0)$ that is close to as small as possible for a range of smoothness classes for the data generating process). While we do not consider this problem in this paper, Armstrong (2014) uses our approach to obtain adaptive one-sided confidence intervals under a monotonicity condition (see Example 4.2 in Section 4 below). For the problem of global estimation and uniform confidence bands Giné and Nickl (2010) propose an approach based on a different type of shape restriction. The latter approach has been generalized in important work by Chernozhukov, Chetverikov, and Kato (2014). The shape restrictions involved in these papers cannot be done away with, as shown by Low (1997). For the problem of adaptive testing, Spokoiny (1996), Fan (1996) and Horowitz and Spokoiny (2001), among others, have used multiple tests involving tuning parameters in other settings. See Armstrong (2014) for additional references.

2 Derivation of the correction in a simple case

This section presents an intuitive derivation of the correction in the case of the conditional mean described in the introduction. To further simplify the exposition, let us first consider an idealized situation in which $Y_i = g(X_i) + \sigma\varepsilon_i$, σ^2 is known, ε_i are i.i.d. with variance one, and the regressors are non-random and equispaced on $[-1/2, 1/2]$. For reasons that will become clear below, it will be easiest if we define $X_i = (i+1)/(2n)$ for i odd and $X_i = -i/(2n)$ for i even (technically, this leads to the regressors not being equally spaced at zero, but this will not matter as $n \to \infty$). Consider the Nadaraya-Watson kernel estimator with a uniform kernel, $k(x) = I(|x| \leq 1/2)$

$$\hat{\theta}(h) = \frac{\sum_{i=1}^{n} k(X_i/h) Y_i}{\sum_{i=1}^{n} k(X_i/h)} = \frac{\sum_{i=1}^{nh} Y_i}{nh}$$

where, for the second equality and throughout the rest of this example, we assume that nh is an even integer for notational convenience. We would like to construct a confidence interval for

$$\theta(h) = E(\hat{\theta}(h)) = \frac{\sum_{i=1}^{nh} g(X_i)}{nh}$$

that will have coverage $1-\alpha$ no matter what bandwidth h we pick, so long as h is in some given range $[\underline{h}, \overline{h}]$. If the conditional mean function $g(x)$ is smooth near 0 and the range of bandwidths is small, so that the difference $\theta(h) - \theta(0)$ is small relative to the variance of $\hat{\theta}(h)$, the confidence interval can be interpreted as a confidence interval for the conditional mean at 0, $g(0)$. For a given bandwidth h, a two-sided t-statistic is given by

$$\sqrt{nh}\frac{|\hat{\theta}(h) - \theta(h)|}{\sigma} = \left| \frac{\sum_{i=1}^{nh} \varepsilon_i}{\sqrt{nh}} \right|. \tag{2}$$

In order to guarantee correct coverage, instead of using critical value that corresponds to the $1-\alpha/2$ quantile of a Normal distribution, we will need to use a critical value that corresponds to the $1-\alpha$ quantile of the distribution of the maximal t-statistic in the range $[\underline{h}, \overline{h}]$. If $n\underline{h} \to \infty$, we can approximate the partial sum $n^{-1/2}\sum_{i=1}^{nh} \varepsilon_i$ by a Brownian motion $\mathbb{B}(h)$, so that in large

samples, we can approximate the distribution of the maximal t-statistic as

$$\sup_{\underline{h}\leq h\leq \overline{h}} \sqrt{nh}\frac{|\hat{\theta}(h)-\theta(h)|}{\sigma} \approx \sup_{\underline{h}\leq h\leq \overline{h}} \left|\mathbb{B}(h)/\sqrt{h}\right| \stackrel{d}{=} \sup_{1\leq h\leq \overline{h}/\underline{h}} \left|\mathbb{B}(h)/\sqrt{h}\right|. \tag{3}$$

Thus, the sampling distribution of the maximal t-statistic will in large samples only depend on the ratio of maximum and minimum bandwidth that we consider, $\overline{h}/\underline{h}$, and can its quantiles can easily be tabulated (see the columns corresponding to uniform kernel in Table 1).

The representation above also explains why $\sqrt{\log\log(\overline{h}/\underline{h})}$ terms pop up in the rates at which the critical values increase if $\overline{h}/\underline{h}\to\infty$. In particular, as $\overline{h}/\underline{h}\to\infty$, the recentered distribution of $\sup_{1\leq h\leq \overline{h}/\underline{h}}|\mathbb{B}(h)/\sqrt{h}|$, scaled by $\sqrt{2\log\log(\overline{h}/\underline{h})}$, can be approximated by the extreme value distribution by the Darling and Erdos (1956) theorem. Moreover, because nh corresponds to an effective sample size, it follows from (2) that looking at kernel estimators with multiple bandwidths is essentially the same problem as computing t-statistics based on multiple sample sizes. Therefore, in this simple example, the critical value adjustment corresponds to the critical value adjustment in the sequential design of experiments, in which one adds more subjects to the experiment and recomputes t-statistics until one finds statistically significant results (see Siegmund, 1985, for an overview of this literature). The law of the iterated logarithm gives the rate at which these critical values must increase in order to control the size of the overall sequential test.

In order to convey the intuition behind our results, the setup in this section has been overly simplistic. In the next section, we show that the approximation of the distribution of the maximal t-statistic by a scaled Brownian motion in (3) still obtains even if these restrictive assumptions are dropped, and holds for more general problems than inference for the conditional mean at a point. The only difference will be that if the kernel is not uniform, then we need to approximate the distribution of the maximal t-statistic by a different Gaussian process.

3 General setup and main result

This section describes our general setup, states our main asymptotic distribution result, and derives critical values based on this result. Readers who are interested only in implementing our procedure can skip to Section 3.1, which explains how to use our tables to find critical values and implement our procedure. We state our result using high level conditions, which we verify

for some applications in Section 5.

We use the following definitions and notation throughout the rest of the paper. For a random vector (X_i, D_i, Y_i) with X_i continuously distributed, $E(Y_i|D_i = d, X_i = \tilde{x}_+) = \lim_{x \downarrow \tilde{x}} E(Y_i|D_i = d, X_i = x)$ and $E(Y_i|D_i = d, X_i = \tilde{x}_-) = \lim_{x \uparrow \tilde{x}} E(Y_i|D_i = d, X_i = x)$. We say that a function f is continuous at $\tilde{x}$ with local modulus of continuity $\ell(x)$ if $\|f(x) - f(\tilde{x})\| \leq \ell(\|x - \tilde{x}\|)$ for $\|x - \tilde{x}\|$ small enough. If this holds for $x > 0$ ($x < 0$) we say that f is right- (left-) continuous at $\tilde{x}$ with local modulus of continuity $\ell(x)$. We use the notation $\#\mathcal{A}$ to denote the number of elements in a set $\mathcal{A}$.

We consider a sample $\{X_i, W_i\}_{i=1}^n$, which we assume throughout the paper to be i.i.d. Here, X_i is a real-valued random variable, and we are interested in a kernel estimate at a particular point, which we normalize to be $x = 0$ for notational convenience. We consider confidence intervals that are uniform in h over some range $[\underline{h}_n, \overline{h}_n]$, where we now make explicit the dependence of $\underline{h}_n$ and $\overline{h}_n$ on n. Our main condition imposes an influence function representation involving a kernel function.

Assumption 3.1. *For some function $\psi(W_i, h)$ and a kernel function k with $E\psi(W_i, h)k(X_i/h) = 0$ and $\frac{1}{h} var(\psi(W_i, h)k(X_i/h)) = 1$,*

$$\frac{\sqrt{nh}(\hat{\theta}(h) - \theta(h))}{\hat{\sigma}(h)} = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \psi(W_i, h)k(X_i/h) + o_P\left(1/\sqrt{\log\log(\overline{h}_n/\underline{h}_n)}\right)$$

uniformly over $h \in [\underline{h}_n, \overline{h}_n]$, where $w \mapsto \psi(w, h)$ and $x \mapsto k(x/h)$ have polynomial uniform covering numbers (as defined in Section A of the appendix) taken as functions over w and x.

Most of the verification of Assumption 3.1 is standard. For most kernel and local polynomial based estimators, these calculations are available in the literature, with the only additional step being that the remainder term must be bounded uniformly over $\underline{h}_n \leq h \leq \overline{h}_n$, and with a $o_P(1/\sqrt{\log\log \overline{h}_n/\underline{h}_n})$ rate of approximation. Section S1.2 in the supplemental appendix provides some results that can be used to obtain this uniform bound. In the local polynomial case, the kernel function k is different from the original kernel, and depends on the order of the polynomial and whether the estimated conditional quantities are at the boundary (see Fan and Gijbels, 1996, and supplemental appendix S2).

We also impose some regularity conditions on k and the data generating process. In applica-

tions, these will typically impose smoothness conditions on the conditional mean and variance of certain variables conditional on X_i.

Assumption 3.2. *(i) The kernel function k is symmetric with finite support $[-A, A]$, is bounded with a bounded, uniformly continuous first derivative on $(0, A)$, and satisfies $\int k(u)\,du \neq 0$.*

(i) $|X_i|$ has a density $f_{|X|}$ with $f_{|X|}(0) > 0$, $\psi(W_i, h)k(X_i/h)$ is bounded uniformly over $h \leq \overline{h}_n$ with $var\left[\psi(W_i, 0)||X_i| = 0\right] > 0$, and, for some deterministic function $\ell(h)$ with $\ell(h)\log\log h^{-1} \to 0$ as $h \to 0$, the following expressions are bounded by $\ell(t)$: $|E\left[\psi(W_i, 0) \mid |X_i| = t\right] - E\left[\psi(W_i, 0) \mid |X_i| = 0\right]|$, $|var\left[\psi(W_i, 0)||X_i| = t\right] - var\left[\psi(W_i, 0)||X_i| = 0\right] |$ and $|(\psi(W_i, t) - \psi(W_i, 0))k(X_i/t)|$.

Note that, while Assumption 3.2 will typically require some smoothness on $\theta(h)$ as a function of h (since it places smoothness on certain conditional means, etc.), the amount of smoothness required is very mild relative to smoothness conditions typically imposed when considering bias-variance tradeoffs. In particular, our conditions require only that certain quantities are slightly smoother than $t \mapsto 1/\log\log t^{-1}$, which does not require differentiability and holds, e.g., for $t \mapsto t^\gamma$ for any $\gamma > 0$. Thus, our confidence bands for $\theta(h)$ are valid under very mild conditions on the smoothness of $\theta(h)$, and our results are valid in settings where the possible lack of smoothness of $\theta(h)$ leads one to examine $\hat{\theta}(h)$ across multiple bandwidths.

We also note that Assumption 3.1 and 3.2 are tailored toward statistics involving conditional means, rather than densities or derivatives of conditional means and densities (for density estimation, we would have $\psi(W_i, h) = 1$, which is ruled out by the assumptions $var\left[\psi(W_i, 0)||X_i| = 0\right] > 0$ and $E\psi(W_i, h)k(X_i/h) = 0$; for estimating derivatives of conditional means or densities, the scaling would be $\sqrt{nh^{1+\nu}}$ where ν is the order of the derivative). This is done only for concreteness and ease of notation, and the results can be generalized to these cases as well. Theorems A.1 and A.3 in Appendix A, which are used in proving Theorem 3.1 below, use high level conditions, which can be verified to give the result in other cases. The only requirement is that a scaled version of $\hat{\theta}(h) - \theta(h)$ be approximated by the Gaussian process $\mathbb{H}$ given in Theorem 3.1 below. For estimating derivatives, the kernel k in the process $\mathbb{H}$ will depend on the order of the derivative as well as the order of the local polynomial.

We are now ready to state the main asymptotic approximation result.

Theorem 3.1. *Let $c_{1-\alpha}(t, k)$ be the $1-\alpha$ quantile of $\sup_{1\leq h\leq t} \mathbb{H}(h)$, and let $c_{1-\alpha,|\cdot|}(t, k)$ be the $1-$*

α quantile of $\sup_{1\leq h\leq t}|\mathbb{H}(h)|$*, where* $\mathbb{H}(h)$ *is a mean zero Gaussian process with covariance kernel* $cov\left(\mathbb{H}(h),\mathbb{H}(h')\right)=\frac{\int k(u/h)k(u/h')\,du}{\sqrt{hh'}\int k(u)^2\,du}=\sqrt{\frac{h'}{h}}\frac{\int k(u(h'/h))k(u)\,du}{\int k(u)^2\,du}$*. Suppose that* $\underline{h}_n\to 0$, $\overline{h}_n=\mathcal{O}_P(1)$*, and* $n\underline{h}_n/[(\log\log n)(\log\log\log n)]^2\to\infty$*. Then, under Assumptions 3.1 and 3.2,*

$$P\left(\frac{\sqrt{nh}\left(\hat{\theta}(h)-\theta(h)\right)}{\hat{\sigma}(h)}\leq c_{1-\alpha}\left(\overline{h}_n/\underline{h}_n,k\right)\ \textit{all}\ h\in[\underline{h}_n\leq h\leq\overline{h}_n]\right)\overset{n\to\infty}{\to}1-\alpha$$

and

$$P\left(\frac{\sqrt{nh}\left|\hat{\theta}(h)-\theta(h)\right|}{\hat{\sigma}(h)}\leq c_{1-\alpha,|\cdot|}\left(\overline{h}_n/\underline{h}_n,k\right)\ \textit{all}\ h\in[\underline{h}_n\leq h\leq\overline{h}_n]\right)\overset{n\to\infty}{\to}1-\alpha.$$

If, in addition, $\overline{h}_n/\underline{h}_n\to\infty$*, the above statements also hold with* $c_{1-\alpha,|\cdot|}(\overline{h}_n/\underline{h}_n,k)$ *replaced by*

$$\frac{-\log\left(-\frac{1}{2}\log(1-\alpha)\right)+b(\overline{h}_n/\underline{h}_n,k)}{\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}}+\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)},$$

and $c_{1-\alpha}(\overline{h}_n/\underline{h}_n,k)$ *replaced by*

$$\frac{-\log\left(-\log(1-\alpha)\right)+b(\overline{h}_n/\underline{h}_n,k)}{\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}}+\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)},$$

where $b(t,k)=\log c_1(k)+(1/2)\log\log\log t$ *if* $k(A)\neq 0$ *and* $b(t,k)=\log c_2(k)$ *if* $k(A)=0$*, with* $c_1(k)=\frac{Ak(A)^2}{\sqrt{\pi}\int k(u)^2\,du}$ *and* $c_2(k)=\frac{1}{2\pi}\sqrt{\frac{\int\left[k'(u)u+\frac{1}{2}k(u)\right]^2du}{\int k(u)^2\,du}}$*.*

Theorem 3.1 shows that the quantiles of $\sup_{\underline{h}_n\leq h\leq\overline{h}_n}\sqrt{nh}(\hat{\theta}(h)-\theta(h))/\hat{\sigma}(h)$ can be approximated by simulating from the supremum of a certain Gaussian process. In addition, Theorem 3.1 provides a further approximation to these critical values based on an extreme value limiting distribution in the case where $\overline{h}_n/\underline{h}_n\to\infty$. In the case where k is the uniform kernel, $\phi(W_i,h)$ does not depend on h and $E[\phi(W_i,h)|X_i=x]=0$ and $var[\phi(W_i,h)|X_i=x]=1$ for all x, the latter result reduces to a well-known theorem of Darling and Erdos (1956) (see also Einmahl and Mason, 1989). For the case where k is not the uniform kernel, or where ψ depends on h, this result is, to our knowledge, new. While the approximations based on $\overline{h}_n/\underline{h}_n\to\infty$ are useful in giving an analytic approximation to how the critical values $c_{1-\alpha,|\cdot|}(\overline{h}_n/\underline{h}_n,k)$ and $c_{1-\alpha}(\overline{h}_n/\underline{h}_n,k)$ change with $\overline{h}_n/\underline{h}_n$, we recommend using $c_{1-\alpha,|\cdot|}(\overline{h}_n/\underline{h}_n,k)$ and $c_{1-\alpha}(\overline{h}_n/\underline{h}_n,k)$ directly, since, given their

asymptotic validity when $\overline{h}_n/\underline{h}_n$ is bounded, we expect these critical values to perform better when $\overline{h}_n/\underline{h}_n$ is not too large.

To outline how Theorem 3.1 obtains, let us again consider the problem of estimating a nonparametric mean at a point described in the introduction. Here, we set $\psi(W_i, h) = (Y_i - \theta(h))/\sqrt{var\left\{[Y_i - \theta(h)]k(X_i/h)/h\right\}}$ so that, for small h, we can approximate the t-statistic as

$$\frac{\sqrt{nh}(\hat{\theta}(h) - \theta(h))}{\hat{\sigma}(h)} \approx \frac{\sum_{i=1}^n [Y_i - \theta(h)]k(X_i/h)}{\sqrt{n \cdot var\left\{[Y_i - \theta(h)]k(X_i/h)\right\}}}.$$

Thus, we expect that the supremum of the absolute value of this display over $h \in [\underline{h}, \overline{h}]$ is approximated by $\sup_{h\in[\underline{h},\overline{h}]} |\mathbb{H}_n(h)|$ where $\mathbb{H}_n(h)$ is a Gaussian process with covariance function

$$cov\left(\mathbb{H}_n(h), \mathbb{H}_n(h')\right) = \frac{cov\left\{[Y_i - \theta(h)]k(X_i/h), [Y_i - \theta(h')]k(X_i/h')\right\}}{\sqrt{var\left\{[Y_i - \theta(h)]k(X_i/h)\right\}}\sqrt{var\left\{[Y_i - \theta(h')]k(X_i/h')\right\}}}. \tag{4}$$

The conditions in Assumption 3.2 ensure that $E(Y_i|X_i = x)$, $var(Y_i|X_i = x)$ and the density $f_X(x)$ of X_i do not vary too much as $x \to 0$, so that, for h and h' close to zero

$$\begin{aligned}
&cov\left\{[Y_i - \theta(h)]k(X_i/h), [Y_i - \theta(h')]k(X_i/h')\right\} \approx E\left\{[Y_i - E(Y_i|X_i)]^2 k(X_i/h)k(X_i/h')\right\} \\
&= \int var(Y_i|X_i = x)k(x/h)k(x/h')f_X(x)\,dx \approx var(Y_i|X_i = 0)f_X(0)\int k(x/h)k(x/h')\,dx \\
&= var(Y_i|X_i = 0)f_X(0)h'\int k\left(u(h'/h)\right)k(u)\,du.
\end{aligned}$$

Using this approximation for the variance terms in the denominator of (4) as well as the covariance in the numerator gives the approximation

$$cov\left(\mathbb{H}_n(h), \mathbb{H}_n(h')\right) \approx \frac{h'\int k\left(u(h'/h)\right)k(u)\,dx}{\sqrt{h'\int k(u)^2\,dx}\sqrt{h\int k(u)^2\,dx}} = \frac{\sqrt{h'/h}\int k\left(u(h'/h)\right)k(u)\,dx}{\int k(u)^2\,dx}.$$

Thus, letting $\mathbb{H}(h)$ be the Gaussian process with the covariance on the right hand side of the above display, we expect that the distribution of $\sup_{h\in[\underline{h},\overline{h}]} \frac{\sqrt{nh}(\hat{\theta}(h)-\theta(h))}{\hat{\sigma}(h)}$ is approximated by the distribution of $\sup_{h\in[\underline{h},\overline{h}]} |\mathbb{H}(h)|$. Since the covariance kernel given above depends only on h'/h, $\sup_{h\in[\underline{h},\overline{h}]} |\mathbb{H}(h)|$ has the same distribution as $\sup_{h\in[\underline{h},\overline{h}]} |\mathbb{H}(h/\underline{h})| = \sup_{h\in[1,\overline{h}/\underline{h}]} |\mathbb{H}(h)|$. As it turns out, this approximation will work under relatively mild conditions so long as $\underline{h} \to 0$ even if $\overline{h}$ does not approach zero, because, in this case, the maximally selected bandwidth will still converge in

probability to zero, yielding the first part of the theorem. For the second part of the theorem, we shown that $\sup_{h\in[\underline{h},\overline{h}]} \frac{\sqrt{nh}|\hat{\theta}(h)-\theta(h)|}{\hat{\sigma}(h)}$ increases proportionally to $\sqrt{2\log\log(\overline{h}/\underline{h})}$, and that a further scaling by $\sqrt{2\log\log(\overline{h}/\underline{h})}$ gives an extreme value limiting distribution. As discussed above, this is related to the classical law of the iterated logarithm and its relation to the sequential design of experiments. To further understand the intuition for this, note that $\mathbb{H}(h)$ is stationary when indexed by $t = \log h$ (since the covariance at $h = e^t$ and $h' = e^{t'}$ depends only on $h'/h = e^{t'-t}$), so, setting $T = \log(\overline{h}/\underline{h})$, we expect the supremum over $[\log 1, \log(\overline{h}/\underline{h})] = [0, T]$ to follow an extreme value limiting with scaling $\sqrt{2\log T} = \sqrt{2\log\log(\overline{h}/\underline{h})}$ so long as dependence dies away quickly enough with T, following classical results (see Leadbetter, Lindgren, and Rootzen, 1983, for a textbook exposition of these results).

3.1 Practical implementation

For convenience, this section gives step-by-step instructions for finding the appropriate critical value in our tables and implementing our procedure. We also provide some analysis of the magnitudes involved in the correction and the undercoverage that can occur from searching over multiple bandwidths without implementing our correction.

Table 1 gives one- and two-sided critical values $c_{1-\alpha}\left(\overline{h}_n/\underline{h}_n, k\right)$ and $c_{1-\alpha,|\cdot|}\left(\overline{h}_n/\underline{h}_n, k\right)$ for several kernel functions k, $\alpha = 0.05$ and selected values of $\overline{h}_n/\underline{h}_n$. Critical values for $\alpha = 0.01$ and $\alpha = 0.10$, and additional kernels are given in Tables S1–S6 in the supplemental appendix. Figure S2 in the supplemental appendix compares these critical values to those based on the extreme value limiting distribution as $\overline{h}/\underline{h} \to \infty$. The critical values can also be obtained using our R package `bandwidth-snooping`, which can be downloaded from `https://github.com/kolesarm/bandwidth-snooping`.

Using these tables, our procedure can be described in the following steps:

1. Compute an estimate $\hat{\sigma}(h)$ of the standard deviation of $\sqrt{nh}(\hat{\theta}(h) - \theta(h))$, where $\hat{\theta}(h)$ is a kernel-based estimate.

2. Let $\underline{h}$ and $\overline{h}$ be the smallest and largest values of the bandwidth h considered, respectively, and let α be the nominal level. Look up the critical value $c_{1-\alpha,|\cdot|}\left(\overline{h}/\underline{h}, k\right)$ (or $c_{1-\alpha}\left(\overline{h}/\underline{h}, k\right)$ for the one-sided case) in Table 1 for $\alpha = 0.05$, or in Tables S1–S6 for $\alpha = 0.01$ and $\alpha = 0.10$.

3. Report $\left\{\hat{\theta}(h) \pm (\hat{\sigma}(h)/\sqrt{nh})c_{1-\alpha,|\cdot|}\left(\overline{h}/\underline{h}, k\right) \;\middle|\; h \in [\underline{h}, \overline{h}]\right\}$ as a uniform confidence band for $\theta(h)$. Or, report $\hat{\theta}(h^*) \pm (\hat{\sigma}(h^*)/\sqrt{nh^*})c_{1-\alpha,|\cdot|}\left(\overline{h}/\underline{h}, k\right)$ for a chosen bandwidth h^* as a confidence interval for $\theta(h^*)$ that takes into account "snooping" over $h \in [\underline{h}, \overline{h}]$.

It is common practice to report an estimate $\hat{\theta}(h^*)$ and a standard error $se(h^*) \equiv \hat{\sigma}(h^*)/\sqrt{nh^*}$ for a value of h^* chosen by the researcher. If one suspects that results reported in this way were obtained after examining the results for h in some set $[\underline{h}, \overline{h}]$ (say, by looking for the value of h for which the corresponding test of $H_0 : \theta(h) = 0$ has the smallest p-value), one can compute a "bandwidth snooping adjusted" confidence interval as described in step 3, so long as the kernel function is reported (as well as the order of the local polynomial).

Figure 1 plots our critical values as a function of $\overline{h}/\underline{h}$ for the two-sided case with $1 - \alpha = .95$. By construction, the critical value is given by the standard Normal quantile 1.96 when $\overline{h}/\underline{h} = 1$, and increases from there. For the kernels and range of $\overline{h}/\underline{h}$ considered, the correction typically amounts to replacing the standard Normal quantile 1.96 with a number between 2.2 and 2.8, depending on the kernel and range of bandwidths considered.

Our results can also be used to quantify undercoverage from entertaining multiple bandwidths without using our correction. Figure 2 plots the true uniform asymptotic coverage of a nominal 95% confidence interval over a range $[\underline{h}, \overline{h}]$ for different values of $\overline{h}/\underline{h}$. This amounts to finding $1 - \tilde{\alpha}$ such that the pointwise critical value 1.96 is equal to $c_{1-\tilde{\alpha},|\cdot|}(\overline{h}/\underline{h}, k)$. For the range of values of $\overline{h}/\underline{h}$ that we consider ($\overline{h}/\underline{h}$ below 10), the true coverage is typically somewhere between 70% and 90%, depending on the kernel and the value of $\overline{h}/\underline{h}$.

4 Examples of bandwidth snooping

This section provides some examples where computing a uniform confidence band for $\theta(h)$ is relevant. In some cases, the justification for using our approach involves the practical realities of empirical work while in others, our approach provides an optimal solution to a well-defined statistical problem. For concreteness, the first two examples in this section use the setup above where $\hat{\theta}(h)$ is a kernel estimator of a conditional mean, but the points made here apply more generally.

Example 4.1. A researcher would like to construct a confidence interval for the conditional mean $\theta(0) = E(Y_i|X_i = 0)$. Automatic methods for bandwidth choice trading off bias and variance lead to a choice of bandwidth $\hat{h}_{\text{opt}}$ such that the asymptotic distribution of $\hat{\theta}(h_{\text{opt}})$ is biased. The researcher therefore evaluates the estimator and CI at a smaller bandwidth h_{small}, such that the bias is negligible under appropriate assumptions on the smoothness of $E(Y_i|X_i = x)$ (this practice is often referred to as "undersmoothing" in the literature). Uncomfortable with these assumptions, the researcher then evaluates the estimator at an even smaller bandwidth h_{smaller}, leading to a confidence region based on $\hat{\theta}(h_{\text{smaller}})$ that is valid under weaker conditions on the smoothness of the conditional mean.

Suppose that the researcher is interested in whether $E(Y_i|X_i = 0) = 0$, and that the CI evaluated at h_{small} contains zero, while the CI evaluated at h_{smaller} does not. Since the confidence interval based on $\hat{\theta}(h_{\text{smaller}})$ is robust under weaker assumptions, the researcher may be tempted to conclude that $\theta(0) = E(Y_i|X_i = 0)$ is nonzero, and that the conclusions of this hypothesis test are robust under even weaker assumptions than the original assumptions the researcher had in mind. Of course, this is not true for the actual hypothesis test that the researcher has performed (looking at both $\hat{\theta}(h_{\text{small}})$ and $\hat{\theta}(h_{\text{smaller}})$), since the α probability of type I error has already been "used up" on the test based on $\hat{\theta}(h_{\text{small}})$. By replacing $z_{1-\alpha/2}$ with the critical value $c_{1-\alpha,|\cdot|}$ derived above for the kernel k and any $\overline{h}/\underline{h}$ with $h_{\text{small}}, h_{\text{smaller}} \in [\underline{h}, \overline{h}]$, the researcher can conclude that $\theta(0) \neq 0$ under the original assumptions that led to bias being negligible under h_{small}, so long as at least one of the two confidence intervals does not contain zero. Appendix B provides some further discussion of cases where the uniform-in-h confidence bands provided in this paper can be useful in sensitivity analysis.

Example 4.2. Suppose that X_i has support $[0, \overline{x}]$, and that $E(Y_i|X_i = x)$ is known to be weakly decreasing, and a Nadaraya-Watson estimator is used with a positive kernel. Then $\theta(h) \leq \theta(0) = E(Y_i|X_i = 0)$ for any h, so the one sided confidence interval $[\hat{\theta}(h) - z_{1-\alpha}\hat{\sigma}(h)/\sqrt{nh}, \infty)$ is asymptotically valid for any h regardless of how fast $h \to 0$ with n (or even if h does not decrease with n at all). One may wish to use this fact to justify reporting the most favorable confidence interval, namely, $[\sup_{h\in[\underline{h},\overline{h}]}(\hat{\theta}(h) - z_{1-\alpha}\hat{\sigma}(h)/\sqrt{nh}), \infty)$ for some $[\underline{h}, \overline{h}]$. Of course, this will not be a valid confidence interval because of the issues with entertaining multiple bandwidths described above. However, using the one-sided version of our critical value, $c_{1-\alpha}$, one can construct the confidence

interval $[\sup_{h\in[\underline{h},\bar{h}]} \hat{\theta}(h) - c_{1-\alpha}\hat{\sigma}(h)/\sqrt{nh}, \infty)$, which will have correct asymptotic coverage.

In fact, this confidence region enjoys an optimality property of being adaptive to certain levels of smoothness of the conditional mean, so long as $\bar{h} \to 0$ slowly enough and $\underline{h} \to 0$ quickly enough. For any $\beta \in (0, 1]$, if $E(Y_i|X_i = x)$ approaches $E(Y_i|X_i = 0)$ at the rate x^β, the lower endpoint of this confidence interval will shrink toward $\theta(0) = E(Y_i|X_i = 0)$ at the same rate as a confidence interval constructed using prior knowledge of β in an optimal way, up to a term involving $\log\log n$. Furthermore, no confidence region can achieve this rate simultaneously for β in a nontrivial interval without giving up this $\log\log n$ term. Since the $\log\log n$ term comes from the multiple bandwidth adjustment in our critical values, this shows that such an adjustment (or something like it), is needed for this form of adaptation. In particular, one cannot estimate the optimal bandwidth accurately enough to do away with our correction (see Armstrong, 2014, for details).

Example 4.3. In many examples in applied econometrics, $\theta(h)$ is an interesting object in its own right. In several problems involving estimation of treatment effects, $\theta(h)$ corresponds to a weighted average treatment effect, where the weights that different individuals receive are determined by h. An application of our procedure yields a uniform confidence band for a set of weighted average treatment effects. This situation arises in estimating treatment effects for the largest set of compliers (Heckman and Vytlacil, 2005; Heckman, Urzua, and Vytlacil, 2006). We apply our results to the problem in Section 5.2.

As another example, consider the problem of estimating treatment effects under unconfoundedness with limited overlap (Crump, Hotz, Imbens, and Mitnik, 2009; Khan and Tamer, 2010). Let $\tau(x)$ denote the treatment effect for individual with observables $X_i = x$. We would like to estimate $\theta(h) = E(\tau(X_i) \mid X_i \in \mathcal{X}_h)$, where $\mathcal{X}_h$ is a subset of the support of X_i, which corresponds to the average treatment effect for the subpopulation with $X_i \in \mathcal{X}_h$. The motivation is that treatment effects for individuals with propensity score $e(X_i) := P(D_i = 1|X_i)$ is close to zero or one cannot be estimated very precisely, so dropping these individuals from $\mathcal{X}_h$ will increase the precision of the resulting estimator $\hat{\theta}(h)$. On the other hand, increasing the set $\mathcal{X}_h$ yields an arguably more interesting estimand. Crump, Hotz, Imbens, and Mitnik (2009) propose to pick the set as $\mathcal{X}_h = \{X_i|h \leq e(X_i) \leq 1 - h\}$, with $\tau(\mathcal{X}_0)$ corresponding to the (unweighted) average treatment effect. In Section 5.3, we generalize our procedure to construct a uniform confidence

interval for $(\tau(X_h))_{h\in[\underline{h},\bar{h}]}$ (for this extension, the form of the adjustment is slightly different and involves the standard errors as well as the bandwidths; see equation (6) in Section 5.3 below).

In both setups, our procedure provides a simple solution to the problem of which particular $\theta(h)$ a researcher should report. With the reported uniform confidence band for $\theta(h)$, the reader can assess how $\theta(h)$ varies with h, or add bias corrections to the confidence interval at particular values of h to obtain a confidence interval for $\theta(0)$ based on the reader's own beliefs about the smoothness of $\theta(h)$.

5 Applications

This section gives primitive conditions for some applications.

5.1 Regression discontinuity with local polynomial estimator

We are interested in a regression discontinuity parameter, where the discontinuity point is normalized to $x = 0$ for convenience of notation. We consider both "sharp" and "fuzzy" regression discontinuity. For fuzzy regression discontinuity, we observe $\{(X_i, D_i, Y_i)\}_{i=1}^n$, and the parameter of interest is given by $\theta(0) = \frac{\lim_{x\downarrow 0} E(Y_i|X_i=x) - \lim_{x\uparrow 0} E(Y_i|X_i=x)}{\lim_{x\downarrow 0} E(D_i|X_i=x) - \lim_{x\uparrow 0} E(D_i|X_i=x)}$. For sharp regression discontinuity, we observe $\{(X_i, Y_i)\}_{i=1}^n$, and the parameter of interest is given by $\theta(0) = \lim_{x\downarrow 0} E(Y_i|X_i = x) - \lim_{x\uparrow 0} E(Y_i|X_i = x)$.

For ease of exposition, we focus on the commonly used local linear estimator. We cover the extension to local polynomial regression of higher order in Appendix S2. Using arguments in the discussion above Theorem 3.1, the results in this section could also be generalized to cover "kink" designs, where the focus is on estimating derivatives of conditional means at a point—in the interest of space, we do not pursue this extension here.

Given ome kernel function k^*, let $\hat{\alpha}_{\ell,Y}(h)$ and $\hat{\beta}_{\ell,Y}(h)$ minimize

$$\sum_{i=1}^{n} (Y_i - \alpha_{\ell,Y} - \beta_{\ell,Y} X_i)^2 \, I(X_i < 0) k^*(X_i/h)$$

and let $(\hat{\alpha}_{u,Y}(h), \hat{\beta}_{u,Y}(h))$ minimize

$$\sum_{i=1}^{n} (Y_i - \alpha_{u,Y} - \beta_{u,Y} X_i)^2 \, I(X_i \geq 0) k^*(X_i/h).$$

The sharp regression discontinuity estimator is given by $\hat{\theta}(h) = \hat{\alpha}_{u,Y}(h) - \hat{\alpha}_{\ell,Y}(h)$. For the fuzzy regression discontinuity estimator, the estimators $(\hat{\alpha}_{\ell,D}(h), \hat{\beta}_{\ell,D}(h))$ and $(\hat{\alpha}_{u,D}(h), \hat{\beta}_{u,D}(h))$ are defined analogously with D_i replacing Y_i, and the estimator is given by $\hat{\theta}(h) = \frac{\hat{\alpha}_{u,Y}(h) - \hat{\alpha}_{\ell,Y}(h)}{\hat{\alpha}_{u,D}(h) - \hat{\alpha}_{\ell,D}(h)}$.

For a given h, we define $\theta(h)$ as the statistic constructed from the population versions of these estimating equations, which leads to $\hat{\theta}(h)$ being approximately unbiased for $\theta(h)$. Let $(\alpha_{\ell,Y}(h), \beta_{\ell,Y}(h))$ minimize

$$E \, (Y_i - \alpha_{\ell,Y} - \beta_{\ell,Y} X_i)^2 \, I(X_i < 0) k^*(X_i/h),$$

and let $(\alpha_{u,Y}(h), \beta_{u,Y}(h))$, $(\alpha_{\ell,D}(h), \beta_{\ell,D}(h))$ and $(\alpha_{u,D}(h), \beta_{u,D}(h))$ be defined analogously. We define $\theta(h) = \frac{\alpha_{u,Y}(h) - \alpha_{\ell,Y}(h)}{\alpha_{u,D}(h) - \alpha_{\ell,D}(h)}$ for fuzzy regression discontinuity, and $\theta(h) = \alpha_{u,Y}(h) - \alpha_{\ell,Y}(h)$ for sharp regression discontinuity. Under appropriate smoothness conditions, $\theta(h)$ will converge to $\theta(0)$ as $h \to 0$.

Let $\mu_{k^*,j} = \int_{u=0}^{\infty} u^j k^*(u)$ for $j = 1, 2$. Under appropriate conditions, Assumption 3.1 holds with $k(u) = (\mu_{k^*,2} - \mu_{k^*,1}|u|)k^*(u)$. Thus, we can perform our procedure by looking up the critical value corresponding to $\overline{h}_n / \underline{h}_n$ and $k(u)$ (rather than the original kernel k^*) in our tables. For convenience, we report critical values for $k(u) = (\mu_{k^*,2} - \mu_{k^*,1}|u|)k^*(u)$ for some common choices of k^* in Table 1 for $\alpha = 0.05$ and Tables S1–S6 for for $\alpha = 0.01$ and $\alpha = 0.10$.

Theorem 5.1. *Suppose that*

(i) $|X_i|$ *has a density* $f_{|X|}(x)$ *at* $x = 0$, Y_i *is bounded, and, for some deterministic function* $\ell(t)$ *with* $\lim_{t \to 0} \log\log t^{-1} \ell(t) = 0$, *the functions* $f_X(x)$, $var((D_i, Y_i)' | X_i = x)$, $E(Y_i | X_i = x)$ *and* $E(D_i | X_i = x)$ *are left- and right-continuous at* 0 *with local modulus of continuity* $\ell(t)$.

(ii) $P(D_i = 1 | X_i = 0_+) - P(D_i = 1 | X_i = 0_-) \neq 0$ *and* $var(Y_i | D_i = d, X_i = 0_+) \neq 0$ *or* $var(Y_i | D_i = d, X_i = 0_-) \neq 0$ *for* $d = 0$ *or* 1.

Then, for $\hat{\theta}(h)$ *and* $\theta(h)$ *given above and* $\hat{\sigma}(h)$ *given in the appendix, if the kernel function* k^* *satisfies part (i) of Assumption 3.2, then Assumptions 3.1 and Assumption 3.2 hold with* $k(u) = (\mu_{k^*,2} -$

$\mu_{k^*,1}|u|)k^*(u)$, *so long as* $\overline{h}_n$ *is bounded by a small enough constant and* $n\underline{h}_n/(\log\log \underline{h}_n^{-1})^3 \to \infty$.

5.2 LATE on the largest sets of compliers

We observe (Z_i, D_i, Y_i) where Z_i is an exogenous variable shifting a zero-one treatment variable D_i, and Y_i is an outcome variable. Let $[\underline{z}, \overline{z}]$ be the support of Z_i, and assume, for simplicity, that $\underline{z}$ and $\overline{z}$ are finite (this does not involve much loss in generality, since Z_i can always be transformed to the unit interval by redefining Z_i as its percentile rank).

Given sets $\mathcal{A}$ and $\mathcal{B}$, define

$$\Delta^{\text{LATE}}(\mathcal{A}, \mathcal{B}) = \frac{E(Y_i|Z_i \in \mathcal{A}) - E(Y_i|Z_i \in \mathcal{B})}{P(D_i = 1|Z_i \in \mathcal{A}) - P(D_i = 1|Z_i \in \mathcal{B})}.$$

Under certain exogeneity and monotonicity assumptions, $\Delta^{\text{LATE}}(\mathcal{A}, \mathcal{B})$ gives the average effect on Y_i of treating an individual i for a certain subpopulation, where the subpopulation depends on $\mathcal{A}$ and $\mathcal{B}$. In the literature, this is called the "local average treatment effect" on this subpopulation, and the subpopulation is termed "compliers" (see Heckman and Vytlacil, 2005; Heckman, Urzua, and Vytlacil, 2006; Imbens and Angrist, 1994). Suppose that $P(D_i = 1|Z_i = z)$ is increasing in z. In this case, $\Delta^{\text{LATE}}(\mathcal{A}, \mathcal{B})$ is often of particular interest for $\mathcal{A} = [\underline{z}, \underline{z} + h]$ and $\mathcal{B} = [\overline{z} - h, \overline{z}]$ for small h, since, under certain monotonicity restrictions, the subpopulation associated with $\Delta^{\text{LATE}}([\underline{z}, \underline{z} + h], [\overline{z} - h, \overline{z}])$ approaches the largest possible subpopulation for which the LATE is identified as $h \to 0$ (see Frölich, 2007; Heckman and Vytlacil, 2005; Heckman, Urzua, and Vytlacil, 2006). Let $\theta(h) = \Delta^{\text{LATE}}([\underline{z}, \underline{z} + h], [\overline{z} - h, \overline{z}])$, and suppose that h is small enough that these sets are nonoverlapping. We estimate $\theta(h)$ with the sample analogue

$$\hat{\theta}(h) = \frac{\frac{1}{\#\{Z_i \in [\underline{z}, \underline{z}+h]\}} \sum_{Z_i \in [\underline{z}, \underline{z}+h]} Y_i - \frac{1}{\#\{Z_i \in [\overline{z}-h, \overline{z}]\}} \sum_{Z_i \in [\overline{z}-h, \overline{z}]} Y_i}{\frac{1}{\#\{Z_i \in [\underline{z}, \underline{z}+h]\}} \sum_{Z_i \in [\underline{z}, \underline{z}+h]} D_i - \frac{1}{\#\{Z_i \in [\overline{z}-h, \overline{z}]\}} \sum_{Z_i \in [\overline{z}-h, \overline{z}]} D_i}.$$

It can be shown that $\hat{\theta}(h)$ is numerically identical to the instrumental variables estimator for β in the equation $Y_i = \alpha + D_i\beta + \varepsilon$, where the sample is restricted to observations with $Z_i \in [\underline{z}, \underline{z} + h] \cup [\overline{z} - h, \overline{z}]$ and the instrument is $I(Z_i \geq \overline{z} - h)$. We define $\hat{\sigma}^2(h)/h$ to be the robust variance estimate for $\sqrt{n}(\hat{\beta} - \beta)$ from this IV regression, so that $\hat{\sigma}(h)/\sqrt{nh} = se(h)$ is the standard error for $\hat{\theta}(h)$.

Since $\hat{\theta}(h)$ is composed of kernel based estimators with the uniform kernel (e.g. with the uniform kernel, $\frac{1}{\#\{Z_i \in [\underline{z}, \underline{z}+h]\}} \sum_{Z_i \in [\underline{z}, \underline{z}+h]} Y_i$ is an estimate of $E[Y_i | Z_i = \underline{z}]$), we expect that our results hold with k given by the uniform kernel $k(u) = I(|u| \leq 1)$. The following theorem shows that this holds under appropriate regularity conditions.

Theorem 5.2. *Suppose that*

(i) Z_i *has a density* $f_Z(z)$ *at* $z = \underline{z}$ *and* $z = \overline{z}$, Y_i *is bounded and, for some function* $\ell(t)$ *with* $\lim_{t \to 0} \log\log t^{-1} \ell(t) = 0$, f_Z, $var((D_i, Y_i)' | Z_i = z)$, $E(Y_i | Z_i = z)$ *and* $E(Z_i | Z_i = z)$ *are continuous at* $\underline{z}$ *and* $\overline{z}$ *with local modulus of continuity* $\ell(t)$.

(ii) $P(D_i = 1 | Z_i = \overline{z}) - P(D_i = 1 | Z_i = \underline{z}) \neq 0$ *and* $var(Y_i | D_i = d, z_i = \underline{z}) \neq 0$ *or* $var(Y_i | D_i = d, Z_i = \overline{z}) \neq 0$ *for* $d = 0$ *or* 1.

Then, for $\hat{\theta}(h)$, $\theta(h)$ *and* $\hat{\sigma}(h)$ *given above, Assumptions 3.1 and Assumption 3.2 hold with* $k(u) = I(|u| \leq 1)$, *so long as* $\overline{h}_n$ *is bounded by a small enough constant and* $n\underline{h}_n / (\log\log \underline{h}_n^{-1})^3 \to \infty$.

Thus, one can compute critical values based on Table 1 and Tables S1–S6, corresponding to the uniform kernel.

In contrast to the regression discontinuity setup of Section 5.1, in which $\theta(h)$ was of interest mainly as a biased estimate of $\theta(0)$, the parameter $\theta(h) = \Delta^{\text{LATE}}([\underline{z}, \underline{z} + h], [\overline{z} - h, \overline{z}])$ has an interpretation for fixed h as the average treatment effect on a subset of the population, where the subset depends on h. Our procedure provides a simple way of summarizing the estimates of $\theta(h)$ for a range of values of h and their statistical accuracy, while formally taking into account that one has looked at multiple estimates.

5.3 Trimmed average treatment effects under unconfoundedness

We extend our setting to obtain uniform confidence bands for average treatment effects (ATEs) on certain subpopulations under an unconfoundedness assumption. Here, the adjustment is slightly different, but can still be computed using our tables along with quantities that are routinely reported in applied research. We explain this further below.

We observe $\{(X_i, D_i, Y_i)\}_{i=1}^n$ iid, where $Y_i = Y_i(D_i)$, D_i is a Bernoulli random variable conditional on X_i, and $E(Y_i(d) | X_i, D_i) = E(Y_i(d) | X_i)$. Let $\mu_d(x) = E(Y_i | X_i = x, D_i = d)$, and let $\tau(x) =$

$E(Y_i(1) - Y_i(0)|X_i = x) = E(Y_i|X_i = x, D_i = 1) - E(Y_i(0)|X_i = x, D_i = 0) = \mu_1(x) - \mu_0(x)$. Let $e(x) = P(D_i = 1|X_i = x)$. We consider inference on the conditional average treatment effect for the set $\mathcal{X}_h = \{h \leq e(X_i) \leq 1 - h\}$ (where $0 \leq h < 1/2$), given by

$$\theta(h) = E(Y_i(1) - Y_i(0)|X_i \in \mathcal{X}_h) = E(\tau(X_i)|X_i \in \mathcal{X}_h).$$

As discussed above in Example 4.3, the motivation for looking at $\theta(h)$ rather than the average treatment for the entire population ($\theta(0)$ in our notation), is that the average treatment effect will be difficult to estimate when $e(X_i)$ is close to zero or one with nonnegligible probability. On the other hand, it is often the ATE on the full sample that is of interest, and in which case reporting $\hat{\theta}(h)$ for $h > 0$ gives a more accurate estimator, but a less interesting estimand. Our approach of reporting a uniform confidence band allows the researcher to avoid the issue of which trimmed estimate to report and simply report a range of estimates. See Crump, Hotz, Imbens, and Mitnik (2009), Hill (2013) and Khan and Tamer (2010) for further discussion of these issues.

Let $\hat{\theta}(h)$ be an estimator of $\theta(h)$ with influence function representation

$$\sqrt{n}(\hat{\theta}(h) - \theta(h)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{[\tilde{Y}_i - \theta(h)]I(X_i \in \mathcal{X}_h)}{P(X_i \in \mathcal{X}_h)} + o_P(1), \tag{5}$$

where the $o_P(1)$ term is uniform over $\underline{h} \leq h \leq \overline{h}$ and $\tilde{Y}_i := D_i \frac{Y_i - \mu_1(X_i)}{e(X_i)} - (1 - D_i)\frac{Y_i - \mu_0(X_i)}{1 - e(X_i)} + \mu_1(X_i) - \mu_0(X_i)$ (see Crump, Hotz, Imbens, and Mitnik, 2009, for references to the literature for estimators that satisfy this condition). Note that $E(\tilde{Y}_i|X_i) = \tau(X_i)$ so that $E(\tilde{Y}_i|X_i \in \mathcal{X}_h) = \theta(h)$. Let

$$\sigma(h)^2 = var\left\{\frac{[\tilde{Y}_i - \theta(h)]I(X_i \in \mathcal{X}_h)}{P(X_i \in \mathcal{X}_h)}\right\} = \frac{var\left\{[\tilde{Y}_i - \theta(h)]I(X_i \in \mathcal{X}_h)\right\}}{P(X_i \in \mathcal{X}_h)^2},$$

and let $\hat{\sigma}(h)$ be a uniformly (over $\underline{h} \leq h \leq \overline{h}$) consistent estimator of $\sigma(h)$.

In contrast to the previous applications, we assume that $\underline{h}$ and $\overline{h}$ are fixed. In settings where $e(X_i)$ is close to zero or one with high probability, the variance bound for the ATE, $\theta(0)$, may be infinite, and a sequence of trimming points $h_n \to 0$ can be used to obtain estimators that converge to the ATE at a slower than root-n rate (see Khan and Tamer, 2010). We expect that our results can be extended to this case under appropriate regularity conditions, but we leave this

question for future research.

To describe the adjustment in this setting, let $N(h) = \#\{i|X_i \in \mathcal{X}_h\}$ be the number of untrimmed observations for a given h, and let $se(h) = \hat{\sigma}(h)/\sqrt{n}$ be the standard error for a given h. We form our uniform confidence band as

$$\left\{\hat{\theta}(h) \pm c_{1-\alpha,|\cdot|}(\hat{t}, k_{\text{uniform}}) \cdot se(h) \middle| h \in [\underline{h}, \overline{h}]\right\} \text{ where } \hat{t} = \frac{se(\underline{h})^2 N(\underline{h})^2}{se(\overline{h})^2 N(\overline{h})^2} \tag{6}$$

(here, k_{uniform} denotes the uniform kernel).

The critical value given above comes from an approximation by a scaled Brownian motion where the "effective sample size" is proportional to a quantity that can be estimated by $se(h)^2 N(h)^2$. See Section S3.2 in the supplemental appendix for details.

The following theorem proves the validity of this confidence band. In the interest of space, we state only the two-sided version.

Theorem 5.3. *Let $0 \leq \underline{h} < \overline{h} < 1/2$. Suppose that*

(i) the influence function representation (5) holds uniformly over $\underline{h} \leq h \leq \overline{h}$, and $se(h) = \hat{\sigma}(h)/\sqrt{n}$ where $\hat{\sigma}(h)$ is consistent for $\sigma(h)$ unfiromly over $\underline{h} \leq h \leq \overline{h}$

(ii) $\theta(h)$ is bounded uniformly over $\underline{h} \leq h \leq \overline{h}$ and $E[\tilde{Y}_i^2|X_i]$ is bounded uniformly over $\underline{h} \leq e(X_i) \leq 1 - \underline{h}$ and

(iii) $v(\overline{h}) > 0$ where $v(h) = E\{[\tilde{Y}_i - \theta(h)]^2 I(X_i \in \mathcal{X}_h)\}$.

Let $\hat{t} = \frac{se(\underline{h})^2 N(\underline{h})^2}{se(\overline{h})^2 N(\overline{h})^2}$ as defined in (6). Then

$$\liminf_n P\left(\frac{\sqrt{n}\left|\hat{\theta}(h) - \theta(h)\right|}{\hat{\sigma}(h)} \leq c_{1-\alpha,|\cdot|}(\hat{t}, k_{uniform}) \text{ all } h \in [\underline{h}, \overline{h}]\right) \geq 1 - \alpha$$

where $k_{uniform}$ is the uniform kernel. If, in addition, $v(h)$ is continuous, the above display holds with the lim inf *replaced by* $\lim_{n\to\infty}$ *and* $\geq$ *replaced by* $=$.

As an example, Crump, Hotz, Imbens, and Mitnik (2009) report estimates based on a study of right heart catheterization (the variable D_i being 1 if patient i received this treatment), with controls X_i reported in that paper and an indicator for 30 day survival as the outcome variable Y_i.

They report an estimate of the average treatment effect (on the full population) of -0.0593 with a standard error of .0167. They also report an estimate of -0.0590 with a standard error of 0.0143 for the average treatment effect conditional on covariates X_i such that $.1 \leq e(X_i) \leq .9$. They report that the data set contains 5735 observations, of which 4728 are in the smaller subsample with $.1 \leq e(X_i) \leq .9$. This gives

$$\hat{t} = \frac{se(\underline{h})^2 N(\underline{h})^2}{se(\overline{h})^2 N(\overline{h})^2} = \frac{0.0167^2 \cdot 5735^2}{0.0143^2 \cdot 4728^2} \approx 2.007.$$

For $\alpha = .05$ the two-sided critical value $c_{.95,|\cdot|}(2.007, k_{\text{uniform}})$ is approximately 2.50 (using the column corresponding to the uniform kernel in Table 1). Thus, the snooping adjusted confidence intervals for the (unconditional) average treatment effect and the average treatment effect conditional on $.1 \leq e(X_i) \leq .9$ are $-0.0593 \pm 2.50 \cdot 0.0167 = [-0.1011, -0.0176]$ and $-0.0590 \pm 2.50 \cdot 0.0143 = [-0.0950, -0.0233]$ respectively. The pointwise CIs are $-0.0593 \pm 1.96 \cdot 0.0167 = [-0.0920, -0.0266]$ and $-0.0590 \pm 1.96 \cdot 0.0143 = [-0.0870, -0.0310]$. Note that the adjusted confidence intervals reported above allow for snooping over $0 \leq h \leq .1$, so they are conservative if we tie our hands to look only at $h = 0$ or $h = .1$.

6 Empirical illustrations

6.1 U.S. House elections

Our first empirical example is based on Lee (2008), who is interested in the effect of an incumbency advantage in U.S. House elections. Given the inherent uncertainty in final vote counts, the party that wins is essentially randomized in elections that are decided by a narrow margin, which suggests using a sharp regression discontinuity design to identify the incumbency advantage.

In particular, the running variable X_i is the Democratic margin of victory in a given election i. Thus, if Democrats won election i, X_i will be positive, and it will be negative if they lost. The outcome variable Y_i is the Democratic vote share in the next election. The parameter $\theta(0)$ is then the incumbency advantage for Democrats—the impact of being the current incumbent party in a congressional district on the probability of winning the next election.

There are $6,558$ observations in this dataset, spanning House elections between 1946 and 1998. The average difference in vote share is 0.13 for Democrats, with standard deviation 0.46.

To analyze the data, Lee (2008) uses a global fourth degree polynomial, which yields a point estimate of 7.7%. However, because estimates may be sensitive to the degree of polynomial, and may give large weights to observations far away from the threshold, global polynomial estimates may be misleading (Gelman and Imbens, 2014). We therefore reanalyze the data using local linear regression with a triangular kernel. Figure 3 plots the results for bandwidths between 0.02 and 0.4. The vertical line corresponds to estimates based on bandwidth selector proposed by Imbens and Kalyanaraman (2012, IK), which yields a point estimate of 7.99%, close to Lee's original estimate. The incumbency effect remains positive and significant over the entire range, even after using the corrected critical value $c_{0.95}(0.4/0.02, \text{triangular}) = 2.526$. At the IK bandwidth, the unadjusted confidence intervals are given by $(6.49, 9.50)$. Our adjustment widens them slightly to $(6.05, 9.93)$. These results suggest that the estimates are very robust to the choice of bandwidth.

6.2 Progresa / Oportunidades

Our second empirical example examines the effect of the Oportunidades (previously known as Progresa) anti-poverty conditional cash transfer program in Mexico, using a dataset from Calonico, Cattaneo, and Titiunik (2014, CCT). The program started in 1998 under the name of Progresa in rural areas, and expanded to urban areas in 2003. The program is designed to target poverty by providing cash payments to families in exchange for regular school attendance, health clinic visits, and nutritional support. The transfer constituted a significant contribution to the income of eligible families.

We focus on the program treatment effect in the urban areas. Here, unlike in the rural areas, the program was first offered in neighborhoods with the highest density of poor households. In order to accurately target the program to poor households, household eligibility to participate in the program was based on a pre-intervention household poverty index. This eligibility assignment rule naturally leads to sharp (intention-to-treat) regression-discontinuity design.

As in CCT, we focus on the effect of the program on food and non-food consumption expenditures two years after its implementation (consumption is measured in pesos, expressed as monthly expenditures per household member). We normalize the poverty index so that the participation cutoff is zero. There are 2,809 households in the dataset, 691 with index $X_i > 0$, and 2,118 controls with $X_i < 0$. For the effect on food consumption, the IK bandwidth selector sets

$h_{IK} = 1.44$, with 95% confidence interval equal to $(5.0, 72.9)$, suggesting a significantly positive effect (the t-statistic equals 2.25). For non-food consumption, $h_{IK} = 1.09$, and the 95% confidence interval equals $(-0.4, 55.7)$, with p-value equal to 0.053.

As pointed out by CCT, and as we argue in Example 4.1, one may be concerned that the confidence intervals based on the IK bandwidth will not be accurate since the asymptotic distribution of $\hat{\theta}(h_{IK})$ is biased due to the MSE-optimal IK bandwidth being too large. In Figure 4 we plot the estimates, along with pointwise and uniform confidence bands over a range of bandwidths. In contrast to the first empirical example, the figures indicate that the results are sensitive to bandwidth choice: the uniform bands contain zero over the entire range plotted for both outcomes.

7 Monte Carlo evidence

We conduct a small Monte Carlo study of inference in a sharp regression discontinuity design to further illustrate our method and examine how well it works in practice. In each replication, we generated a random sample $\{X_i, \epsilon_i\}_{i=1}^n$, with size $n = 500$, $X_i = 2Z_i - 1$, where Z_i has Beta distribution with parameters 2 and 4, and $\epsilon_i \sim \mathcal{N}(0, 0.1295^2)$. The regression discontinuity point is normalized to zero. The outcome Y_i is given by $Y_i = g_j(X_i) + \epsilon_i$, where the regression function g_j depends on the design. We consider two regression functions. The first one is based on data in Lee (2008),

$$g_1(x) = \begin{cases} 0.48 + 1.27x + 7.18x^2 + 20.21x^3 + 21.54x^4 + 7.33x^5 & \text{if } x < 0, \\ 0.52 + 0.84x - 3.00x^2 + 7.99x^3 - 9.01x^4 + 3.56x^5 & \text{otherwise.} \end{cases}$$

This design corresponds exactly to the data generating process in Imbens and Kalyanaraman (2012, IK) and Calonico, Cattaneo, and Titiunik (2014, CCT). The second regression function corresponds to another design in IK, and is given by

$$\mu(x) = 0.42 + 0.1I(x \geq 0) + 0.84x + 7.99x^3 - 9.01x^4 + 3.56x^5.$$

In addition, we also considered designs in which the error term ϵ_i is heteroscedastic. The results for these designs are very similar and reported in the supplemental appendix.

Figure 5 plots the two conditional expectation functions. In each design, we consider estimates based on local linear and local quadratic regression using the uniform and the triangular kernel. Figure 6 plots the function $\theta(h)$ for estimators based on local linear regression that uses these two kernels. Plots of $\theta(h)$ for the local quadratic estimator based on each kernel are given in Figure S1 in the supplemental appendix.

We use the bandwidth selector proposed by IK to select a baseline bandwidth, and then construct confidence bands for estimators in bandwidth range around this baseline bandwidth.

To examine sensitivity of the results to the choice of variance estimator $\hat{\sigma}^2(h)$, we consider four methods for computing the variance. The first estimator corresponds to the Eicker-Huber-White (EHW) robust standard error estimator that treats the two local linear linear regressions on either side of the cutoff as a standard weighted regression. In Theorem 5.1 above, we show formally that using this estimator leads to uniformly valid confidence intervals. The second estimator corresponds to a modification of the EHW estimator proposed by Calonico, Cattaneo, and Titiunik (2014) that uses a nearest neighbor (NN) estimator to estimate $var(Y_i \mid X_i)$ in the middle part of the Eicker-Huber-White "sandwich," rather than using the regression residuals. The third estimator corresponds to the plug-in estimator of the asymptotic variance proposed by IK.

The fourth method corresponds to a particular case of the robust confidence interval proposed by Calonico, Cattaneo, and Titiunik (2014, CCT). In particular, we run a local quadratic instead of local linear regression to construct a point estimate, and use the NN variance estimator to estimate $\hat{\sigma}(h)$ (results for other variance estimators are similar, and not reported here, but available upon request). As explained in CCT, the rationale for this procedure is since the IK bandwidth is optimal for estimation, it balances squared-bias and variance of the RD estimator. Consequently, the bandwidth will be too large in the sense that $\sqrt{nh}(\theta(h) - \theta(0))$ will not be asymptotically negligible, confidence intervals based around the IK bandwidth are likely to have poor coverage of $\theta(0)$. CCT show that using the IK bandwidth and local quadratic regression is equivalent to recentering the confidence interval based on local linear regression by subtracting an estimate of the asymptotic bias, and rescaling it to account for the bias estimation. Alterna-

tively, since optimal bandwidth for local quadratic regression will in general be larger than the optimal bandwidth for local linear regression, this method of constructing confidence intervals can be viewed as a particular undersmoothing procedure.

Finally, we also report results based on the true (but in practice infeasible) variance, $var(\hat{\theta}(h) - \theta(h))$. The supplemental appendix gives detailed description of these five estimators.

Tables 2 and 3 report empirical coverage of the confidence bands for $\theta(h)$ for the two designs we consider. Our adjustment works well overall, with the empirical coverage being close to 95% for most specifications, in contrast with the naive confidence bands (using the unadjusted 1.96 critical value), which undercover. As plotted in Figure 2, Theorem 3.1 predicts that with $\bar{h}/\underline{h} = 2$, the coverage should be 91.6% for the triangular kernel, and 83.9% for the uniform kernel. When $\bar{h}/\underline{h} = 4$, the coverage of the naive confidence bands should drop to 88.5% and 76.8%, respectively. The Monte Carlo results match these predictions closely.

There are a few specifications in Design 2 with the triangular kernel, in which the empirical coverage of the adjusted confidence bands is below 95%. Comparing their coverage with the coverage of the pointwise confidence intervals for the same range of bandwidths indicates that this problem arises because the pointwise confidence intervals fail to achieve nominal coverage in the first place. Since our method only corrects for the multiple comparisons, it cannot solve this problem. Overall, the adjusted confidence bands have coverage that is as good as the coverage of the underlying pointwise confidence intervals.

Typically in regression discontinuity studies, the primary object of interest is $\theta(0)$, the average treatment effect conditional on $X = 0$, rather than $\theta(h)$. We therefore also report empirical coverage of the confidence bands for $\theta(0)$ in Tables 4 and 5. At larger values of the bandwidth, $\hat{\theta}(h)$ is a biased estimator of $\theta(0)$. The pointwise confidence bands based on the local linear regression do not take this bias into account, and they fail to achieve proper coverage, especially for the bandwidth ranges where $\overline{h}_n$ equals twice the IK bandwidth. Consequently, although our adjustment ensures that the coverage of the adjusted confidence band is within the range of the pointwise confidence intervals, it still falls short of 95% due to the pointwise confidence intervals performing poorly.

On the other hand, so long as we undersmooth, the empirical coverage of $\theta(0)$ remains good, especially when the nearest neighbor variance estimator is used. This is borne out in the simu-

lations that correspond to the bandwidth range $[\hat{h}_{IK}/4, \hat{h}_{IK}/2]$. Similarly, the bias-corrected CCT estimator based on local quadratic regression performs well, especially when $\overline{h}_n$ is no larger than the IK bandwidth.

In conclusion, our adjustment performs well in terms of coverage of $\theta(h)$, with empirical coverage close to nominal coverage for a range of variance estimators and Monte Carlo designs. If our method is combined with undersmoothing (corresponding to bandwidth ranges such that $\overline{h}_n$ is not too large), or bias-correction (such as when the CCT method for constructing confidence intervals is used), so that the underlying pointwise confidence intervals achieve good coverage of $\theta(0)$, our method also achieves good coverage of $\theta(0)$.

8 Conclusion

Nonparametric estimators typically involve a choice of tuning parameter. To ensure robustness of the results to tuning parameter choice, researchers often examine sensitivity of the results to the value of the tuning parameter. However, if the tuning parameter is chosen based on this sensitivity analysis, the resulting confidence intervals may undercover even if the estimator is unbiased.

In this paper, we addressed this problem when the estimator is kernel-based, and the tuning parameter is a bandwidth. We showed that if one uses an adjusted critical value instead of the usual critical value based on quantiles of a Normal distribution, the resulting confidence interval will be robust to this form of "bandwidth snooping."

The adjustment only depends on the kernel and the ratio of biggest to smallest bandwidth that the researcher has tried. Therefore, readers can easily quantify the robustness of reported results to the bandwidth choice, as long as both a point estimate and a standard error have been reported. Our method also allows researchers to report the results for a range of bandwidths along with the adjusted confidence bands as a routine robustness check, allowing readers to select their own bandwidth.

Appendix

This appendix contains the proof of Theorem 3.1 in the main text, as well as auxiliary results. Section A contains the proof of the main result. Section B discusses the use of uniform and pointwise in h confidence regions in sensitivity analysis. Additional results, including verification of our conditions in the applications in Section 5, are in the supplemental appendix.

Throughout this appendix, we use the following additional notation. For a sample $\{Z_i\}_{i=1}^n$ and a function f on the sample space, $E_n f(Z_i) = \frac{1}{n}\sum_{i=1}^n f(Z_i)$ denotes the sample mean, and $\mathbb{G}_n f(Z_i) = \sqrt{n}(E_n - E)f(Z_i) = \sqrt{n}[E_n f(Z_i) - Ef(Z_i)]$ denotes the empirical process. We use $t \vee t'$ and $t \wedge t'$ to denote elementwise maximum and minimum, respectively. We use e_k to denote the kth basis vector in Euclidean space (where the dimension of the space is clear from context).

A Proof of Main Result

A.1 Equivalence Results for Extreme Value Limits

This section proves an equivalence result for extreme value limits of the form proved in this paper. We begin with the following result.

Theorem A.1. *Let h_n^* and $\underline{h}_n$ be sequences with $\underline{h}_n \to 0$, $h_n^* = \mathcal{O}(1)$ and $h_n^*/\underline{h}_n \to \infty$, and let $\mathbb{T}_n(h)$ and $\tilde{\mathbb{T}}_n(h)$ be random processes on $\mathbb{R}$. Suppose that*

$$\sqrt{2\log\log(h_n^*/\underline{h}_n)}\left(\sup_{\underline{h}_n \le h \le h_n^*} \mathbb{T}_n(h) - \sqrt{2\log\log(h_n^*/\underline{h}_n)}\right) - b(\log\log(h_n^*/\underline{h}_n)) \xrightarrow{d} Z \tag{7}$$

for some limiting variable Z and $b(t) = \log c_2$ or $b(t) = \log c_1 + \log\sqrt{2t}$ for some constants c_1 and c_2.

Suppose that

$$\sqrt{\log\log(h_n^*/\underline{h}_n)} \sup_{\underline{h}_n \le h \le h_n^*} \left|\mathbb{T}_n(h) - \tilde{\mathbb{T}}_n(h)\right| \xrightarrow{p} 0. \tag{8}$$

Then (7) holds with $\mathbb{T}_n(h)$ replaced by $\tilde{\mathbb{T}}_n(h)$. If, in addition, for some sequence $\overline{h}_n$ with $\overline{h}_n \ge h_n^*$, $\log\log(h_n^*/\underline{h}_n) - \log\log(\overline{h}_n/\underline{h}_n) \to 0$ and, for some $\varepsilon > 0$,

$$\frac{\sup_{h_n^* \le h \le \overline{h}_n} \tilde{\mathbb{T}}_n(h)}{\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}} \le 1 - \varepsilon \text{ with probability approaching one,} \tag{9}$$

then (7) holds with $\mathbb{T}_n(h)$ replaced by $\tilde{\mathbb{T}}_n(h)$ and h_n^* replaced by $\overline{h}_n$.

Proof. The first claim is immediate from the bound $\left|\sup_{\underline{h}_n \le h \le h_n^*} \mathbb{T}_n(h) - \sup_{\underline{h}_n \le h \le h_n^*} \tilde{\mathbb{T}}_n(h)\right| \le \sup_{\underline{h}_n \le h \le h_n^*} \left|\mathbb{T}_n(h) - \tilde{\mathbb{T}}_n(h)\right|$ and Slutsky's theorem.

For the second claim, note that, since (7) holds for $\tilde{\mathbb{T}}_n$, $\sup_{\underline{h}_n \le h \le h_n^*} \tilde{\mathbb{T}}_n(h) / \sqrt{2 \log\log \overline{h}_n / \underline{h}_n} \xrightarrow{p} 1$ so that, with probability approaching one, $\sup_{\underline{h}_n \le h \le \overline{h}_n} \tilde{\mathbb{T}}_n(h) = \sup_{\underline{h}_n \le h \le h_n^*} \mathbb{T}_n(h)$. By Slutsky's theorem, $a_n X_n - b_n \xrightarrow{d} Z$ implies $a_n' X_n - b_n' \xrightarrow{d} Z$ so long as $b_n - b_n' \to 0$ and $(a_n - a_n') \frac{1 \vee b_n}{a_n} \to 0$ (note that $(a_n - a_n')X_n - (b_n - b_n') = \frac{a_n - a_n'}{a_n}(a_n X_n - b_n) + \frac{b_n}{a_n}(a_n - a_n') - (b_n - b_n')$). Applying this fact with $a_n = \sqrt{2 \log\log(h_n^* / \underline{h}_n)}$, $a_n' = \sqrt{2 \log\log(\overline{h}_n / \underline{h}_n)}$, $b_n = 2 \log\log(h_n^* / \underline{h}_n) + b(\log\log(h_n^* / \underline{h}_n))$ and $b_n' = 2 \log\log(\overline{h}_n / \underline{h}_n) + b(\log\log(\overline{h}_n / \underline{h}_n))$, we have

$$
\begin{aligned}
(a_n - a_n') \frac{1 \vee b_n}{a_n} &= \left(\sqrt{2 \log\log(h_n^* / \underline{h}_n)} - \sqrt{2 \log\log(\overline{h}_n / \underline{h}_n)} \right) \frac{2 \log\log(h_n^* / \underline{h}_n) + b(\log\log(h_n^* / \underline{h}_n))}{\sqrt{2 \log\log(h_n^* / \underline{h}_n)}} \\
&= \left(\sqrt{2 \log\log(h_n^* / \underline{h}_n)} - \sqrt{2 \log\log(\overline{h}_n / \underline{h}_n)} \right) \left(\sqrt{2 \log\log(h_n^* / \underline{h}_n)} + o(1) \right) \\
&= \frac{2 \log\log(h_n^* / \underline{h}_n) - 2 \log\log(\overline{h}_n / \underline{h}_n)}{\sqrt{2 \log\log(h_n^* / \underline{h}_n)} + \sqrt{2 \log\log(\overline{h}_n / \underline{h}_n)}} \left(\sqrt{2 \log\log(h_n^* / \underline{h}_n)} + o(1) \right) \to 0
\end{aligned}
$$

and $b_n - b_n' = b(\log\log(h_n^* / \underline{h}_n)) - b(\log\log(\overline{h}_n / \underline{h}_n)) + o(1) \to 0$ since $|b(t) - b(t')| \le t - t'$ for large enough t and t'. □

To prove our main result, we apply Theorem A.1 twice. First, we show that, under the conditions of Theorem 3.1, for some $\varepsilon > 0$,

$$
\frac{\sup_{h_n^* \le h \le \overline{h}_n} \sqrt{nh}(\hat{\theta}(h) - \theta(h)) / \hat{\sigma}(h)}{\sqrt{2 \log\log \overline{h}_n / \underline{h}_n}} = \frac{\sup_{h_n^* \le h \le \overline{h}_n} \frac{1}{\sqrt{nh}} \sum_{i=1}^n \psi(W_i, h) k(X_i / h)}{\sqrt{2 \log\log \overline{h}_n / \underline{h}_n}} + o_P(1) \le 1 - \varepsilon
$$

with probability approaching one, where

$$
h_n^* = \exp\left[-(\log \underline{h}_n^{-1})^{1/K} \right] \tag{10}
$$

for K large enough (the reasoning behind this choice of h_n^* is explained below; in the case where $\overline{h}_n$ goes to zero more quickly than this choice of h_n^*, this step can be skipped). For this choice of h_n^*, (8) is shown to hold with $\tilde{\mathbb{T}}_n(h)$ given by $\frac{\sqrt{nh}(\hat{\theta}(h) - \theta(h))}{\hat{\sigma}(h)}$ and $\mathbb{T}_n(h)$ given by $\frac{1}{\sqrt{nh}} \sum_{i=1}^n \tilde{Y}_i k(X_i / h)$,

where

$$\tilde{Y}_i = \frac{\psi(W_i, 0) - E[\psi(W_i, 0)||X_i|]}{\sqrt{var(\psi(W_i, 0)||X_i|) f_{|X|}(|X_i|) \int_0^\infty k(u)^2 \, du}}. \tag{11}$$

Next, it is shown that (8) holds for $\tilde{\mathbb{T}}_n(h)$ given by $\frac{1}{\sqrt{nh}} \sum_{i=1}^n \tilde{Y}_i k(X_i/h)$ and $\mathbb{T}_n(h)$ given by a Gaussian process with the same covariance kernel, which can be constructed on the same sample space. Calculating this covariance kernel, we see that

$$\begin{aligned} cov\left(\frac{1}{\sqrt{nh}} \sum_{i=1}^n \tilde{Y}_i k(X_i/h), \frac{1}{\sqrt{nh}} \sum_{i=1}^n \tilde{Y}_i k(X_i/h') \right) &= E \frac{1}{\sqrt{hh'}} E[\tilde{Y}_i^2 ||X_i|] k(|X_i|/h) k(|X_i|/h') \\ &= E\left\{ \frac{1}{\sqrt{hh'}} \left[f_{|X|}(|X_i|) \int_0^\infty k(u)^2 \, du \right]^{-1} k(|X_i|/h) k(|X_i|/h') \right\} = \frac{\int k(x/h) k(x/h') \, dx}{\sqrt{hh'} \int k(u)^2 \, du} \end{aligned}$$

(here, we use the fact that $k(|X_i|/h) = k(X_i/h)$ and $\int k(u)^2 \, du = 2 \int_0^\infty k(u)^2 \, du$, since k is symmetric). The change of variables $u = x/h'$ shows that the covariance kernel depends only on h'/h, so that the Gaussian process is stationary when indexed by $t = \log h$. The result then follows by applying a theorem for limits of stationary Gaussian processes on increasing sets (see Leadbetter, Lindgren, and Rootzen, 1983).

The reasoning behind this choice of h_n^* is as follows. With $h_n^* = \exp\left[-(\log \underline{h}_n^{-1})^{1/K} \right]$, we have $h_n^*/\underline{h}_n = \exp\left[-(\log \underline{h}_n^{-1})^{1/K} + (\log \underline{h}_n^{-1}) \right] = \exp\left\{ (\log \underline{h}_n^{-1})[1 - (\log \underline{h}_n^{-1})^{1/K - 1}] \right\}$ so that $\log\log(h_n^*/\underline{h}_n) = \log\{(\log \underline{h}_n^{-1})[1 - (\log \underline{h}_n^{-1})^{1/K-1}]\} = \log\log \underline{h}_n^{-1} + \log[1 - (\log \underline{h}_n^{-1})^{1/K-1}]$. Since the last term converges to zero, this is equal to $\log\log \underline{h}_n^{-1}$ up to an $o(1)$ term, and the same holds for $\log\log \overline{h}_n/\underline{h}_n$ as required.

To see why this choice of h_n^* is useful for showing (9), note that, if the supremum of $\tilde{\mathbb{T}}_n(h)$ increases at the same rate over $h_n^* \leq h \leq \overline{h}_n$ (as a function of $\overline{h}_n/h_n^*$) as it does over $\underline{h}_n \leq h \leq h_n^*$ (as a function of $h_n^*/\underline{h}_n$), then we will have, for some constant C that does not depend on h_n^*, $\sup_{h_n^* \leq h \leq \overline{h}_n} \tilde{\mathbb{T}}_n(h) \leq C\sqrt{\log\log(\overline{h}_n/h_n^*)}$ with probability approaching one. Thus, (9) will hold so long as $\frac{\log\log(\overline{h}_n/h_n^*)}{\log\log(\overline{h}_n/\underline{h}_n)} = \frac{\log\log h_n^{*-1}}{\log\log \underline{h}_n^{-1}} + o(1)$ can be made arbitrarily small by making K large, which we can do since $\log\log h_n^{*-1} = \log(\log \underline{h}_n^{-1})^{1/K} = (1/K)\log\log \underline{h}_n^{-1}$.

The rest of this section uses Theorem A.1 to prove Theorem 3.1. First, we state some empirical process bounds, which will be used later in the proof.

A.2 Empirical Process Bounds

This section states some empirical process bounds used later in the proof. The proofs of these results are given in Section S1.2 of the supplemental material (see Lemmas S1.4 and S1.5). In these lemmas, the following conditions are assumed to hold for some finite constants B_f, B_k and $\overline{f}_X$. The function $f(w,h,t)$ is assumed to satisfy $|f(W_i,h,t)k(X_i/h)| \leq B_f$ for all $h \leq \overline{h}$ and $t \in T$ with probability one, and the class of functions $\{(x,w) \mapsto f(w,h,t)k(x/h) | 0 \leq h \leq \overline{h}, t \in T\}$ is contained in some larger class $\mathcal{G}$ with polynomial covering number as defined in Section S1.1 in the supplemental appendix. We assume that $k(x)$ is a bounded kernel function with support $[-A,A]$ and $|k(x)| \leq B_k < \infty$, and that X_i is a real valued random variable with density $f_X(x)$ with $f_X(x) \leq \overline{f}_X < \infty$ for all x.

Lemma A.1. *Suppose that the conditions given above hold and let $a(h) = 2\sqrt{K \log\log(1/h)}$ where K is a constant depending only on $\mathcal{G}$ given in Lemma S1.3. Then, for a constant $\varepsilon > 0$ that depends only on K, A and $\overline{f}_X$,*

$$P\left(|\mathbb{G}_n f(W_i,h,t)k(X_i/h)| \geq a(h)h^{1/2}B_f A^{1/2}\overline{f}_X^{1/2} \text{ some } (\log\log n)/(\varepsilon n) \leq h \leq \overline{h},\, t \in T\right) \leq K(\log 2)^{-2} \sum_{(2\overline{h})^{-1} \leq 2^k \leq \infty} k^{-2}.$$

Lemma A.2. *Under the conditions of Lemma A.1,*

$$\sup_{(\log\log n)/(\varepsilon n) \leq h \leq \overline{h}, t \in T} \frac{|\mathbb{G}_n f(W_i,h,t)k(X_i/h)|}{(\log\log h^{-1})^{1/2}h^{1/2}} = \mathcal{O}_P(1)$$

It will be useful to state a slight extension of these results. Suppose that $f(W_i,h,t)k(X_i/h)$ converges to zero as $h \to 0$. In particular, suppose that, for some bounded function $\ell(h)$,

$$f(W_i,h,t)k(X_i/h) \leq \ell(h) \tag{12}$$

with probability one. Then, applying the above results with $f(W_i,h,t)$ replaced by $f(W_i,h,t)/\ell(h)$, we have

$$\sup_{(\log\log n)/(\varepsilon n) \leq h \leq \overline{h}, t \in T} \frac{|\mathbb{G}_n f(W_i,h,t)k(X_i/h)|}{(\log\log h^{-1})^{1/2}h^{1/2}\ell(h)} = \mathcal{O}_P(1).$$

Thus,

$$\sup_{\underline{h}_n \le h \le \overline{h}_n, t \in T} \frac{|\mathbb{G}_n f(W_i, h, t) k(X_i/h)|}{h^{1/2}} = \mathcal{O}_P\left(\sup_{\underline{h}_n \le h \le \overline{h}_n} (\log\log h^{-1})^{1/2} \ell(h) \right)$$
$$= \mathcal{O}_P\left((\log\log \overline{h}_n^{-1})^{1/2} \ell(\overline{h}_n) \right),$$

where the second equality holds if $(\log\log h^{-1})^{1/2}\ell(h)$ is nondecreasing in h.

A.3 Replacing $\psi(W_i, h)$ with $\tilde{Y}_i$

This section shows that (9) holds for $\tilde{\mathbb{T}}_n(h) = \sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$, and that (8) holds for $\mathbb{T}_n(h) = \frac{1}{\sqrt{nh}} \sum_{i=1}^n \tilde{Y}_i k(X_i/h)$.

The following lemma proves (9) for $\sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$.

Lemma A.3. *Suppose that the classes of functions $w \mapsto \psi(w, h)$ and $x \mapsto k(x/h)$ have polynomial uniform covering numbers, $\psi(w, h)k(x/h)$ is bounded, X_i has a bounded density and that k is a bounded kernel function with support $[-A, A]$.*

Let h_n^* be defined as above for some constant K and let $\overline{h}_n$ be a bounded sequence $\overline{h}_n \ge h_n^*$. Then, if K is large enough, (9) will hold for $\tilde{\mathbb{T}}_n(h) = \frac{1}{\sqrt{h}}\mathbb{G}_n \psi(W_i, h)k(X_i/h)$. Thus, under Assumption 3.1, (9) will hold for $\tilde{\mathbb{T}}_n(h) = \sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$.

Proof. Let C be such that, for any $\tilde{h}$,

$$P\left(\sup_{\underline{h}_n \le h \le \tilde{h}} \frac{1}{\sqrt{\log\log h^{-1}}\sqrt{h}} \mathbb{G}_n \psi(W_i, h) k(X_i/h) > C \right) \le C \sum_{(2\tilde{h})^{-1} \le k \le \infty} k^{-2}$$

(this can be done by Lemma A.1). Given $\delta > 0$, let $\tilde{h}_\delta$ be such that the right hand side of this display is less than δ, and let $\tilde{C}_\delta$ be such that $\sup_{\tilde{h}_\delta \le h \le \overline{h}_n} \frac{1}{\sqrt{h}}\mathbb{G}_n \psi(W_i, h)k(X_i/h) \le \tilde{C}_\delta$ with probability at least $1 - \delta$. Then, with probability at least $1 - 2\delta$,

$$\sup_{h_n^* \le h \le \overline{h}_n} \frac{1}{\sqrt{h}} \mathbb{G}_n \psi(W_i, h) k(X_i/h)$$
$$\le \max\left\{ \sqrt{2\log\log h_n^{*-1}} \sup_{h_n^* \le h \le \tilde{h}_\delta} \frac{\mathbb{G}_n \psi(W_i, h) k(X_i/h)}{\sqrt{\log\log h^{-1}}\sqrt{h}}, \sup_{\tilde{h}_\delta \le h \le \overline{h}_n} \frac{1}{\sqrt{h}} \mathbb{G}_n \psi(W_i, h) k(X_i/h) \right\}$$
$$\le C \cdot \sqrt{2\log\log h_n^{*-1}} + \tilde{C}_\delta = C \cdot \sqrt{(2/K)\log\log \underline{h}_n^{-1}} + \tilde{C}_\delta \le C \cdot \sqrt{(3/K)\log\log \underline{h}_n^{-1}}$$

for large enough n. Since δ was arbitrary, it follows that $\frac{\sup_{h_n^* \le h \le \overline{h}_n} \frac{1}{\sqrt{h}} \mathbb{G}_n \psi(W_i,h) k(X_i/h)}{\sqrt{2 \log\log \underline{h}_n^{-1}}} \le C\sqrt{3/(2K)}$ with probability approaching one. Since this can be made less than $1-\varepsilon$ by making K large (and since $\limsup_n \sqrt{2\log\log \underline{h}_n^{-1}} / \sqrt{2\log\log(\overline{h}_n/\underline{h}_n)} \le 1$), the result follows. □

We now show that (8) holds for $\mathbb{T}_n(h) = \frac{1}{\sqrt{nh}} \sum_{i=1}^n \tilde{Y}_i k(X_i/h)$ and $\tilde{\mathbb{T}}_n(h) = \sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$. By Assumption 3.1, it suffices to show this for $\tilde{\mathbb{T}}_n(h) = \frac{1}{\sqrt{h}} \mathbb{G}_n \psi(W_i,h) k(X_i/h)$. To this end, we first prove a general result where $\mathbb{T}_n(h)$ and $\tilde{\mathbb{T}}_n(h)$ are given by $\frac{1}{\sqrt{nh}} \sum_{i=1}^n \psi(W_i,h) k(X_i/h)$ and $\frac{1}{\sqrt{nh}} \sum_{i=1}^n \tilde{\psi}(W_i,h) k(X_i/h)$, and then verify these conditions for $\tilde{\psi}(W_i,h)$ given by $\tilde{Y}_i$.

Lemma A.4. *Suppose that the conditions of Lemma A.3 hold as stated and with ψ replaced by $\tilde{\psi}$. If $|[\tilde{\psi}(W_i,h) - \psi(W_i,h)]k(X_i/h)| \le \ell(h)$ for some function $\ell(h)$ with $\lim_{h\to 0} \ell(h) \log\log h^{-1} = 0$. Then, for h_n^* given in (10),*

$$\sqrt{\log\log(h_n^*/\underline{h}_n)} \sup_{\underline{h}_n \le h \le \overline{h}_n^*} \left| \frac{1}{\sqrt{h}} \mathbb{G}_n \psi(W_i,h) k(X_i/h) - \frac{1}{\sqrt{h}} \mathbb{G}_n \tilde{\psi}(W_i,h) k(X_i/h) \right| \xrightarrow{p} 0.$$

Proof. By Lemma A.2 applied to $[\tilde{\psi}(W_i,h) - \psi(W_i,h)]k(X_i/h)/\ell(h)$, we have

$$\sup_{\underline{h}_n \le h \le h_n^*} \left| \frac{1}{\sqrt{h}} \mathbb{G}_n \psi(W_i,h) k(X_i/h) - \frac{1}{\sqrt{h}} \mathbb{G}_n \tilde{\psi}(W_i,h) k(X_i/h) \right| = \mathcal{O}_P\left(\sup_{\underline{h}_n \le h \le h_n^*} \ell(h) \sqrt{\log\log h^{-1}} \right).$$

Since $\lim_{h\to 0} \ell(h) \log\log h^{-1} = 0$, we can assume without loss of generality that $\ell(h)\log\log h^{-1}$ is nondecreasing and that, therefore, $\ell(h)\sqrt{\log\log h^{-1}}$ is nondecreasing. Thus,

$$\begin{aligned}
&\sqrt{\log\log(h_n^*/\underline{h}_n)} \sup_{\underline{h}_n \le h \le \overline{h}_n^*} \left| \frac{1}{\sqrt{h}} \mathbb{G}_n \psi(W_i,h) k(X_i/h) - \frac{1}{\sqrt{h}} \mathbb{G}_n \tilde{\psi}(W_i,h) k(X_i/h) \right| \\
&= \mathcal{O}_P\left(\ell(h_n^*) \sqrt{\log\log h_n^{*-1}} \sqrt{\log\log(h_n^*/\underline{h}_n)} \right) \\
&= \mathcal{O}_P\left(\ell(h_n^*) \log\log h_n^{*-1} \frac{\sqrt{\log\log(h_n^*/\underline{h}_n)}}{\sqrt{\log\log h_n^{*-1}}} \right).
\end{aligned}$$

The result follows since $\ell(h_n^*)\log\log h_n^{*-1} \to 0$ and $\frac{\sqrt{\log\log(h_n^*/\underline{h}_n)}}{\sqrt{\log\log h_n^{*-1}}} \le \frac{\sqrt{\log\log \underline{h}_n^{-1}}}{\sqrt{\log\log h_n^{*-1}}} = \frac{\sqrt{\log\log \underline{h}_n^{-1}}}{\sqrt{(1/K)\log\log \underline{h}_n^{-1}}} = \sqrt{K}$. □

We now show that the conditions of Lemma A.4 hold for $\tilde{\psi}(W_i,h)$ given by $\tilde{Y}_i$ under the

conditions of Theorem 3.1.

Lemma A.5. *Under the conditions of Theorem 3.1,* $|[\psi(W_i,h) - \tilde{Y}_i]k(X_i/h)| \leq \ell(h)$ *for some function* $\ell(h)$ *with* $\lim_{h\to 0} \ell(h) \log\log h^{-1} = 0$.

Proof. Let $\tilde{\sigma}^2(x) = var[\psi(W_i,0)||X_i| = x]$, $a(x) = [\tilde{\sigma}^2(x) f_{|X|}(x) \int_0^\infty k(u)^2\,du]^{-1/2}$, and $\tilde{\mu}(x) = E[\psi(W_i,0)||X_i| = x]$. We have

$$
\begin{aligned}
&[\psi(W_i,h) - \tilde{Y}_i]k(X_i/h) \\
&= [\psi(W_i,h) - \psi(W_i,0)]k(X_i/h) + \{\psi(W_i,0) - a(|X_i|)\,[\psi(W_i,0) - \tilde{\mu}(|X_i|)]\}\,k(X_i/h) \\
&= [\psi(W_i,h) - \psi(W_i,0)]k(X_i/h) + \psi(W_i,0)[1 - a(|X_i|)]k(X_i/h) + a(|X_i|)\tilde{\mu}(|X_i|)k(X_i/h)
\end{aligned}
$$

The first term is bounded by a function $\ell(h)$ with $\lim_{h\to 0} \ell(h) \log\log h^{-1} = 0$ by assumption.

The second term is bounded by a constant times $\sup_{0\leq x\leq Ah} |1 - a(x)|$, and the last term is bounded by a constant times $\sup_{0\leq x\leq Ah} |\tilde{\mu}(x)|$ once $a(x)$ is shown to be bounded. To deal with these terms, note that $a(0) = 1$ and $\tilde{\mu}(0) = 0$ by construction (this is shown below in Lemma A.6). Thus,

$$
\begin{aligned}
&\sup_{0\leq x\leq Ah} |1 - a(x)| = \sup_{0\leq x\leq Ah} |a(0) - a(x)| \\
&= \left[\int_0^\infty k(u)^2\,du\right]^{-1/2} \sup_{0\leq x\leq Ah} \left|[\tilde{\sigma}^2(0) f_{|X|}(0)]^{-1/2} - [\tilde{\sigma}^2(x) f_{|X|}(x)]^{-1/2}\right|.
\end{aligned}
$$

By continuous differentiability of $(s,t) \mapsto (st)^{-1/2}$ at $s = \tilde{\sigma}^2(0)$ and $t = f_{|X|}(0)$ along with Assumption 3.2, this is bounded by a constant times $\sup_{0\leq x\leq Ah} \ell(x)$ for a function $\ell(h)$ with $\ell(h) \log\log h^{-1} \to 0$ as $h \to 0$. Since $[\log\log h^{-1}] \sup_{0\leq x\leq Ah} \ell(x) \leq \sup_{0\leq x\leq Ah} [\log\log x^{-1}]\ell(x)$, this bound satisfies the required conditions. The last term is bounded by a constant times $\sup_{0\leq x\leq Ah} |\tilde{\mu}(x) - \tilde{\mu}(0)|$, and this term is bounded by a function $\ell(h)$ with $\ell(h) \log\log h^{-1} \to 0$ as $h \to 0$ by assumption. □

The following lemma is used in the proof of Lemma A.5.

Lemma A.6. *Under the conditions of Theorem 3.1,* $a(0) = 1$ *and* $\tilde{\mu}(0) = 0$, *where* $a(x)$ *and* $\tilde{\mu}(x)$ *are defined in Lemma A.5.*

Proof. Note that

$$0 = \frac{1}{h}E\psi(W_i,h)k(X_i/h) = \frac{1}{h}E\psi(W_i,0)k(X_i/h) + \frac{1}{h}E[\psi(W_i,h)-\psi(W_i,0)]k(X_i/h)$$
$$= \tilde{\mu}(0)\frac{1}{h}Ek(X_i/h) + \frac{1}{h}E(\tilde{\mu}(X_i)-\tilde{\mu}(0))k(X_i/h) + \frac{1}{h}E[\psi(W_i,h)-\psi(W_i,0)]k(X_i/h).$$

As $h \to 0$, $\frac{1}{h}Ek(X_i/h) \to f_{|X|}(0)\int_0^\infty k(u)\,du > 0$, $\frac{1}{h}E(\tilde{\mu}(x)-\tilde{\mu}(0))k(X_i/h) \to 0$ and $\frac{1}{h}E[\psi(W_i,h) - \psi(W_i,0)]k(X_i/h) \to 0$, so taking limits in the above display shows that $\tilde{\mu}(0) = 0$. Similarly,

$$1 = \frac{1}{h}var(\psi(W_i,h)k(X_i/h))$$
$$= \frac{1}{h}var(\psi(W_i,0)k(X_i/h)) + \frac{1}{h}var([\psi(W_i,h)-\psi(W_i,0)]k(X_i/h))$$
$$+ \frac{2}{h}cov([\psi(W_i,h)-\psi(W_i,0)]k(X_i/h), \psi(W_i,0)k(X_i/h)).$$

As $h \to 0$, the last two terms converge to zero, since they are bounded by $\ell(h)$ or $\ell(h)^2$ times terms of the form $Ek(X_i/h)/h$ and $Ek(X_i/h)^2/h$. The first term is

$$\frac{1}{h}\int_0^\infty \tilde{\sigma}^2(x)k(x/h)^2 f_{|X|}(x)\,dx + \frac{1}{h}var(\mu(|X_i|)k(|X_i|/h)),$$

which converges to $\tilde{\sigma}^2(0)f_{|X|}(0)\int_0^\infty k(u)^2\,du$ as $h \to 0$ (the last term is bounded by a constant times $\ell(h)^2$). Thus, $\tilde{\sigma}^2(0) = \left(f_{|X|}(0)\int_0^\infty k(u)^2\,du\right)^{-1}$ so that, with $a(x)$ defined above, $a(0) = 1$. □

A.4 Gaussian Approximation

This section states shows that $\frac{1}{\sqrt{h}}\mathbb{G}_n\tilde{Y}_i k(X_i/h) = \frac{1}{\sqrt{nh}}\sum_{i=1}^n \tilde{Y}_i k(X_i/h)$ is approximated by a Gaussian process with the same covariance kernel. The proof of the result is given in Section S1.3 of the supplemental appendix.

We consider a general setup with $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^n$ iid, with $\tilde{X}_i \geq 0$ a.s. such that $\tilde{X}_i$ has a density $f_{\tilde{X}}(x)$ on $[0,\overline{x}]$ for some $\overline{x} \geq 0$, with $f_{\tilde{X}}(x)$ bounded away from zero and infinity on this set. We assume that $\tilde{Y}_i$ is bounded almost surely, with $E(\tilde{Y}_i|\tilde{X}_i) = 0$ and $var(\tilde{Y}_i|\tilde{X}_i = x) = f_{\tilde{X}}(x)^{-1}$. We assume that the kernel function k has finite support $[0,A]$ and is differentiable on its support with bounded derivative. For ease of notation, we assume in this section that $\int k(u)^2\,du = 1$. The result applies to our setup with $\tilde{Y}_i$ given in (11) and $\tilde{X}_i$ given by $|X_i|$.

Let

$$\hat{\mathbb{H}}_n(h) = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \tilde{Y}_i k(\tilde{X}_i/h).$$

Theorem A.2. *Under the conditions above, there exists, for each n, a process $\mathbb{H}_n(h)$ such that, conditional on $(\tilde{X}_1, \ldots, \tilde{X}_n)$, $\mathbb{H}_n$ is a Gaussian process with covariance kernel*

$$cov\left(\mathbb{H}_n(h), \mathbb{H}_n(h')\right) = \frac{1}{\sqrt{hh'}} \int k(x/h)k(x/h')\,dx$$

and

$$\sup_{\underline{h}_n \le h \le \overline{x}/A} \left|\hat{\mathbb{H}}_n(h) - \mathbb{H}_n(h)\right| = \mathcal{O}_P\left((n\underline{h}_n)^{-1/4}[\log(n\underline{h}_n)]^{1/2}\right)$$

for any sequence $\underline{h}_n$ with $n\underline{h}_n / \log\log \underline{h}_n^{-1} \to \infty$.

For our purposes, we need $(n\underline{h}_n)^{-1/4}[\log(n\underline{h}_n)]^{1/2} \cdot (\log\log \underline{h}_n^{-1})^{1/2} \to 0$, so that the rate in the above theorem is $o_P(1/\sqrt{\log\log \underline{h}_n})$. For this, the condition that $n\underline{h}_n/[(\log\log n)(\log\log\log n)]^2 \to \infty$ given in the conditions of Theorem 3.1, is sufficient, since this implies, for some $a_n \to \infty$, $(n\underline{h}_n)^{1/4} \ge a_n(\log\log n)^{1/2}(\log\log\log n)^{1/2}$ and this implies, for large enough n,

$$\begin{aligned}
(n\underline{h}_n)^{-1/4}[\log(n\underline{h}_n)]^{1/2} &\le a_n^{-1} \frac{\{\log[a_n(\log\log n)^{1/2}(\log\log\log n)^{1/2}]^4\}^{1/2}}{(\log\log n)^{-1/2}(\log\log\log n)^{-1/2}} \\
&= a_n^{-1} \frac{\{4[\log a_n + (1/2)\log\log\log n + (1/2)\log\log\log\log n]\}^{1/2}}{(\log\log n)^{-1/2}(\log\log\log n)^{-1/2}} \\
&\le 2a_n^{-1}(\log a_n + 1)^{1/2}(\log\log n)^{-1/2}.
\end{aligned}$$

A.5 Limit Theorem for the Gaussian Approximation

This section derives the limiting distribution of the approximating Gaussian process as $\overline{h}_n/\underline{h}_n$ increases.

Theorem A.3. *Let $\mathbb{H}(h)$ be a Gaussian process with mean zero and covariance kernel*

$$cov\left(\mathbb{H}(h), \mathbb{H}(h')\right) = \frac{\int k(u/h)k(u/h')\,du}{\sqrt{hh'}\int k(u)^2\,du} = \sqrt{\frac{h'}{h}}\frac{\int k(u(h'/h))k(u)\,du}{\int k(u)^2\,du},$$

where k is a bounded symmetric kernel with bounded derivative and support $[-A, A]$. Let $c_1 = \frac{Ak(A)^2}{\sqrt{\pi}\int k(u)^2\,du}$, $c_2 = \frac{1}{2\pi}\sqrt{\frac{\int\left[k'(u)u + \frac{1}{2}k(u)\right]^2 du}{\int k(u)^2\,du}}$, and let $b(t) = \log c_2$ if $k(A) = 0$ and $b(t) = \log c_1 + \frac{1}{2}\log t$ if $k(A) \ne 0$.

Let $\underline{h}_n$ and $\overline{h}_n$ be sequences with $\overline{h}_n/\underline{h}_n \to \infty$. Then

$$\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}\left(\sup_{\underline{h}_n \le h \le \overline{h}_n} \mathbb{H}(h) - \sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}\right) - b(\log\log(\overline{h}_n/\underline{h}_n)) \xrightarrow{d} Z$$

and

$$\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}\left(\sup_{\underline{h}_n \le h \le \overline{h}_n} |\mathbb{H}(h)| - \sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}\right) - b(\log\log(\overline{h}_n/\underline{h}_n)) \xrightarrow{d} Z \vee Z'$$

where Z and Z' are independent extreme value random variables.

Proof. We use Theorem 12.3.5 of Leadbetter, Lindgren, and Rootzen (1983) applied to the process $\mathbb{X}(t) = \mathbb{H}(e^t)$, which is stationary, with, in the case where $k(A) \neq 0$, $\alpha = 1$ and $C = \frac{Ak(A)^2}{\int k(u)^2\,du}$ and, in the case where $k(A) = 0$, $\alpha = 2$ and $C = \frac{\int\left[k'(u)u + \frac{1}{2}k(u)\,du\right]^2 du}{2\int k(u)^2\,du}$. The calculations and verification of the conditions for this theorem follow from elementary calculus and are given in Section S1.4 of the supplemental appendix. □

A.6 Proof of Theorem 3.1

We are now ready to prove Theorem 3.1.

proof of Theorem 3.1. By arguing along subsequences, we can assume without loss of generality that $\overline{h}_n/\underline{h}_n \to h^*$ for some $h^* \in [0,\infty)$ or $h^* = \infty$. In the first case,

$$\sup_{\underline{h}_n \le h \le \overline{h}_n} \frac{\sqrt{nh}(\hat{\theta}(h) - \theta(h))}{\hat{\sigma}(h)} = \sup_{1 \le t \le \overline{h}_n/\underline{h}_n} \mathbb{H}_n(t\underline{h}_n) + r_n$$

where $r_n \xrightarrow{p} 0$ and $\mathbb{H}_n(h)$ is, conditional on $\{|X_i|\}_{i=1}^n$, a Gaussian process with the same distribution as $\mathbb{H}(h)$. Since multiplying h by a constant does not change the distribution of $\mathbb{H}(h)$, it follows that

$$\sup_{1 \le t \le \overline{h}_n/\underline{h}_n} \mathbb{H}_n(t\underline{h}_n) \overset{d}{=} \sup_{1 \le h \le \overline{h}_n/\underline{h}_n} \mathbb{H}(h) \xrightarrow{d} \sup_{1 \le h \le h^*} \mathbb{H}(h),$$

where the last step follows from stochastic equicontinuity of $\mathbb{H}(h)$ on compact intervals. The result then follows by continuity of the distribution of $\sup_{1 \le h \le h^*} \mathbb{H}(h)$ at $c_{1-\alpha}(h^*, k)$ (which follows

from Proposition 3.2 in Pitt and Tran, 1979), and a similar argument applies in the two-sided case.

In the case where $\overline{h}_n/\underline{h}_n \to \infty$, let h_n^* be given by (10) for some K which will be chosen large enough to satisfy conditions given below. We can assume without loss of generality that either $\overline{h}_n > h_n^*$ for all n large enough or that $\overline{h}_n \leq h_n^*$ for all n large enough (again, by arguing along subsequences). In the former case, we apply Lemma A.3 to show that condition (9) holds for $\sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$ (or $\sqrt{nh}|\hat{\theta}(h) - \theta(h)|/\hat{\sigma}(h)$ in the two-sided case) so long as K is chosen large enough in the definition of h_n^*. Thus, by Theorem A.1, it suffices to consider the latter case where $\overline{h}_n \leq h_n^*$.

By Lemmas A.4 and A.5, (8) holds for $\sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$ and $\frac{1}{\sqrt{nh}}\sum_{i=1}^n \tilde{Y}_i k(X_i/h)$. It therefore follows from Theorem A.1 that it suffices to consider $\frac{1}{\sqrt{nh}}\sum_{i=1}^n \tilde{Y}_i k(X_i/h)$. By Theorem A.2, this can be replaced by $\mathbb{H}_n(h)$, where $\mathbb{H}_n(h)$ is the Gaussian process conditional on $\{|X_i|\}_{i=1}^n$ defined in the proof of that theorem. By Theorem A.3,

$$\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}\left(\sup_{\underline{h}_n \leq h \leq \overline{h}_n} \mathbb{H}_n(h) - \sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}\right) - b(\log\log(\overline{h}_n/\underline{h}_n)) \xrightarrow{d} Z.$$

Thus, by Theorems A.1 and A.2, the same holds with $\mathbb{H}_n(h)$ replaced by $\sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$. Since $c_{1-\alpha}(\overline{h}_n/\underline{h}_n, k)$ is the $1-\alpha$ quantile of a distribution that converges in distribution to Z by Theorem A.2, and since the cdf of Z is continuous, the result follows for the one-sided case. The two-sided case follows from the same arguments with $\sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$ replaced by $\sqrt{nh}|\hat{\theta}(h) - \theta(h)|/\hat{\sigma}(h)$, etc. The last two displays in the statement of the theorem follow directly from these extreme value limits. □

B Specification Searches and Sensitivity Analysis

This section discusses the use of uniform-in-the-tuning-parameter confidence bands in sensitivity analysis and compares them to pointwise-in-the-tuning-parameter confidence bands. The points made here apply to any sensitivity analysis of some parameter $\theta(h)$ to a tuning parameter h (e.g., h may be the subset of included covariates, as in Leamer, 1983).

Consider a setup where an estimate $\hat{\theta}(h)$ depends on a tuning parameter h and, for a given h, is an approximately unbiased estimate of a parameter $\theta(h)$. Suppose that there is some "true"

parameter θ^*, and different readers may disagree on how $\theta(h)$ relates to θ^* as h varies. We have the option of reporting pointwise-in-h confidence sets $\mathcal{C}_{\text{pointwise}}(h)$ satisfying

$$P\left(\theta(h) \in \mathcal{C}_{\text{pointwise}}(h)\right) = 1 - \alpha \quad \text{for all } h \in \mathcal{H}$$

or uniform-in-h confidence sets $\mathcal{C}_{\text{uniform}}(h)$ satisfying

$$P\left(\theta(h) \in \mathcal{C}_{\text{uniform}}(h) \text{ all } h \in \mathcal{H}\right) = 1 - \alpha.$$

If each reader has in mind a particular h such that $\hat{\theta}(h)$ and $\mathcal{C}_{\text{pointwise}}(h)$ are best, in some sense, for estimating and performing inference on θ^*, and, if given access to the original data, would not perform any other analysis, then the researcher can simply report $\hat{\theta}(h)$ and $\mathcal{C}_{\text{pointwise}}(h)$ for a range of values of h. Then, individual readers can simply choose which $\mathcal{C}_{\text{pointwise}}(h)$ to use and perform the analysis they would have performed with the data and their prior belief about the best h. The confidence region $\mathcal{C}_{\text{pointwise}}(h)$ selected by the reader (which the reader would have always selected regardless of the data) will have the correct coverage for $\theta(h)$ for the given h, and this will be satisfactory for the given reader.

If, however, the researcher has some liberty in choosing which $\hat{\theta}(h)$ to report and/or emphasize (e.g. by reporting some results in the abstract or main text and others in an appendix), reporting $\mathcal{C}_{\text{pointwise}}(h)$ can lead to undercoverage, if one interprets coverage as "coverage conditional on being reported/emphasized in the main text." In this setting, reporting $\mathcal{C}_{\text{uniform}}(h)$ solves the problem of undercoverage of $\theta(h)$, so long as the set $\mathcal{H}$ includes all values of h considered by the researcher in choosing which $\hat{\theta}(h)$ to report. This becomes particularly important when readers are less informed about the subject matter or details of the data than the researcher, since, in this case, readers may defer to the researcher on the choice of h. Indeed, even if they were to go into the appendix, it may not be clear what patterns they should look for in the other estimates that would go against the results in the main text.

To get at these ideas in another way, let us consider some hypothesis testing problems that a

researcher might have in mind in performing a sensitivity analysis:

$$
\begin{aligned}
&H_{0,a}\colon \theta(h) \leq 0 \quad \text{some } h \in \mathcal{H},\\
&H_{0,b}\colon \theta(h) \leq 0 \quad \text{for all } h \in \mathcal{H},\\
&H_{0,c}\colon \theta(h) \text{ has the same sign for all all } h \in \mathcal{H}.
\end{aligned}
$$

One may consider formalizing the notion of "concluding that θ is greater than zero in a robust sense" in one of the following ways:

$$\text{rejecting } H_{0,a} \text{ (and therefore also accepting } H_{0,c} \text{ in the sense of rejecting its complement)} \quad (13)$$

or

$$\text{rejecting } H_{0,b} \text{ and failing to reject } H_{0,c}. \quad (14)$$

Clearly, (13) is a more stringent requirement than (14). Note that rejecting only when $\mathcal{C}_{\text{pointwise}}(h) \subseteq (0,\infty)$ for all h provides a valid test of $H_{0,a}$ since, under $H_{0,a}$, $\theta(h^*) \leq 0$ for some h^* and, for this h^*, $P\left(\mathcal{C}_{\text{pointwise}}(h) \subseteq (0,\infty) \text{ all } h\right) \leq P\left(\mathcal{C}_{\text{pointwise}}(h^*) \subseteq (0,\infty)\right) \leq P\left(\theta^* \notin \mathcal{C}_{\text{pointwise}}(h^*)\right)$.

Thus, if one takes (13) as a criterion for "concluding that θ is greater than zero in a robust sense," one can perform this test using the pointwise-in-h confidence bands. However, this approach is likely to be conservative in many practically relevant situations. In our case, where $\hat{\theta}(h)$ is a kernel based estimate with bandwidth h, the confidence interval will be very large for small h and will contain zero for these values even if $\theta(h)$ is large.

If, instead, one takes (14) as the criterion for "concluding that θ is greater than zero in a robust sense," one can perform such a test by looking at the uniform confidence band, and concluding (14) only if $\mathcal{C}_{\text{uniform}}(h) \subseteq (0,\infty)$ for some h, and $\mathcal{C}_{\text{uniform}}(h) \cap (0,\infty) \neq 0$ for all h. Note that performing this analysis with $\mathcal{C}_{\text{pointwise}}(h)$ does not provide a test of $H_{0,c}$ with correct size, and therefore may lead the researcher to conclude that $\theta(h)$ changes signs when in fact it does not. Thus, according to this formulation, examining whether the qualitative conclusions of an analysis (such as the sign of θ) are affected by the choice of the tuning parameter requires a uniform-in-h confidence band. One can view this approach as a way of formulating a confidence statement for procedures such as those proposed by Imbens and Lemieux (2008) that examine whether the

sign of of a kernel estimator changes over a range of bandwidths.

References

Andrews, D. W. K., and M. M. A. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65(3), 497–517.

Armstrong, T. B. (2014): "Adaptive Testing of a Regression Function at a Point," *Unpublished Manuscript, Yale University, New Haven, CT*.

Bickel, P. J., and M. Rosenblatt (1973): "On some global measures of the deviations of density function estimates," *The Annals of Statistics*, pp. 1071–1095.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, forthcoming.

Card, D., C. Dobkin, and N. Maestas (2009): "Does Medicare save lives?," *Quarterly Journal of Economics*, 124(2), 597–636.

Chamberlain, G. (1986): "Asymptotic efficiency in semi-parametric models with censoring," *Journal of Econometrics*, 32(2), 189–218.

Chernozhukov, V., D. Chetverikov, and K. Kato (2013): "Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors," *The Annals of Statistics*, 41(6), 2786–2819.

——— (2014): "Anti-concentration and honest, adaptive confidence bands," *The Annals of Statistics*, 42(5), 1787–1818.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, p. asn055.

Darling, D., and P. Erdos (1956): "A limit theorem for the maximum of normalized sums of independent random variables," 23, 143–156.

DiNardo, J., and D. S. Lee (2011): "Program Evaluation and Research Designs," in *Handbook of Labor Economics*, ed. by O. Ashenfelter, and D. Card, vol. 4a, pp. 463–536. Elsevier.

Einmahl, U., and D. M. Mason (1989): "Darling-Erdos theorems for martingales," *Journal of Theoretical Probability*, 2(4), 437–460.

Fan, J. (1996): "Test of Significance Based on Wavelet Thresholding and Neyman's Truncation," *Journal of the American Statistical Association*, 91(434), 674–688.

Fan, J., and I. Gijbels (1996): *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. CRC Press.

Frölich, M. (2007): "Nonparametric IV estimation of local average treatment effects with covariates," *Journal of Econometrics*, 139(1), 35–75.

Gelman, A., and G. W. Imbens (2014): "Why high-order polynomials should not be used in regression discontinuity designs," .

Giné, E., and R. Nickl (2010): "Confidence bands in density estimation," *The Annals of Statistics*, 38(2), 1122–1170.

Hahn, J., P. Todd, and W. Van der Klaauw (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

Heckman, J. (1990): "Varieties of Selection Bias," *The American Economic Review*, 80(2), 313–318.

Heckman, J. J., S. Urzua, and E. Vytlacil (2006): "Understanding Instrumental Variables in Models with Essential Heterogeneity," *Review of Economics and Statistics*, 88(3), 389–432.

Heckman, J. J., and E. Vytlacil (2005): "Structural Equations, Treatment Effects, and Econometric Policy Evaluation1," *Econometrica*, 73(3), 669–738.

Hill, J. B. (2013): "Robust Estimation for Average Treatment Effects," *Available at SSRN 2260573*.

Horowitz, J. L., and V. G. Spokoiny (2001): "An Adaptive, Rate-Optimal Test of a Parametric Mean-Regression Model Against a Nonparametric Alternative," *Econometrica*, 69(3), 599–631.

Imbens, G., and K. Kalyanaraman (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *The Review of Economic Studies*, 79(3), 933–959.

Imbens, G. W., and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

Imbens, G. W., and T. Lemieux (2008): "Regression discontinuity designs: A guide to practice," *Journal of Econometrics*, 142(2), 615–635.

Khan, S., and E. Tamer (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78(6), 2021–2042.

Leadbetter, M. R., G. Lindgren, and H. Rootzen (1983): *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York, 1 edition edn.

Leamer, E. E. (1983): "Let's Take the Con Out of Econometrics," *The American Economic Review*, 73(1), 31–43.

Lee, D. S. (2008): "Randomized experiments from non-random selection in U.S. House elections," *Journal of Econometrics*, 142(2), 675–697.

Lee, D. S., and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48(2), 281–355.

Lehmann, E. L., and J. P. Romano (2005): *Testing statistical hypotheses*. Springer.

Lemieux, T., and K. Milligan (2008): "Incentive effects of social assistance: A regression discontinuity approach," *Journal of Econometrics*, 142(2), 807–828.

Low, M. G. (1997): "On nonparametric confidence intervals," *The Annals of Statistics*, 25(6), 2547–2554.

Ludwig, J., and D. L. Miller (2007): "Does Head Start improve children's life chances? Evidence from a regression discontinuity design," *Quarterly Journal of Economics*, 122(1), 159–208.

Miller, R., and D. Siegmund (1982): "Maximally Selected Chi Square Statistics," *Biometrics*, 38(4), 1011–1016.

Pagan, A., and A. Ullah (1999): *Nonparametric econometrics*. Cambridge University Press.

Pitt, L. D., and L. T. Tran (1979): "Local Sample Path Properties of Gaussian Fields," *The Annals of Probability*, 7(3), 477–493.

Romano, J. P., and M. Wolf (2005): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73(4), 1237–1282.

Sakhanenko, A. I. (1985): "Convergence rate in the invariance principle for non-identically distributed variables with exponential moments," *Advances in Probability Theory: Limit Theorems for Sums of Random Variables*, pp. 2–73.

Selvin, H. C., and A. Stuart (1966): "Data-Dredging Procedures in Survey Analysis," *The American Statistician*, 20(3), 20–23.

Shao, Q.-M. (1995): "Strong Approximation Theorems for Independent Random Variables and Their Applications," *Journal of Multivariate Analysis*, 52(1), 107–130.

Siegmund, D. (1985): *Sequential Analysis: Tests and Confidence Intervals*. Springer.

Spokoiny, V. G. (1996): "Adaptive hypothesis testing using wavelets," *The Annals of Statistics*, 24(6), 2477–2498.

Sun, Y. (2005): "Adaptive Estimation of the Regression Discontinuity Model," SSRN Scholarly Paper ID 739151, Social Science Research Network, Rochester, NY.

van Der Klaauw, W. (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression-Discontinuity Approach," *International Economic Review*, 43(4), 1249–1287.

White, H. (2000): "A Reality Check for Data Snooping," *Econometrica*, 68(5), 1097–1126.

	One-sided						Two-sided					
	Nadaraya-Watson			Local linear			Nadaraya-Watson			Local linear		
$\overline{h}/\underline{h}$	Unif	Tri	Epa	Unif	Tri	Epa	Unif	Tri	Epa	Unif	Tri	Epa
1.0	1.65	1.65	1.65	1.65	1.65	1.65	1.96	1.96	1.96	1.96	1.96	1.96
1.2	1.95	1.71	1.73	1.94	1.73	1.74	2.25	2.02	2.03	2.25	2.04	2.05
1.4	2.04	1.75	1.78	2.04	1.78	1.81	2.35	2.06	2.08	2.34	2.09	2.11
1.6	2.10	1.78	1.81	2.10	1.82	1.85	2.42	2.09	2.12	2.40	2.12	2.16
1.8	2.16	1.81	1.85	2.15	1.86	1.89	2.46	2.12	2.16	2.46	2.16	2.19
2.0	2.19	1.84	1.88	2.19	1.89	1.92	2.50	2.15	2.18	2.49	2.19	2.22
3.0	2.31	1.91	1.97	2.32	1.97	2.02	2.61	2.23	2.27	2.61	2.28	2.32
4.0	2.38	1.97	2.02	2.39	2.03	2.08	2.67	2.27	2.32	2.68	2.33	2.38
5.0	2.43	2.01	2.06	2.43	2.07	2.12	2.71	2.30	2.36	2.72	2.36	2.41
6.0	2.46	2.03	2.09	2.46	2.09	2.15	2.74	2.33	2.39	2.75	2.39	2.44
7.0	2.49	2.06	2.11	2.49	2.12	2.18	2.76	2.35	2.41	2.77	2.41	2.46
8.0	2.51	2.08	2.13	2.51	2.14	2.20	2.78	2.37	2.42	2.79	2.43	2.48
9.0	2.52	2.09	2.15	2.53	2.15	2.21	2.80	2.38	2.44	2.81	2.44	2.50
10.0	2.54	2.10	2.16	2.54	2.17	2.23	2.81	2.39	2.46	2.82	2.45	2.51
20.0	2.63	2.17	2.24	2.63	2.24	2.31	2.89	2.47	2.53	2.90	2.53	2.58
50.0	2.71	2.26	2.33	2.72	2.32	2.39	2.98	2.54	2.60	2.98	2.60	2.66
100.0	2.77	2.31	2.38	2.77	2.38	2.44	3.03	2.58	2.66	3.04	2.65	2.71

Table 1: Critical values for level-5% tests for the Naradaya-Watson estimator, and local linear estimator at a boundary for the Uniform (Unif, $k(u) = \frac{1}{2}I(|u| \leq 1)$), Triangular (Tri, $(1-|u|)I(|u| \leq 1)$) and Epanechnikov (Epa, $3/4(1-u^2)I(|u| \leq 1)$) kernels.
Critical values correspond to 0.95 quantiles of $\sup_{1 \leq h \leq \overline{h}/\underline{h}} \mathbb{H}(h)$ for one-sided confindence intervals and to $\sup_{1 \leq h \leq \overline{h}/\underline{h}} |\mathbb{H}(h)|$ for two-sided confidence intervals.

		Uniform Kernel			Triangular Kernel		
$(\underline{h}, \bar{h})$	$\hat{\sigma}(h)$	Pointwise	Naive	Adjusted	Pointwise	Naive	Adjusted
Local Linear regression							
$(1/2\hat{h}_{IK}, \hat{h}_{IK})$	exact	(94.7, 95.5)	86.7	95.9	(94.7, 95.5)	92.1	95.3
	EHW	(94.2, 94.7)	85.0	95.0	(93.9, 94.5)	90.5	94.0
	plugin	(96.2, 96.8)	90.1	97.0	(96.7, 97.7)	94.9	97.0
	NN	(95.3, 96.1)	87.9	96.3	(94.9, 95.9)	92.3	95.3
$(1/2\hat{h}_{IK}, 2\hat{h}_{IK})$	exact	(90.9, 95.5)	76.7	94.5	(92.8, 95.5)	87.2	94.2
	EHW	(90.4, 94.7)	74.8	93.4	(92.1, 94.5)	85.6	93.0
	plugin	(96.2, 99.2)	88.2	97.7	(96.7, 99.6)	94.6	97.7
	NN	(91.8, 96.1)	77.4	94.4	(93.4, 95.9)	88.2	94.4
$(1/4\hat{h}_{IK}, 1/2\hat{h}_{IK})$	exact	(95.2, 95.5)	87.5	96.4	(95.3, 95.5)	92.3	95.4
	EHW	(92.7, 94.4)	83.7	93.9	(91.8, 94.0)	88.5	92.3
	plugin	(96.2, 96.4)	90.6	97.4	(96.5, 96.7)	94.2	96.6
	NN	(94.6, 95.8)	87.3	95.3	(94.2, 95.3)	91.2	94.2
Local quadratic regression							
$(1/2\hat{h}_{IK}, \hat{h}_{IK})$	NN	(94.8, 95.7)	87.1	95.5	(94.5, 95.4)	91.3	94.6
$(1/2\hat{h}_{IK}, 2\hat{h}_{IK})$	NN	(87.1, 96.2)	74.9	92.9	(91.3, 96.0)	84.5	92.5
$(1/4\hat{h}_{IK}, 1/2\hat{h}_{IK})$	NN	(93.8, 94.8)	85.2	94.3	(93.2, 94.5)	89.0	93.0

Table 2: Monte Carlo study of regression discontinuity. Design 1. Empirical coverage of $\theta(h)$ for nominal 95% confidence bands around IK bandwidth. "Pointwise" refers to range of coverages of pointwise confidence intervals. "Naive" refers to the coverage of the naive confidence band that uses the unadjusted critical value equal to 1.96. "Adjusted" refers to confidence bands using adjusted critical values based on Theorem 3.1. Variance estimators are described in the text. 50,000 Monte Carlo draws (10,000 for NN-based variance estimators), 100 grid points for h.

		Uniform Kernel			Triangular Kernel		
$(\underline{h}, \bar{h})$	$\hat{\sigma}(h)$	Pointwise	Naive	Adjusted	Pointwise	Naive	Adjusted
Local Linear regression							
$(1/2\hat{h}_{IK}, \hat{h}_{IK})$	exact	(94.6, 95.2)	86.2	95.8	(94.5, 95.2)	91.5	94.9
	EHW	(91.3, 92.7)	80.3	91.9	(90.2, 92.1)	86.0	90.3
	plugin	(96.4, 96.7)	91.1	97.5	(96.9, 97.3)	94.8	96.9
	NN	(94.0, 94.6)	85.3	94.3	(93.5, 94.2)	90.2	93.2
$(1/2\hat{h}_{IK}, 2\hat{h}_{IK})$	exact	(89.2, 95.2)	76.0	93.9	(87.2, 95.2)	83.5	91.4
	EHW	(88.5, 92.8)	70.6	90.4	(86.3, 92.1)	78.8	87.7
	plugin	(96.4, 98.0)	87.9	97.2	(96.9, 99.4)	94.2	97.4
	NN	(85.0, 94.8)	73.0	91.3	(81.0, 94.2)	76.0	85.3
$(1/4\hat{h}_{IK}, 1/2\hat{h}_{IK})$	exact	(94.9, 95.2)	86.9	96.1	(95.0, 95.2)	91.6	94.9
	EHW	(85.6, 91.4)	73.7	86.5	(83.0, 90.0)	78.3	83.2
	plugin	(96.9, 97.8)	93.8	99.0	(97.1, 98.2)	95.5	97.8
	NN	(93.3, 94.3)	84.6	93.9	(92.7, 93.5)	88.7	92.1
Local quadratic regression							
$(1/2\hat{h}_{IK}, \hat{h}_{IK})$	NN	(93.5, 94.4)	84.0	93.8	(92.9, 93.9)	88.4	92.4
$(1/2\hat{h}_{IK}, 2\hat{h}_{IK})$	NN	(93.5, 96.1)	78.5	93.9	(92.9, 95.4)	85.4	92.7
$(1/4\hat{h}_{IK}, 1/2\hat{h}_{IK})$	NN	(93.5, 94.8)	84.5	93.6	(92.8, 94.0)	87.7	91.6

Table 3: Monte Carlo study of regression discontinuity. Design 2. Empirical coverage of $\theta(h)$ for nominal 95% confidence bands around IK bandwidth. "Pointwise" refers to range of coverages of pointwise confidence intervals. "Naive" refers to the coverage of the naive confidence band that uses the unadjusted critical value equal to 1.96. "Adjusted" refers to confidence bands using adjusted critical values based on Theorem 3.1. Variance estimators are described in the text. 50,000 Monte Carlo draws (10,000 for NN-based variance estimators), 100 grid points for h.

		Uniform Kernel			Triangular Kernel		
$(\underline{h}, \bar{h})$	$\hat{\sigma}(h)$	Pointwise	Naive	Adjusted	Pointwise	Naive	Adjusted
Local Linear regression							
$(1/2\hat{h}_{IK}, \hat{h}_{IK})$	exact	(74.7, 91.1)	63.6	82.5	(78.1, 91.5)	75.9	82.8
	EHW	(73.7, 89.8)	62.0	80.7	(76.7, 89.5)	74.0	80.9
	plugin	(79.1, 91.2)	66.7	83.4	(85.1, 92.1)	81.4	86.8
	NN	(77.1, 92.0)	66.9	84.2	(80.1, 91.9)	78.0	84.1
$(1/2\hat{h}_{IK}, 2\hat{h}_{IK})$	exact	(74.3, 91.1)	55.7	82.7	(77.8, 91.5)	71.1	83.0
	EHW	(73.2, 89.8)	54.3	80.9	(76.5, 89.5)	69.4	81.1
	plugin	(79.1, 97.8)	62.5	84.9	(85.1, 97.8)	80.5	88.8
	NN	(76.7, 92.0)	59.8	84.9	(79.9, 91.9)	74.2	84.9
$(1/4\hat{h}_{IK}, 1/2\hat{h}_{IK})$	exact	(91.7, 95.1)	84.0	94.7	(92.0, 95.1)	89.4	93.4
	EHW	(90.3, 92.7)	79.9	92.0	(90.0, 92.1)	85.5	90.1
	plugin	(91.7, 95.7)	85.3	94.8	(92.5, 95.9)	90.5	93.9
	NN	(92.4, 94.7)	84.5	94.4	(92.4, 94.1)	89.2	92.7
Local quadratic regression							
$(1/2\hat{h}_{IK}, \hat{h}_{IK})$	NN	(92.4, 95.3)	84.8	94.7	(91.5, 94.7)	88.4	92.8
$(1/2\hat{h}_{IK}, 2\hat{h}_{IK})$	NN	(78.6, 95.3)	62.2	86.3	(82.2, 94.7)	75.3	86.5
$(1/4\hat{h}_{IK}, 1/2\hat{h}_{IK})$	NN	(93.8, 94.7)	85.1	94.4	(93.2, 94.3)	88.8	93.0

Table 4: Monte Carlo study of regression discontinuity. Design 1. Empirical coverage of $\theta(0)$ for nominal 95% confidence bands around IK bandwidth. "Pointwise" refers to range of coverages of pointwise confidence intervals. "Naive" refers to the coverage of the naive confidence band that uses the unadjusted critical value equal to 1.96. "Adjusted" refers to confidence bands using adjusted critical values based on Theorem 3.1. Variance estimators are described in the text. 50,000 Monte Carlo draws (10,000 for NN-based variance estimators), 100 grid points for h.

		Uniform Kernel			Triangular Kernel		
$(\underline{h}, \bar{h})$	$\hat{\sigma}(h)$	Pointwise	Naive	Adjusted	Pointwise	Naive	Adjusted
Local Linear regression							
$(1/2\hat{h}_{IK}, \hat{h}_{IK})$	exact	(94.3, 95.2)	86.0	95.7	(94.1, 95.1)	91.1	94.7
	EHW	(91.3, 92.6)	80.2	91.8	(90.2, 91.9)	85.7	90.1
	plugin	(96.6, 97.1)	91.9	97.9	(97.0, 97.9)	95.3	97.3
	NN	(94.0, 94.6)	85.1	94.3	(93.5, 94.1)	89.9	93.1
$(1/2\hat{h}_{IK}, 2\hat{h}_{IK})$	exact	(60.0, 95.2)	51.8	78.0	(54.7, 95.1)	50.9	63.7
	EHW	(59.3, 92.6)	47.8	75.3	(53.9, 91.9)	47.4	60.7
	plugin	(96.6, 99.4)	90.5	98.4	(97.0, 100.0)	95.2	98.0
	NN	(63.1, 94.6)	54.0	79.8	(57.6, 94.1)	52.9	65.5
$(1/4\hat{h}_{IK}, 1/2\hat{h}_{IK})$	exact	(94.9, 95.2)	86.9	96.1	(95.0, 95.2)	91.5	94.9
	EHW	(85.6, 91.4)	73.7	86.5	(83.0, 90.1)	78.3	83.2
	plugin	(97.0, 97.8)	93.8	99.1	(97.2, 98.2)	95.5	97.9
	NN	(93.3, 94.3)	84.6	93.9	(92.7, 93.5)	88.7	92.1
Local quadratic regression							
$(1/2\hat{h}_{IK}, \hat{h}_{IK})$	NN	(93.5, 94.4)	84.1	93.8	(92.9, 93.8)	88.4	92.4
$(1/2\hat{h}_{IK}, 2\hat{h}_{IK})$	NN	(93.5, 95.8)	78.3	93.8	(92.9, 95.1)	84.6	92.4
$(1/4\hat{h}_{IK}, 1/2\hat{h}_{IK})$	NN	(93.6, 94.8)	84.4	93.6	(92.8, 94.0)	87.7	91.6

Table 5: Monte Carlo study of regression discontinuity. Design 2. Empirical coverage of $\theta(0)$ for nominal 95% confidence bands around IK bandwidth. "Pointwise" refers to range of coverages of pointwise confidence intervals. "Naive" refers to the coverage of the naive confidence band that uses the unadjusted critical value equal to 1.96. "Adjusted" refers to confidence bands using adjusted critical values based on Theorem 3.1. Variance estimators are described in the text. 50,000 Monte Carlo draws (10,000 for NN-based variance estimators), 100 grid points for h.

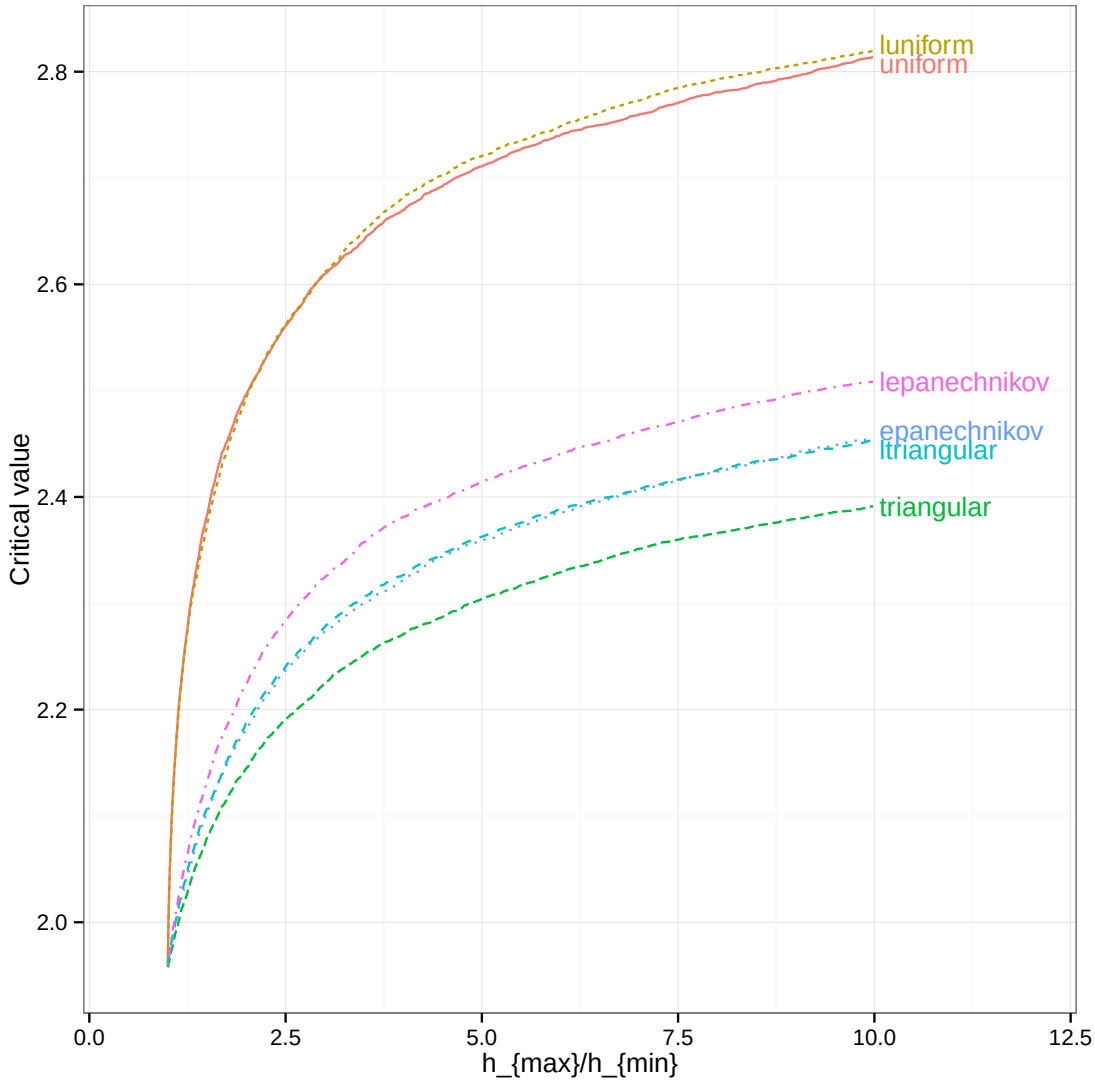


Figure 1: Two-sided 95% critical values for different kernels. luniform, ltriangular, and lepanechnikov refer to equivalent uniform and triangular kernels for local linear regression.

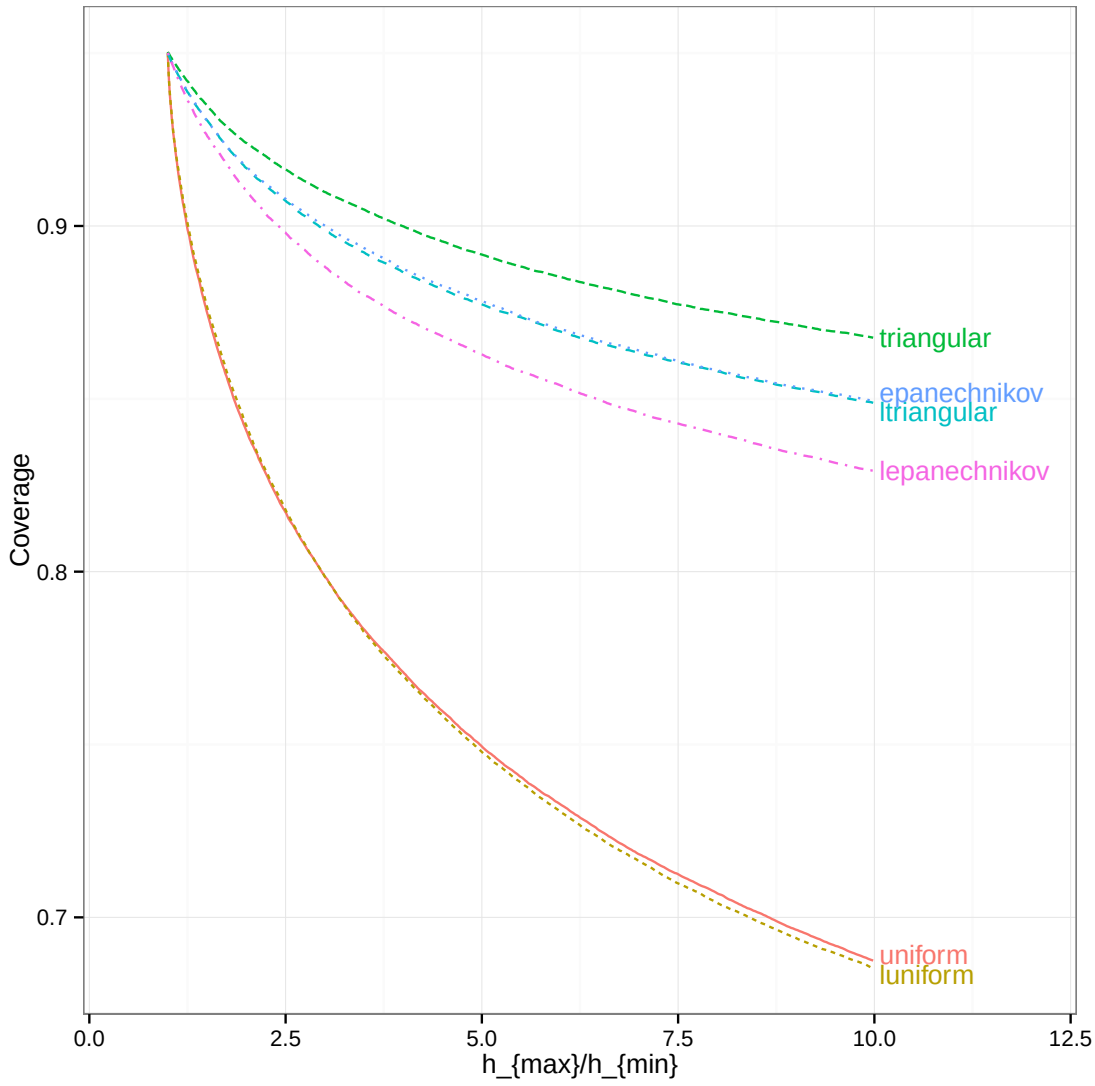


Figure 2: Coverage of unadjusted 95% confidence bands (i.e. using critical values equal to 1.96) for different kernels. luniform, ltriangular, and lepanechnikov refer to equivalent uniform and triangular kernels for local linear regression.

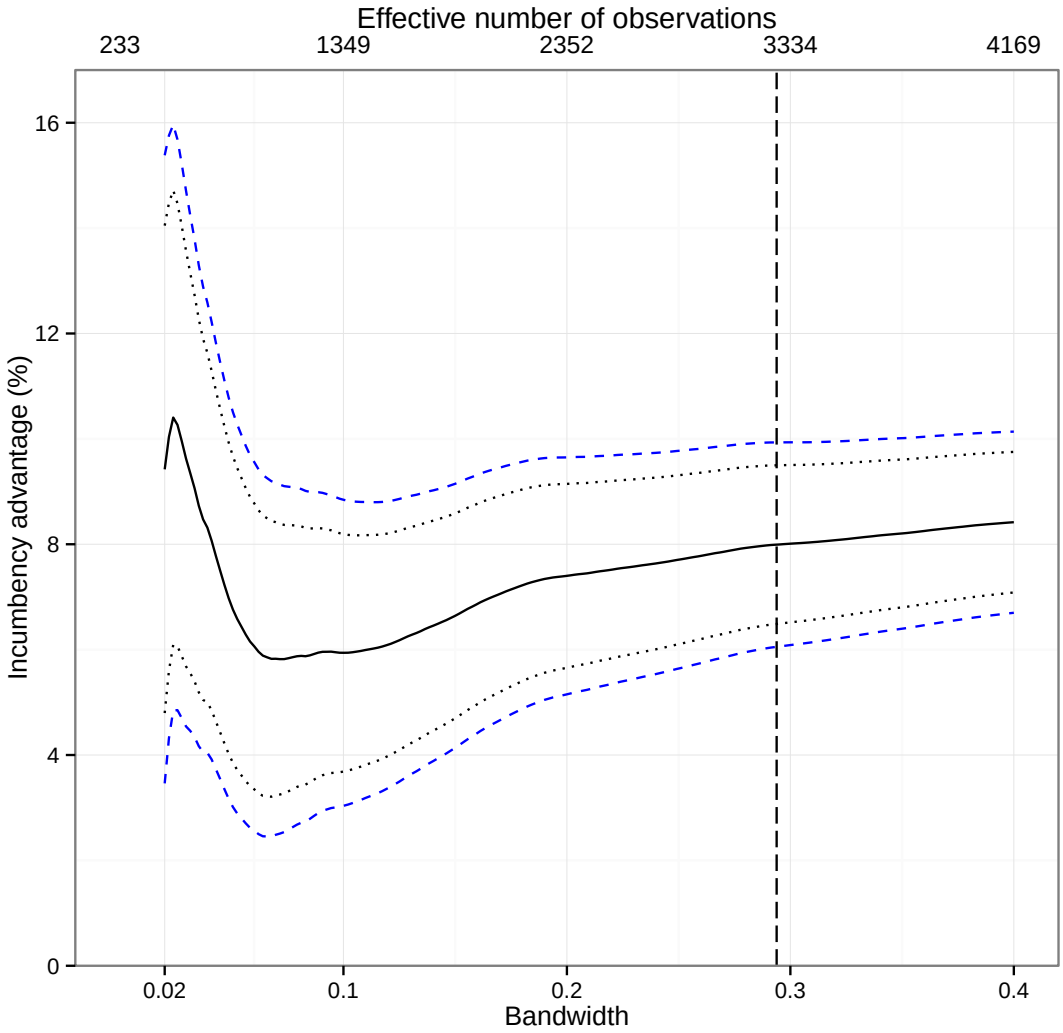


Figure 3: Effect of incumbency on percentage vote share in the next election. Data are from Lee (2008). Local linear regression with triangular kernel. Point estimate $\hat{\theta}(h)$ (solid line), pointwise (dotted), and uniform (dashed) confidence bands as function of the bandwidth h. The range of bandwidths plotted is $(0.02, 0.40)$, so that $\overline{h}/\underline{h} = 20$, and the adjusted critical value is 2.526. Vertical dashed line corresponds to estimates using Imbens and Kalyanaraman (2012) bandwidth. Effective number of observations refers to number of observations that receive non-zero kernel weight, $\sum_{i=1}^{n} 1(K(X_i/h) > 0)$.

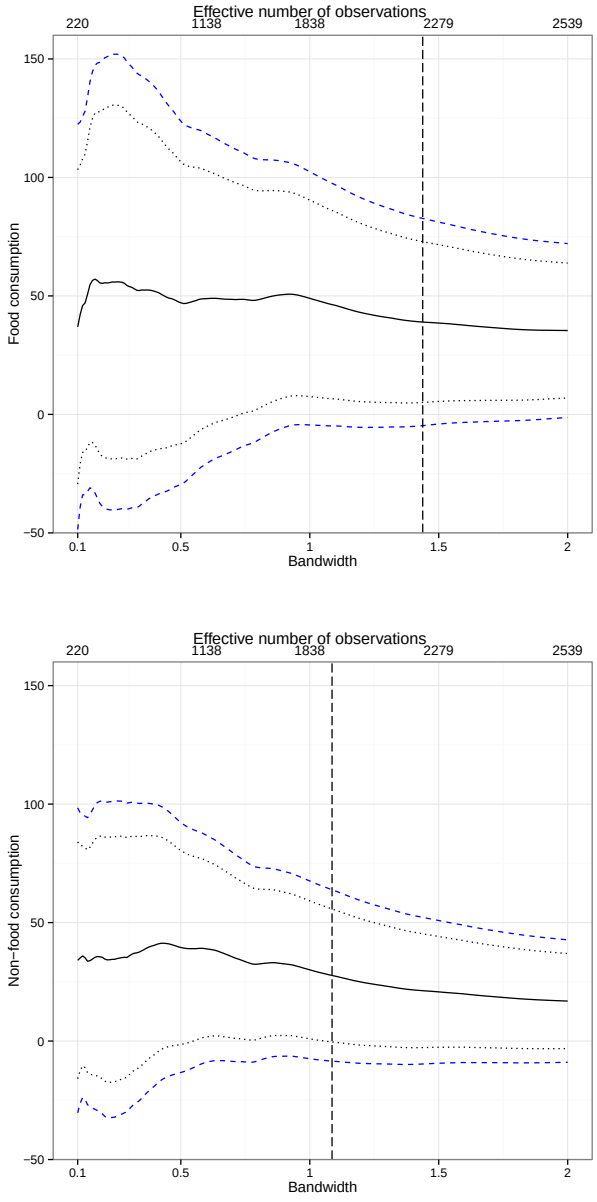


Figure 4: Effect of the Oportunidades cash transfer program on food and non-food consumption. Data are from Calonico, Cattaneo, and Titiunik (2014). Local linear regression with triangular kernel. Point estimate $\hat{\theta}(h)$ (solid line), pointwise (dotted), and uniform (dashed) confidence bands as function of the bandwidth h. The range of bandwidths plotted is $(0.1, 2)$, so that $\overline{h}/\underline{h} = 20$, and the adjusted critical value is 2.526. Vertical dashed line corresponds to estimates using Imbens and Kalyanaraman (2012) bandwidth. Effective number of observations refers to number of observations that receive non-zero kernel weight, $\sum_{i=1}^{n} 1(K(X_i/h) > 0)$.

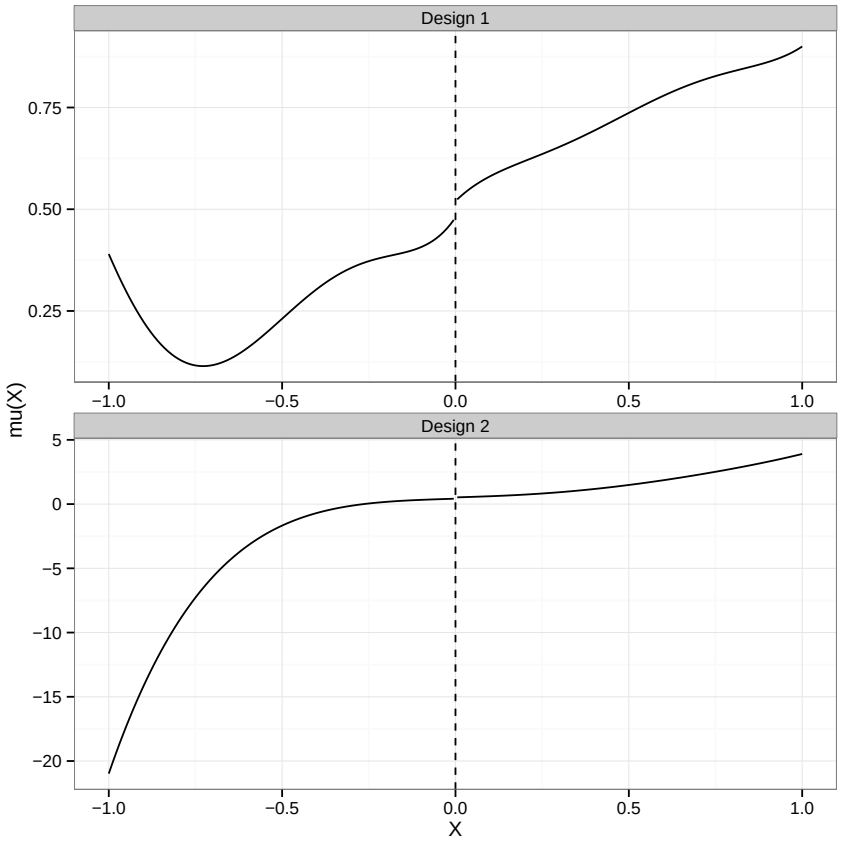


Figure 5: Monte Carlo study of regression discontinuity. Regression function $g(X)$ for designs we consider.

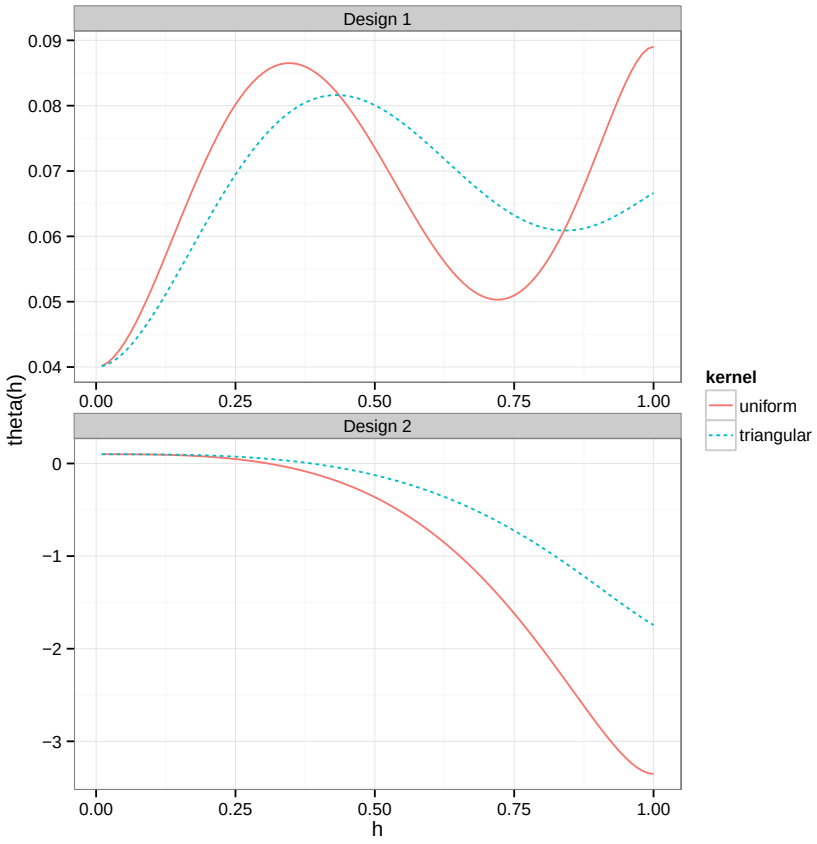


Figure 6: Monte Carlo study of regression discontinuity. Function $\theta(h)$ for designs we consider.

Name	$k^*(u)$	Order	$k(u)$
Uniform	$\frac{1}{2}I(\lvert u\rvert \leq 1)$	0	$\frac{1}{2}I(\lvert u\rvert \leq 1)$
		1	$(4-6\lvert u\rvert)I(\lvert u\rvert \leq 1)$
		2	$(9-36\lvert u\rvert+30u^2)I(\lvert u\rvert \leq 1)$
Triangular	$(1-\lvert u\rvert)_+$	0	$(1-\lvert u\rvert)_+$
		1	$6(1-2\lvert u\rvert)(1-\lvert u\rvert)_+$
		2	$12(1-5\lvert u\rvert+5u^2)(1-\lvert u\rvert)_+$
Epanechnikov	$\frac{3}{4}(1-u^2)_+$	0	$\frac{3}{4}(1-u^2)_+$
		1	$\frac{6}{19}(16-30\lvert u\rvert)(1-u^2)_+$
		2	$\frac{1}{8}(85-400\lvert u\rvert+385u^2)(1-u^2)_+$

Table 6: Definitions of kernels and equivalent kernels for regression discontinuity / estimation at a boundary. Order refers to the order of the local polynomial.