A Simple Adjustment for Bandwidth Snooping*

Timothy B. Armstrong† Yale University

Michal Kolesár‡ Princeton University

June 28, 2017

Abstract

Kernel-based estimators such as local polynomial estimators in regression discontinuity designs 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. Our results also apply to other settings involving trimmed data, such as trimming to ensure overlap in treatment effect estimation. We illustrate our approach with three empirical applications.

*We thank Joshua Angrist, Matias Cattaneo, Victor Chernozhukov, Kirill Evdokimov, Bo Honoré, Chris Sims, numerous seminar and conference participants, four anonymous referees and the editor for helpful comments and suggestions. We also thank Matias Cattaneo for sharing the Progresa dataset. All remaining errors are our own. The research of the first author was supported by National Science Foundation Grant SES-1628939. The research of the second author was supported by National Science Foundation Grant SES-1628878.

†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 sensitivity of the results to the chosen bandwidth, it is common to compute estimates and confidence intervals for several bandwidths, or plot them against a continuum of bandwidths. For example, in regression discontinuity designs—a leading application of non-parametric methods in econometrics—this approach is recommended in several surveys (Imbens and Lemieux, 2008; Lee and Lemieux, 2010; DiNardo and Lee, 2011) and implemented widely in applied work.[1] However, such practice leads to a well-known problem that if the bandwidth choice is influenced by these results, the confidence interval at the chosen 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). Undercoverage can also occur from honest attempts to report a confidence interval with good statistical properties. In settings in which one does not know the smoothness of the estimated function, it is typically *necessary* to examine multiple bandwidths to obtain confidence intervals that are optimal (see Section 4.1.2 for details and Armstrong (2015) for a formal statement). We use the term "bandwidth snooping" to refer to any situation in which a researcher considers multiple bandwidths in reporting confidence intervals.

This paper proposes a simple adjustment to account for bandwidth snooping: replace the usual critical value based on a quantile of a standard 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 kernels, and show how our adjustment can be applied in regression discontinuity designs, as well as estimation of average treatment effects under unconfoundedness after trimming, and estimation of local average treatment effects.

To explain the adjustment in a simple setting, consider the problem of estimating the conditional mean $E[Y_i \mid X_i = x]$ at a point x, which we normalize to zero. Given and i.i.d. sample

[1] For prominent examples see, for instance, van Der Klaauw (2002), Lemieux and Milligan (2008), Ludwig and Miller (2007), or Card, Dobkin, and Maestas (2009).

$\{(X_i, Y_i)\}_{i=1}^n$, a kernel k, and a bandwidth h, the Nadaraya-Watson kernel estimator is given by

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

and it is approximately unbiased for the pseudo-parameter

$$\theta(h) = \frac{E[Y_i k(X_i/h)]}{E[k(X_i/h)]}.$$

Under appropriate smoothness conditions, if we take $h \to 0$ with the sample size, $\hat{\theta}(h)$ will converge to $\theta(0) := \lim_{h\to 0} \theta(h) = E(Y_i \mid X_i = 0)$. Given a standard error $\hat{\sigma}(h)/\sqrt{nh}$, the t-statistic $\sqrt{nh}(\hat{\theta}(h) - \theta(h))/\hat{\sigma}(h)$ is approximately standard normal. 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 the bias $|\theta(h) - \theta(0)|$ is small enough relative to the standard error, such as when the bandwidth h "undersmooths", the standard confidence interval is also an approximate confidence interval for $\theta(0)$, the conditional mean at zero.

However, if the selected bandwidth $\hat{h}$ is based on examining $\hat{\theta}(h)$ over h in some interval $[\underline{h}, \overline{h}]$, the standard confidence interval around $\hat{\theta}(\hat{h})$ may undercover even if there is no bias. To address this problem, we propose confidence intervals that cover $\theta(h)$ simultaneously for all h in a given interval $[\underline{h}, \overline{h}]$ with a prespecified probability. In particular, we derive a critical value $c_{1-\alpha}$ such that as $n \to \infty$,

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

In other words, our critical values allow for a uniform confidence band for $\theta(h)$. Thus, the confidence interval for the selected bandwidth, $[\hat{\theta}(\hat{h}) \pm c_{1-\alpha}\hat{\sigma}(\hat{h})/\sqrt{n\hat{h}}]$, will achieve correct coverage of $\theta(\hat{h})$ no matter what selection rule was used to pick $\hat{h}$.

Our main contribution is to give a coverage result of the form (1) for a large class of kernel-based estimators $\hat{\theta}(h)$, as well as a similar statement showing coverage of $\theta(0)$. The latter follows under additional conditions that allow the bias to be mitigated through undersmoothing or bias-correction. These conditions are essentially the same as those needed for pointwise coverage: 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. In particular, we show how our approach can be combined with a popular bias-correction method proposed by Calonico, Cattaneo, and Titiunik (2014).

Since our confidence bands cover $\theta(h)$ under milder smoothness conditions than those needed for coverage of $\theta(0)$, they are particularly well-suited for sensitivity analysis. Suppose that a particular method for bias correction or undersmoothing implies that, in a given data set, the bias is asymptotically negligible if $h \leq 3$. If 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 are not supported by the data. Our confidence bands can thus be used to formalize certain conclusions about confidence intervals being "sensitive" to bandwidth choice.[2]

In many applications, $\theta(h)$, taken as a function indexed by the bandwidth, is an interesting parameter in its own right, in which case our confidence bands are simply confidence bands for this function. As we discuss in detail in Section 4, 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. In the latter case, h corresponds to a trimming parameter such that observations with propensity score within distance h to 0 or 1 are discarded, and $\theta(h)$ corresponds to average treatment effects for the remaining subpopulation with moderate values of the propensity score.

A key 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, it also depends on the order of the polynomial and whether the point is on the boundary of the support). In practice, researchers often report a point estimate $\hat{\theta}(\hat{h})$ and a standard error $\hat{\sigma}(\hat{h})/\sqrt{n\hat{h}}$. As long as the kernel and order of the local polynomial are also reported, a reader can use our critical values 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. Alternatively, one can assess the sensitivity of the conclusions of the analysis to bandwidth specification search by, say, computing the largest value

[2] An alternative approach to sensitivity analysis is to reject a particular null hypothesis regarding $\theta(0)$ only when one rejects the corresponding hypothesis test based on $\hat{\theta}(h)$ for all values h that one has examined. The CI for $\theta(0)$ is then given by the union of the CIs based on $\hat{\theta}(h)$ as h varies over all values that one has examined. This form of sensitivity analysis does not require a snooping correction, but is typically very conservative. See Section C of the appendix for further discussion.

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%.

We also derive analytic results showing that the critical values grow very slowly with $\overline{h}/\underline{h}$, at the rate $\sqrt{\log\log(\overline{h}/\underline{h})}$. Thus, from a practical standpoint, examining a wider range of bandwidths carries only a very small penalty (relative to examining a moderate range): while using our correction is important for obtaining correct coverage, the critical values increase quite slowly once $\overline{h}/\underline{h}$ is above 5. A Monte Carlo study in the supplemental appendix confirms that these critical values lead to uniform coverage of $\theta(h)$ that is close to the nominal level. Uniform coverage of $\theta(0)$ is also good so long as our method is combined with bias correction or undersmoothing.

We illustrate our results with three empirical applications. First, we apply our method to the regression discontinuity study of the effect of Progresa from Calonico et al. (2014), and we find that the significance of the results is sensitive to bandwidth snooping. The second empirical application is the regression discontinuity study of Lee (2008). Here, in contrast, 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. Finally, we consider an application to estimating treatment effects under unconfoundedness from Connors Jr et al. (1996). Here, we find that the results are again quite robust to the choice of trimming parameter, providing additional evidence supporting the study's conclusions.

The rest of the paper is organized as follows. Section 1.1 discusses related literature. Section 2 gives a heuristic 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 works out applications of our results to several economet-

ric models. Section 5 presents an illustration of our approach in three empirical applications. Section 6 concludes. Proofs and auxiliary results, as well as additional tables and figures and a Monte Carlo study, are given in the appendix and a supplemental appendix.[3] 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—see Lehmann and Romano (2005, Chapter 9) for an overview and early contributions, and White (2000) for an application to econometrics. On a technical level, our results borrow from the literature on Gaussian approximations to empirical processes and extreme value limits for suprema of Gaussian processes. To obtain an approximation of the kernel estimator by a Gaussian process, we use an approximation of Sakhanenko (1985). For the case $\overline{h}/\underline{h} \to \infty$, we then use extreme value theory, and our derivation is similar in spirit to Bickel and Rosenblatt (1973), who consider kernel estimation of a density under a fixed sequence of bandwidths $h = h_n$ and derive confidence bands that are uniform in the point x at which the density is evaluated. For the case with bounded $\overline{h}/\underline{h}$, classical empirical process results such as those given in van der Vaart and Wellner (1996) could be used instead of the Sakhanenko (1985) approximation, which we use in our proof since it covers both cases. In both cases, our results require the (to our knowledge novel) insight that the approximating Gaussian process is stationary when indexed by $t = \log h$, and depends only on the kernel used to compute the estimator. This leads to simple critical values that depend only on the kernel and bandwidth ratio. In other settings in which snooping does not lead to a pivotal asymptotic distribution, one could use the general bootstrap approach of Chernozhukov, Chetverikov, and Kato (2013), which allows one to obtain uniform confidence bands without obtaining an asymptotic distribution.

In addition to Bickel and Rosenblatt (1973), numerous authors have used extreme value limiting theorems for suprema of Gaussian processes to derive confidence bands for a density or conditional mean function that are uniform in the point x at which the function is evaluated,

[3]The supplemental appendix is available at `http://arxiv.org/abs/1412.0267`

with the bandwidth sequence $h = h_n$ fixed (see, among others Johnston, 1982; Härdle, 1989; Liu and Wu, 2010). In the special case where the Nadaraya-Watson estimator with uniform kernel is used, extreme value limiting results in Armstrong and Chan (2016) lead to confidence bands that are uniform in both x and h. In contrast, our case corresponds to fixing x and requiring that coverage be uniform over h.

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 (2015) uses our approach to obtain adaptive one-sided confidence intervals under a monotonicity condition (see Section 4.1.2 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 (2014a).

2 Derivation of the correction in a simple case

This section presents a heuristic derivation of the correction in the simple problem of inference on the conditional mean described in the introduction. To further simplify the exposition, 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 given by $X_i = (i+1)/(2n)$ for i odd and $X_i = -i/(2n)$ for i even. In this case, the Nadaraya-Watson kernel estimator with a uniform kernel, $k(x) = I(|x| \leq 1/2)$, reduces to

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

where, for the second equality and throughout the rest of this example, we assume that nh is an even integer for notational convenience. Consider first the problem of constructing 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 $[\underline{h}, \overline{h}]$. For a given bandwidth h, a two-sided t-statistic is given by

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

In order to guarantee correct coverage, instead of using a critical value equal to the $1-\alpha/2$ quantile of a normal distribution, we will need to use a critical value equal 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

$$\sup_{\underline{h}\leq h\leq\overline{h}}\sqrt{nh}\frac{|\hat{\theta}(h)-\theta(h)|}{\sigma} \approx \sup_{\underline{h}\leq h\leq\overline{h}}|\mathbb{B}(h)/\sqrt{h}| \stackrel{d}{=} \sup_{1\leq h\leq\overline{h}/\underline{h}}|\mathbb{B}(h)/\sqrt{h}|. \tag{2}$$

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 its quantiles can easily be tabulated (see the columns corresponding to uniform kernel in Table 1). 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 Erdös (1956) theorem. Thus, as $\overline{h}/\underline{h} \to \infty$, the critical values increase very slowly, at the rate $\sqrt{\log\log(\overline{h}/\underline{h})}$.

To guarantee that the resulting confidence interval achieves coverage for $\theta(0) = g(0)$, the conditional mean at zero, we also need to ensure that the bias $|\theta(h)-\theta(0)|$ is small relative to the standard error $\sigma/\sqrt{nh}$, uniformly over $h \in [\underline{h}, \overline{h}]$. If the conditional mean function is twice differentiable with a bounded second derivative and $\overline{h}/\underline{h}$ is bounded, a sufficient condition is that $n\overline{h}^5 \to 0$, that is, we "undersmooth".

In the next section, we show that the approximation of the distribution of the maximal t-statistic by a scaled Brownian motion in (2) still obtains even if the restrictive assumptions in this section 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 can be verified for particular applications. For applications in Section 4, we verify these conditions in Appendix B.

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. To keep statements of theoretical results simple, all of our results are pointwise in the underlying distribution (we show that, for any data generating process satisfying certain assumptions, coverage of the uniform-in-h CI converges to $1 - \alpha$). However, versions of these results in which coverage is shown to converge to $1 - \alpha$ uniformly in some class of underlying distributions could be derived from similar arguments, using uniform versions of the bounds in our assumptions. 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 $h \in [\underline{h}_n, \overline{h}_n]$.

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 $h \in [\underline{h}_n, \overline{h}_n]$, and with a $o_P(1/\sqrt{\log\log(\overline{h}_n/\underline{h}_n)})$ rate of approximation. Supplemental Appendix S1.2 provides some results that can be used to obtain this uniform bound. For example, in the case of the

Nadaraya-Watson kernel estimator $\hat{\theta}(h) = \sum_{i=1}^{n} Y_i k(X_i/h) / \sum_{i=1}^{n} k(X_i/h)$, Assumption 3.1 holds with $\psi(W_i, h) = (Y_i - \theta(h)) / \sqrt{var\{[Y_i - \theta(h)]k(X_i/h)/h\}}$. In the local polynomial case, the kernel function k corresponds to the equivalent kernel, and depends on the order of the polynomial and whether the estimated conditional quantities are at the boundary (see Section 4.1 and Supplemental Appendix S2 for details, including a discussion of how our results can be extended to cover cases in which the boundary of the support of X_i is local to 0).

We also impose some regularity conditions on k and the data generating process. In applications, 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 $[-A, A]$, bounded with a bounded, uniformly continuous first derivative on $(0, A)$, and satisfies $\int k(u)\,du \neq 0$.*

(ii) $|X_i|$ 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 $var\,(\psi(W_i, 0) \mid |X_i| = 0) > 0$, and, for some deterministic function $\ell(h)$ with $\ell(h)\log\log(h^{-1}) \to 0$ as $h \to 0$, the absolute values of the following expressions are bounded by $\ell(t)$: $f_{|X|}(t) - f_{|X|}(0)$, $E\,[\psi(W_i, 0) \mid |X_i| = t] - E\,[\psi(W_i, 0) \mid |X_i| = 0]$, $(\psi(W_i, t) - \psi(W_i, 0))k(X_i/t)$, and $var(\psi(W_i, 0) \mid |X_i| = t) - var(\psi(W_i, 0) \mid |X_i| = 0)$.

(iii) Taken as classes of functions varying over $h > 0$, $w \mapsto \psi(w, h)$ and $x \mapsto k(x/h)$ have polynomial uniform covering numbers (as defined in Appendix A).

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.). For inference on $\theta(h)$, rather than $\theta(0)$, the amount of smoothness required is very mild relative to smoothness conditions typically imposed when considering bias-variance tradeoffs. In particular, Assumption 3.2 only requires 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 $\theta(h)$, which is useful in settings where the possible lack of smoothness of $\theta(h)$ leads one to examine $\hat{\theta}(h)$ across multiple bandwidths.

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 $var\,[\psi(W_i, 0) \mid |X_i| = 0] > 0$

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 by verifying the high level conditions in Theorems A.1 and A.3 in Appendix A, which is used in proving Theorem 3.1 below. 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 correspond to the equivalent kernel, and it will depend on the order of the derivative as well as the order of the local polynomial.

Finally, note that Assumption 3.2 requires that $\psi(W_i,h)k(X_i/h)$ be bounded, which typically requires a bounded outcome variable in applications. We conjecture that this assumption could be relaxed at the expense of imposing stronger assumptions on $\underline{h}_n$ and $\overline{h}_n$ (see Section A.4).

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 $\sup_{1\leq h\leq t}|\mathbb{H}(h)|$, where $\mathbb{H}(h)$ is a mean zero Gaussian process with covariance kernel cov $(\mathbb{H}(h),\mathbb{H}(h'))=\frac{\int k(u/h)k(u/h')\,du}{\sqrt{hh'}\int k(u)^2\,du}=\sqrt{\frac{h'}{h}}\frac{\int k(u(h'/h))k(u)\,du}{\int k(u)^2\,du}$. 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,*

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

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

$$\frac{-\log\left(-\frac{1}{2}\log(1-\alpha)\right)+b(\overline{h}_n/\underline{h}_n,k)}{\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}}+\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)},\tag{3}$$

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 coverage of $\theta(h)$. Often, however, $\theta(0)$ is of interest. We now state a corollary showing coverage of $\theta(0)$ under an additional condition.

Corollary 3.1. *If $\sup_{h\in[\underline{h}_n,\overline{h}_n]}\frac{\sqrt{nh}|\theta(h)-\theta(0)|}{\hat{\sigma}(h)}=o_P\big((\log\log(\overline{h}_n/\underline{h}_n))^{-1/2}\big)$, and the conditions of Theo-*

rem 3.1 hold, then

$$P\left(\theta(0) \in \left\{\hat{\theta}(h) \pm \hat{\sigma}(h) \cdot c_{1-\alpha}(\overline{h}_n/\underline{h}_n, k)/\sqrt{nh}\right\} \text{ all } h \in [\underline{h}_n \leq h \leq \overline{h}_n]\right) \overset{n\to\infty}{\to} 1-\alpha.$$

Corollary 3.1 uses the additional condition that the bias $\theta(h) - \theta(0)$ is negligible relative to the standard error $\hat{\sigma}(h)/\sqrt{nh}$ uniformly over the range of bandwidths considered.[4] Typically, it is ensured by bias-correction or undersmoothing, as long as the smoothness conditions in Assumption 3.2 are appropriately strengthened.[5] In Section 4.1, we discuss how, in a regression discontinuity setting, our approach can be applied with a bias-correction proposed by Calonico et al. (2014), and illustrate this approach in empirical examples in Section 5. Critical values for constructing one-sided confidence intervals robust to bandwidth snooping are analogous to the two-sided case—see Supplemental Appendix S4 for details.

If the bandwidth choice is a priori tied to a pre-specified set, it is possible to further tighten the critical values. For example, Imbens and Lemieux (2008) suggest examining estimates at half and twice the original bandwidth $\hat{h}$, which yields the set $\{\hat{h}/2, \hat{h}, 2\hat{h}\}$. One can extend our approach to obtain critical values under such discrete snooping. However, such critical values will depend on the entire discrete set, and will often not be much tighter than $c_{1-\alpha}(\overline{h}_n/\underline{h}_n, k)$ with $\overline{h}_n$ and $\underline{h}_n$ given by the biggest and smallest bandwidths in the set (so long as the triangular or Epanechnikov kernel is used). For example, for the discrete bandwidth set $\{\hat{h}/2, \hat{h}, 2\hat{h}\}$ the critical value for the triangular kernel can be shown to equal 2.23, while $c_{1-\alpha}(4, k) = 2.26$.

In addition to providing the critical values $c_{1-\alpha}$, Theorem 3.1 provides a further approximation in (3) to the quantiles of $\sup_{\underline{h}_n \leq h \leq \overline{h}_n} \sqrt{nh}|\hat{\theta}(h) - \theta(h)|/\hat{\sigma}(h)$ based on an extreme value limiting distribution, provided that $\overline{h}_n/\underline{h}_n \to \infty$. In the case where k is the uniform kernel, $\psi(W_i, h)$ does not depend on h and $E[\psi(W_i, h)|X_i = x] = 0$ and $var[\psi(W_i, h)|X_i = x] = 1$ for all x, the latter result reduces to a well-known theorem of Darling and Erdös (1956) (see also Einmahl

[4] If $\overline{h}_n/\underline{h}_n$ is bounded, the condition in Corollary 3.1 is the same as the condition that is needed for pointwise-in-h coverage of conventional CIs that do not adjust for snooping. If $\overline{h}_n/\underline{h}_n \to \infty$, there is an additional log log term in the rate at which the bias must decrease, which arises for technical reasons. However, this term is small enough that this condition is still guaranteed by bias-correction methods such as the one proposed by Calonico et al. (2014).

[5] Alternatively, if a bound $\overline{b}(h)$ on the bias $|\theta(h) - \theta(0)|$ is available, one can allow the bias to be of the same order of magnitude as standard deviation by adding and subtracting $\hat{\sigma}(h) \cdot c_{1-\alpha}(\overline{h}_n/\underline{h}_n, k)/\sqrt{nh} + \overline{b}(h)$. This has the advantage of allowing for weaker conditions on the bandwidth sequence, including cases where the undersmoothing condition does not hold. See Chernozhukov et al. (2014a), Schennach (2015) and Donoho (1994) for applications of this idea in different settings.

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. We do not recommend using the critical value in (3) value in practice, as critical values based on extreme value results have been known to perform poorly in related settings (see Hall, 1991). Instead we recommend using the critical value $c_{1-\alpha}(\overline{h}_n/\underline{h}_n, k)$, which does not suffer from these issues because it is based directly on the Gaussian process approximation, and it remains valid even for fixed $\overline{h}_n/\underline{h}_n$ (see Figure S1 in the supplemental appendix for a comparison of these critical values). Thus, we report only this critical value in Table 1 below.

The main practical value of the approximation in (3) is that it demonstrates that critical value grows very slowly with $\overline{h}_n/\underline{h}_n$, at rate $\sqrt{\log\log(\overline{h}_n/\underline{h}_n)}$, so that the cost of examining a wider range of bandwidths relative to examining a moderate range is rather small. Indeed, while using our correction is important for maintaining correct coverage, as can be seen from Table 1, once $\overline{h}_n/\underline{h}_n$ is above 5, widening the range of bandwidths that one examines increases the critical value by only a small amount.

To outline how Theorem 3.1 obtains, consider again the problem of estimating a nonparametric mean at a point described in the introduction. Here the influence function is given by $\psi(W_i, h)k(X_i/h)$ where $\psi(W_i, h) = (Y_i - \theta(h))/\sqrt{var\left\{[Y_i - \theta(h)]k(X_i/h)/h\right\}}$ 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 bandwidth that achieves the supremum will still converge in probability to zero, yielding the first part of the theorem. For the second part of the theorem, we show 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. 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 the critical values $c_{1-\alpha}(\overline{h}_n/\underline{h}_n, k)$ 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 $\alpha = 0.10$, are given in Table S1 in the supplemental appendix. The critical values can also be obtained using our R package `BWSnooping`, which can be downloaded from `https://github.com/kolesarm/BWSnooping`. For local polynomial estimators, the critical value depends on the order of the local polynomial, as well as whether the point of interest is at the boundary (including the case of regression discontinuity) or in the interior of the support of X_i. We report values for Nadaraya-Watson (local constant) and local linear estimators. Note that the critical values for Nadaraya-Watson kernel regression are the same whether or not the point of interest is in the interior or at the boundary. For local linear regression in the interior, the equivalent kernel is the same as the original kernel, and therefore the critical value is the same as that for Nadaraya-Watson kernel regression. For local linear regression at the boundary, including inference in regression discontinuity designs, the critical value is different because the equivalent kernel is different (see Supplemental Appendix S2 for details).

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. Appropriate choice of $\underline{h}$ and $\overline{h}$ will depend on the application; Section 4 discusses this choice for the applications we consider. Look up the critical value $c_{1-\alpha}(\overline{h}_n/\underline{h}_n, k)$ in Table 1 for $\alpha = 0.05$, or in Table S1 for $\alpha = 0.01$ and $\alpha = 0.10$.

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

It is common practice to report an estimate $\hat{\theta}(\hat{h})$ and a standard error $se(\hat{h}) \equiv \hat{\sigma}(\hat{h})/\sqrt{n\hat{h}}$ for a value of $\hat{h}$ 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 $1 - \alpha = 0.95$. 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 $c_{1-\tilde{\alpha}}(\overline{h}_n/\underline{h}_n, k)$. For $\overline{h}/\underline{h}$ below 10, the true coverage is typically somewhere between 70% and 90%, depending on the kernel and the exact value of $\overline{h}/\underline{h}$.

4 Applications

This section applies the main results from Section 3 to three econometric models. In the first example, $\theta(0)$ is of primary interest, while in the other examples, $\theta(h)$ is an interesting economic object in its own right. Technical details for this section are relegated to Appendix B.

4.1 Regression discontinuity

We are interested in a regression discontinuity (RD) parameter, where the discontinuity point is normalized to $x = 0$ for convenience of notation. We consider both "sharp" and "fuzzy" regression discontinuity. Using arguments in the discussion preceding Theorem 3.1, the results in this section could also be generalized to cover "kink" designs (Card, Lee, Pei, and Weber, 2015), 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.

For fuzzy RD, 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 RD, we observe $\{(X_i, Y_i)\}_{i=1}^n$, and the parameter of inter-

est 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 (see, e.g., Porter, 2003).[6] Given a kernel function k^* and a bandwidth h, let $\hat{\alpha}_{\ell,Y}(h)$ and $\hat{\beta}_{\ell,Y}(h)$ denote the intercept and slope from a weighted linear regression of Y_i on X_i in the subsample with $X_i < 0$, weighted by $k(X_i/h)$. That is, $\hat{\alpha}_{\ell,Y}(h)$ and $\hat{\beta}_{\ell,Y}(h)$ minimize

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

Let $(\hat{\alpha}_{u,Y}(h), \hat{\beta}_{u,Y}(h))$ denote the regression coefficients from a regression in the subsample with $X_i \geq 0$. For the fuzzy case, define $(\hat{\alpha}_{\ell,D}(h), \hat{\beta}_{\ell,D}(h))$ and $(\hat{\alpha}_{u,D}(h), \hat{\beta}_{u,D}(h))$ analogously with D_i replacing Y_i. The sharp RD local linear estimator is then given by $\hat{\theta}(h) = \hat{\alpha}_{u,Y}(h) - \hat{\alpha}_{\ell,Y}(h)$. The fuzzy RD estimator is given by $\hat{\theta}(h) = \frac{\hat{\alpha}_{u,Y}(h) - \hat{\alpha}_{\ell,Y}(h)}{\hat{\alpha}_{u,D}(h) - \hat{\alpha}_{\ell,D}(h)}$.

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

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

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 RD, and $\theta(h) = \alpha_{u,Y}(h) - \alpha_{\ell,Y}(h)$ for sharp RD. Under appropriate smoothness conditions, $\theta(h)$ will converge to $\theta(0)$ as $h \to 0$.

Theorem B.1 in Appendix B shows that under appropriate conditions, Theorem 3.1 applies with $k(u)$ given by the equivalent kernel $k(u) = (\mu_{k^*,2} - \mu_{k^*,1}|u|)k^*(u)$, where $\mu_{k^*,j} = \int_{u=0}^{\infty} u^j k^*(u)$ for $j = 1, 2$ (rather than the original kernel k^*). 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 Table S1 in the supplemental appendix for $\alpha = 0.01$ and $\alpha = 0.10$.

In most RD applications, $\theta(0)$, rather than $\theta(h)$, is of primary interest. Let h_{ll}^* denote a bandwidth that minimizes the mean-squared error $E[(\hat{\theta}(h) - \theta(0))^2]$ of the local linear estimator (or an asymptotic approximation of it), such as the Imbens and Kalyanaraman (2012) bandwidth selector. Then, as is well-known, the bias of $\hat{\theta}(h_{ll}^*)$ will not be asymptotically negligible, and confidence intervals around $\hat{\theta}(h_{ll}^*)$ will have poor coverage of $\theta(0)$, even without any snooping.

In an important paper, Calonico et al. (2014, CCT) show that one can address this issue

[6] We cover the extension to local polynomial regression of higher order in Appendix S2.

by recentering the confidence interval by subtracting an estimate of the asymptotic bias, and rescaling it to account for the additional noise induced by the bias estimation. CCT show that the remaining bias is asymptotically negligible so that this alternative confidence interval will achieve proper coverage of $\theta(0)$, provided the conditional mean functions are smooth enough on each side of the cutoff. If the pilot bandwidth used to estimate the bias equals h^*_{ll}, this procedure is equivalent to constructing the usual confidence interval around a local quadratic estimator with bandwidth h^*_{ll}. Since the MSE optimal bandwidth for local quadratic regression is of larger order than the optimal bandwidth for local linear regression, this method of constructing confidence intervals can also be viewed as a particular undersmoothing procedure. Consequently, if one uses a local quadratic estimator and $\overline{h} = \mathcal{O}(h^*_{ll})$, Corollary 3.1 applies, so that our adjusted confidence intervals will also achieve correct coverage of the RD parameter $\theta(0)$. We apply this method in two empirical examples in Section 5, and investigate its finite-sample properties in a Monte Carlo exercise in Supplemental Appendix S5.

In the remainder of this subsection, we discuss two cases in which our computing our adjusted confidence interval is relevant. We also discuss the choice of $\overline{h}$ and $\underline{h}$.

4.1.1 Sensitivity Analysis

A researcher implements the CCT bias-correction method by calculating the local quadratic estimator of the sharp RD parameter $\theta(0) = \lim_{x \downarrow 0} E(Y_i|X_i = x) - \lim_{x \uparrow 0} E(Y_i|X_i = x)$ at the bandwidth $h = h^*_{ll}$. To check the robustness of the results, the researcher also evaluates the estimator at a bandwidth $h_{\text{smaller}} < h^*_{ll}$. Suppose that the CI evaluated at h^*_{ll} contains zero, while the CI evaluated at h_{smaller} does not (in any given sample, this may happen even if both estimators are exactly unbiased). Arguing that the bias of the estimator at h_{smaller} is negligible 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 valid under even weaker assumptions than the original assumptions needed for validity of the CCT confidence interval. Unfortunately, this is not true for the actual hypothesis test that the researcher has performed (looking at both h^*_{ll} and h_{smaller}), since the α probability of type I error has already been "used up" on the test based on h^*_{ll}. By replacing $z_{1-\alpha/2}$ with the critical value $c_{1-\alpha}(h^*_{ll}/h_{\text{smaller}}, k)$, the researcher can conclude that $\theta(0) \neq 0$ under the original assumptions, so long as at least one of the two confidence in-

tervals does not contain zero. Appendix C provides further discussion of cases in which the uniform-in-h confidence bands can be useful in sensitivity analysis.

4.1.2 Adaptive inference

Suppose that it is known from the economics of the problem that the conditional mean function $E(Y_i|X_i = x)$ is weakly decreasing. Then a Nadaraya-Watson (local constant) estimator $\hat{\theta}_{NW}(h)$ of the sharp RD parameter must be biased downward for any bandwidth h. Because any downward bias will make the one-sided confidence interval $[\hat{\theta}_{NW}(h) - z_{1-\alpha}\hat{\sigma}_{NW}(h)/\sqrt{nh}, \infty)$ only more conservative, it is asymptotically valid for any h regardless of how fast $h \to 0$ with n (even if h does not decrease with n at all), so long as $nh \to \infty$ so that a central limit theorem applies to $\hat{\theta}_{NW}(h)$.

One may wish to use this fact to "snoop" by reporting the most favorable confidence interval, namely, $[\sup_{h\in[\underline{h},\overline{h}]}(\hat{\theta}_{NW}(h) - z_{1-\alpha}\hat{\sigma}_{NW}(h)/\sqrt{nh}), \infty)$ for some $[\underline{h}, \overline{h}]$. Because it involves entertaining multiple bandwidths, this is not a valid confidence interval. Replacing $z_{1-\alpha}$ with one-sided version of our critical value, $c^{\text{os}}_{1-\alpha}$ (see Supplemental Appendix S4), leads to a confidence interval $[\sup_{h\in[\underline{h},\overline{h}]}(\hat{\theta}(h) - c^{\text{os}}_{1-\alpha}\hat{\sigma}(h)/\sqrt{nh}), \infty)$, which will have correct asymptotic coverage.

In fact, this confidence interval enjoys an optimality property of being adaptive to certain levels of smoothness of the conditional mean, that is, it is almost as tight as the tightest confidence interval if the smoothness of the conditional mean was known. More formally, suppose $E(Y_i|X_i = x)$ approaches $E(Y_i|X_i = 0)$ at the rate x^β for some $\beta \in (0, 1]$. Then, so long as $\overline{h} \to 0$ slowly enough and $\underline{h} \to 0$ quickly enough, 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, 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 necessary for this form of adaptation. In particular, one cannot estimate the optimal bandwidth accurately enough to do away with our correction (see Armstrong, 2015, for details).

4.1.3 Choice of $\underline{h}$ and $\overline{h}$

Let us discuss some general considerations for a choice of the smallest and largest bandwidth in the context of sensitivity analysis in RD (for adaptive inference under monotonicity, the appropriate choice depends on the range of smoothness levels of the conditional mean, see Armstrong (2015) for details). A conservative approach is to set $\underline{h}$ to the smallest value such that enough observations are used for the central limit theorem to give a good approximation (say, 50 effective observations). If one is interested in inference on $\theta(0)$, using the CCT bias-correction discussed above, $\overline{h}$ can be set to be of the same order as h_{ll}^*, such as $\overline{h} = 3h_{ll}^*/2$ or $\overline{h} = 2h_{ll}^*$. Alternatively, one can take an even more conservative approach of setting $\overline{h}$ to include all of the data, so long as one keeps in mind that CIs with h much larger than h_{ll}^* may not contain $\theta(0)$ due to bias. Given that the critical value increases slowly with $\overline{h}/\underline{h}$ for moderate to large values of $\overline{h}/\underline{h}$, the resulting critical value will not be much larger than under a more moderate choice of $\underline{h}$ and $\overline{h}$.

Implementations of the MSE optimal bandwidth such as those in Imbens and Kalyanaraman (2012) and Calonico et al. (2014) typically yield a random bandwidth, so if $\overline{h}$ depends on it, it will also be random. While we state our results for nonrandom $[\underline{h}, \overline{h}]$, our results can be extended to this case without the need for additional corrections so long as $\overline{h}/\overline{h}^* \xrightarrow{p} 1$ and $\underline{h}/\underline{h}^* \xrightarrow{p} 1$ for some nonrandom sequences $\overline{h}^*$ and $\underline{h}^*$ satisfying our conditions. Imbens and Kalyanaraman (2012) exhibit a nonrandom sequence h_{IK}^* such that their bandwidth selector $\hat{h}_{IK}^*$ satisfies $\hat{h}_{IK}^*/h_{IK}^* \xrightarrow{p} 1$ under certain conditions, so that one can take, for example, $\overline{h} = \hat{h}_{IK}^*$ or $\overline{h} = 2\hat{h}_{IK}^*$ under these conditions.[7] Note, however, that bandwidth selectors such as $\hat{h}_{IK}^*$ can, in practice, exhibit substantial variability and dependence on tuning parameters chosen by the user.[8] To the extent that a data-dependent bandwidth is highly variable in finite samples, or can be "gamed" using tuning parameters, it is safer to make use a conservative choice of $[\underline{h}, \overline{h}]$ that contains the data-dependent bandwidth with probability one regardless of the tuning

[7] While this argument applies to certain data-dependent bandwidth selectors, one cannot use arbitrary data-dependent rules to choose $[\underline{h}, \overline{h}]$ in our setup. As an extreme example, choosing $\overline{h} = \underline{h}$ to minimize the p-value for a particular hypothesis (and then arguing that a snooping correction is not needed since $\overline{h} = \underline{h}$) is clearly not compatible with our setup. Rather, one would have to define $[\underline{h}, \overline{h}]$ to be the range over which the p-value was minimized.

[8] For example, as Imbens and Kalyanaraman (2012) point out, the bandwidth that is "optimal" according to their definition is infinite in certain cases. Their procedure uses tuning parameters to ensure that the selected bandwidth goes to zero in these cases, resulting in a bandwidth sequence that depends on tuning parameters asymptotically. This can lead to substantial differences between different implementations of their approach such as the original implementation in Imbens and Kalyanaraman (2012) and the implementation in Calonico et al. (2014).

parameters.

4.2 Trimmed average treatment effects under unconfoundedness

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

Let $Y_i(0)$ and $Y_i(1)$ denote the potential outcomes associated with a binary treatment D_i that is as good as randomly assigned, conditional on covariates X_i, so that $E(Y_i(d)|X_i, D_i) = E(Y_i(d)|X_i)$. We observe and i.i.d. sample $\{(X_i, D_i, Y_i)\}_{i=1}^n$, where $Y_i = Y_i(1)D_i + Y_i(0)(1 - D_i)$ denotes the observed outcome. Let $\tau(x) = E(Y_i(1) - Y_i(0) \mid X_i = x) = \mu_1(x) - \mu_0(x)$ denote the average treatment effect for individuals with $X_i = x$, where $\mu_d(x) = E(Y_i|X_i = x, D_i = d)$. Let $e(x) = P(D_i = 1|X_i = x)$ denote the propensity score.

Typically, we are interested in the ATE for the whole population, $\theta(0) = E[Y_i(1) - Y_i(0)]$. However, since effects for individuals with propensity score $e(X_i)$ close to zero or one cannot be estimated very precisely, in samples with limited overlap (i.e. in which the number of such individuals is high), estimates of the ATE $\theta(0)$ will be too noisy. To deal with this problem, it is common in empirical applications to trim the sample by discarding observations with extreme values of the propensity score.[9] Doing so, however, changes the estimand. In particular, if the sample is restricted to individuals with moderate values of the propensity score, $\mathcal{X}_h = \{X_i \colon h \leq e(X_i) \leq 1 - h\}$ for some $0 \leq h < 1/2$, then, as discussed in Crump, Hotz, Imbens, and Mitnik (2009), the estimand changes from $\theta(0)$ to

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

One therefore faces the trade-off between increasing h from 0 to increase the precision of the estimator at the cost of making the estimand $\theta(h)$ arguably less interesting. See Crump et al. (2009), Chaudhuri and Hill (2016) and Khan and Tamer (2010) for a detailed discussion of these issues. Crump et al. (2009) propose a rule for picking the trimming parameter h that minimizes

[9] For prominent examples, see Heckman, Ichimura, and Todd (1997), Galiani, Gertler, and Schargrodsky (2005), or Bailey and Goodman-Bacon (2015).

the variance of the resulting estimator. In practice, one may want to resolve this trade-off in other ways. 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. With the reported confidence band for $\theta(h)$, the reader can pick their preferred trimming value, assess treatment effect heterogeneity by examining how $\theta(h)$ varies with h, or obtain a confidence interval for $\theta(0)$ based on the reader's own beliefs about the smoothness of $\theta(h)$.

When forming this confidence band, one can choose a trimming range $[\underline{h}, \overline{h}]$ that is wide enough to include different suggestions in the literature about the appropriate amount of trimming, such as $[0, 0.1]$ or $[0, 0.2]$. This allows the reader to pick their preferred trimming amount as well as assess the sensitivity of the results to the amount of trimming.

To describe the adjustment to critical values in this setting, let $\hat{\theta}(h)$ be an efficient estimator of $\theta(h)$ (in the sense of satisfying condition (12) in Appendix B), and let $se(h)$ denote its standard error. Let $N(h)$ be the number of untrimmed observations for a given h (i.e. number of observations i such that $X_i \in \mathcal{X}_h$). In contrast to the previous applications, assume that $\underline{h}$ and $\overline{h}$ are fixed. If $e(X_i)$ 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. We form our uniform confidence band as

$$\left\{\hat{\theta}(h) \pm c_{1-\alpha}(\hat{t}, k_{\text{uniform}}) \cdot se(h) \middle| h \in [\underline{h}, \overline{h}]\right\}, \quad \text{where} \quad \hat{t} = \frac{se(\underline{h})^2 N(\underline{h})^2}{se(\overline{h})^2 N(\overline{h})^2}, \tag{5}$$

and k_{uniform} denotes the uniform kernel. In Theorem B.2 in Appendix B, we show that this confidence band is asymptotically valid under appropriate regularity conditions. 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 proof of Theorem B.2 in Supplemental Appendix S3.2 for details.

4.3 LATEs for different sets of compliers

We observe (Z_i, D_i, Y_i) where Z_i is an exogenous instrument shifting a binary 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). Suppose that $P(D_i = 1 \mid Z_i = z)$ is increasing in z, and for $h \leq (\overline{z} - \underline{z})/2$ define

$$\theta(h) = \frac{E(Y_i \mid Z_i \in [\overline{z} - h, \overline{z}]) - E(Y_i | Z_i \in [\underline{z}, \underline{z} + h])}{P(D_i = 1 \mid Z_i \in [\overline{z} - h, \overline{z}]) - P(D_i = 1 | Z_i \in [\underline{z}, \underline{z} + h])}.$$

Under certain exogeneity and monotonicity assumptions, $\theta(h)$ gives the average effect for the subpopulation of "compliers", individuals who change their treatment status if their instrument shifts from $Z_i \in [\underline{z}, \underline{z} + h]$ to $Z_i \in [\overline{z} - h, \overline{z}]$. In the literature, this is called the "local average treatment effect", or LATE (see Imbens and Angrist, 1994; Heckman and Vytlacil, 2005; Heckman, Urzua, and Vytlacil, 2006). It can be estimated with the sample analogue

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

where $\#\mathcal{A}$ denotes the number of elements in a set $\mathcal{A}$. The estimator $\hat{\theta}(h)$ 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 $I(Z_i \geq \overline{z} - h)$. Let $\hat{\sigma}^2(h)/h$ 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 $\hat{\theta}(h)$.

The parameter $\theta(0) = \lim_{h \to 0} \theta(0)$ is typically of particular interest since it corresponds to the LATE for the largest subpopulation for which the LATE is identified (see Frölich, 2007; Heckman and Vytlacil, 2005; Heckman et al., 2006). In finite samples one faces a trade-off similar to that in the trimmed ATE application in Section 4.2: increasing h increases the precision of the estimate, but decreases the size of the complier subpopulation associated with the estimand.

Theorem B.3 in Appendix B shows that under appropriate regularity conditions, the confidence band $[\hat{\theta}(h) \pm c_{1-\alpha}(\overline{h}/\underline{h}, k_{\text{uniform}}) se(h)]_{h \in [\underline{h}, \overline{h}]}$, where k_{uniform} denotes the uniform kernel, is a valid confidence band for $\theta(h)$. This result follows from the fact that $\hat{\theta}(h)$ is composed of kernel-based estimators with the uniform kernel (e.g. $\frac{1}{\#\{Z_i \in [\underline{z}, \underline{z} + h]\}} \sum_{Z_i \in [\underline{z}, \underline{z} + h]} Y_i$ is a uniform kernel estimate of $E[Y_i \mid Z_i = \underline{z}]$). This confidence band 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. This allows the reader to assess treatment

effect heterogeneity by examining how $\theta(h)$ varies with h, or obtain a confidence interval for $\theta(0)$ based on their own beliefs about the smoothness of $\theta(h)$.

In addition to the trimmed ATE and LATE applications, similar extensions are possible to other econometric models that are "identified at infinity" (see, among others Chamberlain, 1986; Heckman, 1990; Andrews and Schafgans, 1998). In the interest of brevity, we do not pursue such extensions here.

5 Empirical illustrations

5.1 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, so that the incumbency advantage can be identified using a sharp regression discontinuity design.

In particular, the running variable X_i is the Democratic margin of victory in a given election i. 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.

To analyze the data, Lee (2008) uses a global fourth degree polynomial, which yields a point estimate of 7.7%. However, global polynomial estimates may give large weights to observations far away from the threshold and be sensitive to the degree of the polynomial (Gelman and Imbens, 2014). We therefore reanalyze the data using local linear and local quadratic regression with a triangular kernel. We consider bandwidths between 2 and 40, which includes the Imbens and Kalyanaraman (2012, IK) optimal bandwidth selector for local linear regression, equal to 29.4. Figure 3 plots the results. Because the IK bandwidth is designed to minimize the mean squared error of the local linear estimator, as discussed in Section 4.1, the bias at bandwidths of this order is not asymptotically negligible. Panel (a) of Figure 3 should therefore be interpreted as a confidence band for $\theta(h)$. As discussed in that section, one can interpret the local quadratic estimator as implementing the Calonico et al. (2014) bias-correction method, so that panel (b) can

be interpreted as giving results for $\theta(0)$.

The incumbency effect remains positive and significant over the entire range, even after using the corrected critical value, and after implementing the Calonico et al. (2014) bias correction. At the IK bandwidth, the confidence interval is given by $(4.49, 8.87)$ for the local quadratic (bias-corrected) estimator. Our adjustment widens it slightly to $(3.82, 9.54)$. These results suggest that the estimates are very robust to the choice of bandwidth.

5.2 Progresa / Oportunidades

Our second empirical example examines the effect of the Oportunidades anti-poverty conditional cash transfer program in Mexico, using a dataset from Calonico et al. (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 around the local linear estimator equal to $(6.7, 71.2)$, and to $(4.6, 102.7)$ for the local quadratic estimator, suggesting a significantly positive effect. For non-food consumption, $h_{IK} = 1.09$, and the 95% confidence intervals are given by $(1.6, 53.7)$ for the local linear estimator, and by $(4.5, 79.3)$ for the local quadratic estimator. To examine sensitivity of these results to snooping, we plot the estimates, along with pointwise and uniform confidence bands over a range of bandwidths in Figures 4 and 5. In contrast to the previous

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.

5.3 Right heart catheterization

Our final example uses data from Connors Jr et al. (1996) to examine the effect of receiving right heart catheterization (RHC) on 30-day mortality. The data contain information on 5,735 adult patients who were critically ill upon admission to the hospital ICU, 2,184 treated and 3,551 controls. The treatment, an indicator for receiving RHC withing 24 hours of admission, is assumed to be as good as randomly conditional on 72 covariates (see Connors Jr et al. (1996) for a detailed description).

The original analysis by Connors Jr et al. (1996) matched on the propensity score estimated by a logistic regression, with each unit matched at most once. It found that RHC appeared to lead to lower survival than not performing RHC, contradicting a popular perception among practitioners that RHC was beneficial. To estimate the treatment effect, we follow the procedure in reanalysis of this data by Crump et al. (2009). First, we estimate the propensity score by logistic regression. We then take the difference between the treated and control units weighted by the estimated propensity score. Standard errors are computed by the bootstrap.

Due to limited overlap, Crump et al. (2009) trim the data by setting the trimming parameter to $h = 0.1$, discarding individuals with propensity score lower than 0.1 and higher than 0.9. To examine sensitivity of the results to the amount of trimming, we consider a range of trimming parameters from 0 to 0.1. This leads to an effective bandwidth ratio $\hat{t} = 2.00$. Figure 6 plots the results. Without trimming, the unadjusted 95% confidence interval is given by $(0.027, 0.092)$. Trimming at $h = 0.1$ reduces it to $(0.031, 0.087)$. Adjusting the confidence intervals for snooping widens them to $(0.018, 0.100)$ and $(0.024, 0.094)$, respectively. Overall, the results are stable over the trimming range, with the precision of the estimates increasing with trimming. The conclusion that RHC negatively impacts survival is robust to snooping, with RHC lowering the 30-day survival probability by about 6%.

6 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. Appendix A contains the proof of the main result. Appendix B gives formal regularity conditions applying the main result to models considered in Section 4. Appendix C discusses the use of uniform and pointwise in h confidence regions in sensitivity analysis. Additional results and proofs 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.

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*

$$\sqrt{2\log\log(h_n^*/\underline{h}_n)}\left(\sup_{\underline{h}_n \le h \le h_n^*} \mathbb{T}_n(h) - \sqrt{2\log\log(h_n^*/\underline{h}_n)}\right) - b(\log\log(h_n^*/\underline{h}_n)) \xrightarrow{d} Z \tag{6}$$

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

$$\sqrt{\log\log(h_n^*/\underline{h}_n)} \sup_{\underline{h}_n \le h \le h_n^*} \left|\mathbb{T}_n(h) - \tilde{\mathbb{T}}_n(h)\right| \xrightarrow{p} 0. \tag{7}$$

Then (6) 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* $\varepsilon > 0$,

$$\frac{\sup_{h_n^* \le h \le \overline{h}_n} \tilde{\mathbb{T}}_n(h)}{\sqrt{2\log\log(\overline{h}_n/\underline{h}_n)}} \le 1 - \varepsilon \text{ with probability approaching one,} \tag{8}$$

then (6) 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 (6) 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 $\varepsilon > 0$,

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

with probability approaching one, where

$$h_n^* = \exp\left[-(\log \underline{h}_n^{-1})^{1/K}\right] \tag{9}$$

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^*, (7) 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

$$\tilde{Y}_i = \frac{\psi(W_i,0) - E[\psi(W_i,0)||X_i|]}{\sqrt{var(\psi(W_i,0)||X_i|)f_{|X|}(|X_i|)\int_0^\infty k(u)^2\,du}}. \tag{10}$$

Next, it is shown that (7) 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 the absolute value of 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 et al., 1983).

The reasoning behind this choice of h_n^* is as follows. With $h_n^* = \exp[-(\log \underline{h}_n^{-1})^{1/K}]$, we have $h_n^*/\underline{h}_n = \exp[-(\log \underline{h}_n^{-1})^{1/K} + (\log \underline{h}_n^{-1})] = \exp\{(\log \underline{h}_n^{-1})[1 - (\log \underline{h}_n^{-1})^{1/K-1}]\}$, 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 (8), 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, (8) 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 Supplemental Appendix S1.2 (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 Supplemental Appendix S1.1. 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 $\ell(h)$,

$$f(W_i,h,t)k(X_i/h) \leq \ell(h) \tag{11}$$

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

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

Thus,

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

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 (8) holds for $\tilde{\mathbb{T}}_n(h)=\sqrt{nh}|\hat{\theta}(h)-\theta(h)|/\hat{\sigma}(h)$, and that (7) 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 (8) 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, (8) 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, (8) 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 $1-2\delta$,

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

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 (7) 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* (9)*,*

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

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,

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

$$= \mathcal{O}_P\left(\ell(h_n^*)\log\log h_n^{*-1}\frac{\sqrt{\log\log(h_n^*/\underline{h}_n)}}{\sqrt{\log\log h_n^{*-1}}}\right).$$

The result follows since $\ell(h_n^*)\log\log h_n^{*-1} \to 0$ and

$$\frac{\sqrt{\log\log(h_n^*/\underline{h}_n)}}{\sqrt{\log\log h_n^{*-1}}} \leq \frac{\sqrt{\log\log \underline{h}_n^{-1}}}{\sqrt{\log\log h_n^{*-1}}} = \frac{\sqrt{\log\log \underline{h}_n^{-1}}}{\sqrt{(1/K)\log\log \underline{h}_n^{-1}}} = \sqrt{K}.$$

□

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 $\sup_{0\le x\le Ah}\ell(x)$ for a function $\ell(h)$ with $\ell(h)\log\log h^{-1}\to 0$ as $h\to 0$. Since $[\log\log h^{-1}]\sup_{0\le x\le Ah}\ell(x)\le \sup_{0\le x\le Ah}[\log\log x^{-1}]\ell(x)$, this bound satisfies the required conditions. The last term is bounded by a constant times $\sup_{0\le x\le Ah}|\tilde{\mu}(x)-\tilde{\mu}(0)|$, 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

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

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,

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

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

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

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 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 this result, given in Supplemental Appendix S1.3, uses an application of a Gaussian approximation theorem of Sakhanenko (1985) along with arguments similar to those in Bickel and Rosenblatt (1973). It is worth noting that other Gaussian approximation results could be used, with potentially different regularity conditions. For example, one could use results from Chernozhukov, Chetverikov, and Kato (2014b), which would allow us to replace the assumption of a bounded outcome variable with assumptions bounding the higher moments, at the expense of stronger conditions on $\underline{h}_n$ (this would also require additional truncation arguments elsewhere in the proofs).

We consider a general setup with $\{(\tilde{X}_i, \tilde{Y}_i)\}_{i=1}^n$ i.i.d., 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 (10) 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

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

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 $n\underline{h}_n/[(\log\log n)(\log\log\log n)]^2 \to \infty$ given in the conditions of Theorem 3.1, is sufficient, since this implies, for some $a_n \to \infty$,

$(n\underline{h}_n)^{1/4} \geq a_n(\log\log n)^{1/2}(\log\log\log n)^{1/2}$ and this implies, for large enough n,

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

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 a 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 $k(A) \neq 0$. Let $\underline{h}_n$ and $\overline{h}_n$ be sequences with $\overline{h}_n/\underline{h}_n \to \infty$. Then

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

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

Proof. We use Theorem 12.3.5 of Leadbetter et al. (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 Supplemental Appendix S1.4. □

A.6 Proof of Theorem 3.1

We are now ready to prove Theorem 3.1. Before proceeding, we recall a result regarding absolute continuity of suprema of Gaussian processes, which will be used in the proof.

Lemma A.7. *Let $\mathbb{X}(t)$ be a Gaussian process on a countable index set $\mathcal{T}$ with $P(\sup_{t\in\mathcal{T}} |\mathbb{X}(t)| < \infty) = 1$ and $\inf_{t\in\mathcal{T}} var(\mathbb{X}(t)) > 0$. Then $\sup_{t\in\mathcal{T}} \mathbb{X}(t)$ has an absolutely continuous distribution with bounded density.*

Proof. See Proposition 3.2 in Pitt and Tran (1979). □

It follows from Lemma A.7 that the distribution of $\sup_{h\in[\underline{h},\overline{h}]} \mathbb{H}(h)$ is absolutely continuous for any $0 < \underline{h} \leq \overline{h} < \infty$ (the supremum is equal to the supremum over a countable subset with probability one by continuity of the sample paths). It also follows that $\sup_{h\in[\underline{h},\overline{h}]} |\mathbb{H}(h)|$ is absolutely continuous, since $\sup_{h\in[\underline{h},\overline{h}]} \mathbb{H}(h)$ and $\sup_{h\in[\underline{h},\overline{h}]} -\mathbb{H}(h)$ are absolutely continuous and absolute continuity of Y and Z implies absolute continuity of $\max\{Y, Z\}$.

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,

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

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

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

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 $\sup_{1\leq h\leq h^*} |\mathbb{H}(h)|$ at $c_{1-\alpha}(h^*, k)$ (which follows from the result in Pitt and Tran 1979 stated in Lemma A.7).

In the case where $\overline{h}_n/\underline{h}_n \to \infty$, let h_n^* be given by (9) 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 (8) holds for $\sqrt{nh}|\hat{\theta}(h) - \theta(h)|/\hat{\sigma}(h)$ 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, (7) 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,

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

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 $Z \vee Z'$ by Theorem A.2, and since the cdf of $Z \vee Z'$ is continuous, the result follows. The last display in the statement of the theorem follow directly from this extreme value limit. □

B Regularity conditions for applications

In this appendix, we state three theorems that give primitive regularity conditions for applications discussed in Section 4. Proofs for these results are in Supplemental Appendix S3.

To state these conditions, we use the following additional notation. For a random vector (X_i, D_i, Y_i) with X_i continuously distributed, let $E(Y_i|D_i = d, X_i = \tilde{x}_+) = \lim_{x\downarrow\tilde{x}} E(Y_i|D_i = d, X_i = x)$ and $E(Y_i|D_i = d, X_i = \tilde{x}_-) = \lim_{x\uparrow\tilde{x}} E(Y_i|D_i = d, X_i = x)$. For a function $f : \mathbb{R} \to \mathbb{R}$ and $\tilde{x} \in \mathbb{R}$, let $f(\tilde{x}_+) = \lim_{x\downarrow\tilde{x}} f(x)$ and $f(\tilde{x}_-) = \lim_{x\uparrow\tilde{x}} f(x)$ when these limits exist. We say that a function f is right-continuous at $\tilde{x}$ with local modulus of continuity $\ell(x)$ if $\|f(x) - f(\tilde{x}_+)\| \leq \ell(\|x - \tilde{x}\|)$ for all $x > \tilde{x}$ with $\|x - \tilde{x}\|$ small enough. We say that a function f is left-continuous at $\tilde{x}$ with local modulus of continuity $\ell(x)$ if $\|f(x) - f(\tilde{x}_-)\| \leq \ell(\|x - \tilde{x}\|)$ for all $x < \tilde{x}$ with $\|x - \tilde{x}\|$ small enough. We say that a function f is continuous at $\tilde{x}$ with local modulus $\ell(x)$ if it is both left- and right-continuous with $f(\tilde{x}_+) = f(\tilde{x}_-) = f(\tilde{x})$. Note that we define left- and right-continuity with respect to the left- and right-hand limits of the function, so that a function may be both left- and right-continuous according to our definition even if these limits are different (as is typically the case in RD).

Theorem B.1. *Consider the regression discontinuity design from Section 4.1. 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 in Section 4.1 and* $\hat{\sigma}(h)$ *corresponding to the Eicker-Huber-White standard error estimator given in the supplemental 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* $n\underline{h}_n/(\log\log \underline{h}_n^{-1})^3 \to \infty$.

Next, consider the problem of constructing uniform confidence bands for average treatment effects under unconfoundedness as in Section 4.2. Let $\hat{\theta}(h)$ be an estimator of $\theta(h)$ with influence function representation

$$\sqrt{n}(\hat{\theta}(h) - \theta(h)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{[\tilde{Y}_i - \theta(h)] I(X_i \in \mathcal{X}_h)}{P(X_i \in \mathcal{X}_h)} + o_P(1), \tag{12}$$

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 et al. (2009) for references to the literature for estimators that satisfy this condition. This condition requires that the trimmed sample is constructed using a known propensity score $e(x)$ (while allowing for a fully nonparametric estimator on the trimmed sample). If instead one uses the trimming rule $\hat{\mathcal{X}}_h = \{x\colon h < \hat{e}(x) \leq 1 - h\}$ based on an estimated propensity score $e(x)$, we conjecture that (12) will hold for $\sqrt{n}(\hat{\theta}(h) - \theta(\widehat{\mathcal{X}}_h))$ under regularity conditions, where, for a set $\mathcal{X}$, $\theta(\mathcal{X})$ is defined as $E(Y_i(1) - Y_i(0) \mid X_i \in \mathcal{X})$.[10] This will lead to uniform confidence bands for the parameter $\theta(\widehat{\mathcal{X}}_h)$, which can be interpreted as the average treatment effect for the random subpopulation $\widehat{\mathcal{X}}_h$. The influence function (12) and the pivotal asymptotic distribution we derive below are specific to estimators of trimmed average treatment effects: other classes of estimators and trimming rules with different influence function representations may not lead to a snooping adjusted critical value based on a pivotal asymptotic

[10] In the pointwise-in-h case, similar results are given in the the working paper version (Crump, Hotz, Imbens, and Mitnik, 2006) of Crump et al. (2009); verifying this conjecture essentially involves verifying that their results can be generalized to hold uniformly over $\underline{h} \leq h \leq \overline{h}$.

distribution. 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

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

The following theorem proves the validity of the confidence band given in Section 4.2.

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

(i) the influence function representation (12) 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 $\sigma(h)$ uniformly 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 (5). Then

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

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 $=$.*

The final result gives the regularity conditions for inference on local average treatment effects.

Theorem B.3. *Consider the setup and notation from Section 4.3. 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 $n\underline{h}_n/(\log\log \underline{h}_n^{-1})^3 \to \infty$.

C 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 determine the subset of included covariates, as in Leamer, 1983), provided we have an estimator $\hat{\theta}(h)$ that, for a given h, is approximately unbiased for $\theta(h)$.

Suppose that different readers may disagree on how $\theta(h)$ relates to $\theta(0)$, the parameter of interest, as h varies. We can report 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 believes that a particular h is most suitable for estimating and performing inference on $\theta(0)$, and, if given access to the original data, would only perform analysis based on this h, then the researcher can simply report $\hat{\theta}(h)$ and $\mathcal{C}_{\text{pointwise}}(h)$ for a range of values of h. The reader would then select an estimate $\hat{\theta}(h)$ and a confidence set $\mathcal{C}_{\text{pointwise}}(h)$ that correspond to their prior belief about the most appropriate h. The confidence set $\mathcal{C}_{\text{pointwise}}(h)$ 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.

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.

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:

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

One may consider formalizing the notion of "concluding that $\theta(0)$ is greater than zero in a robust sense" by either

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

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). Rejecting only when $\mathcal{C}_{\text{pointwise}}(h) \subseteq (0,\infty)$ for all h provides a valid test of $H_{0,a}$ since, under $H_{0,a}$ there exists a h_0 such that $\theta(h_0) \leq 0$, so that $P\left(\mathcal{C}_{\text{pointwise}}(h) \subseteq (0,\infty) \text{ all } h\right) \leq P\left(\mathcal{C}_{\text{pointwise}}(h_0) \subseteq (0,\infty)\right) \leq P\left(\theta(h_0) \notin \mathcal{C}_{\text{pointwise}}(h_0)\right)$. Thus, under the criterion (13), it is sufficient to use pointwise-in-h confidence bands. However, this approach is likely to be conservative in many practically relevant situations: the confidence set for $\theta(0)$ is effectively given by the union of all pointwise sets $\mathcal{C}_{\text{pointwise}}(h)$. In our case, where $\hat{\theta}(h)$ is a kernel based estimate with bandwidth h, such confidence interval will be very large since pointwise confidence intervals for small h can be very wide.

If, instead, one takes (14) as the criterion for "concluding that $\theta(0)$ 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. In contrast, performing this analysis with $\mathcal{C}_{\text{pointwise}}(h)$ does not provide a test of $H_{0,c}$ with correct size: this formulation of robustness of the results to 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, 497–517.

Armstrong, T. (2015): "Adaptive testing on a regression function at a point," *The Annals of Statistics*, 43, 2086–2101.

Armstrong, T. B. and H. P. Chan (2016): "Multiscale adaptive inference on conditional moment inequalities," *Journal of Econometrics*, 194, 24–43.

Bailey, M. J. and A. Goodman-Bacon (2015): "The War on Poverty's Experiment in Public Medicine: Community Health Centers and the Mortality of Older Americans," *American Economic Review*, 105, 1067–1104.

Bickel, P. J. and M. Rosenblatt (1973): "On some global measures of the deviations of density function estimates," *The Annals of Statistics*, 1, 1071–1095.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 82, 2295–2326.

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

Card, D., D. S. Lee, Z. Pei, and A. Weber (2015): "Inference on Causal Effects in a Generalized Regression Kink Design," *Econometrica*, 83, 2453–2483.

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

Chaudhuri, S. and J. B. Hill (2016): "Robust Estimation and Inference for Average Treatment Effects," Unpublished manuscript, University of North Carolina at Chapel Hill.

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, 2786–2819.

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

——— (2014b): "Gaussian approximation of suprema of empirical processes," *The Annals of Statistics*, 42, 1564–1597.

Connors Jr, A. F., T. Speroff, N. V. Dawson, C. Thomas, F. E. Harrell Jr, D. Wagner, N. Desbiens, L. Goldman, A. W. Wu, R. M. Califf, W. J. Fulkerson Jr, H. Vidaillet, S. Broste, P. Bellamy, J. Lynn, and W. A. Knaus (1996): "The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill patients," *The Journal of the American Medical Association*, 276, 889–897.

Crump, R. K., V. J. Hotz, G. W. Imbens, and O. A. Mitnik (2006): "Moving the Goalposts: Addressing Limited Overlap in the Estimation of Average Treatment Effects by Changing the Estimand," Working Paper 330, National Bureau of Economic Research.

——— (2009): "Dealing with limited overlap in estimation of average treatment effects," *Biometrika*, 96, 187–199.

Darling, D. and P. Erdös (1956): "A limit theorem for the maximum of normalized sums of independent random variables," *Duke Mathematical Journal*, 23, 143–155.

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

Donoho, D. L. (1994): "Statistical Estimation and Optimal Recovery," *The Annals of Statistics*, 22, 238–270.

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

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

Galiani, S., P. Gertler, and E. Schargrodsky (2005): "Water for Life: The Impact of the Privatization of Water Services on Child Mortality," *Journal of Political Economy*, 113, 83–120.

Gelman, A. and G. W. Imbens (2014): "Why high-order polynomials should not be used in regression discontinuity designs," Working Paper 20405, National Bureau of Economic Research.

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

Hall, P. (1991): "On convergence rates of suprema," *Probability Theory and Related Fields*, 89, 447–455.

Härdle, W. (1989): "Asymptotic maximal deviation of M-smoothers," *Journal of Multivariate Analysis*, 29, 163–179.

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

Heckman, J. J., H. Ichimura, and P. E. Todd (1997): "Matching Evidence Job As An Econometric Estimator: Evidence from Evaluating a Job Training Programme," *Review of Economic Studies*, 64, 605–654.

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

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

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

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

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

Johnston, G. J. (1982): "Probabilities of maximal deviations for nonparametric regression function estimates," *Journal of Multivariate Analysis*, 12, 402–414.

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

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

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

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

Lee, D. S. and T. Lemieux (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 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, 807–828.

Liu, W. and W. B. Wu (2010): "Simultaneous nonparametric inference of time series," *The Annals of Statistics*, 38, 2388–2421.

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, 159–208.

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

Porter, J. (2003): "Estimation in the Regression Discontinuity Model," Unpublished manuscript, University of Wisconsin.

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

Schennach, S. M. (2015): "A bias bound approach to nonparametric inference," Working Paper CWP71/15, Cemmap.

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

van der Vaart, A. W. and J. A. Wellner (1996): *Weak convergence and empirical processes*, Springer.

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

	NW / LL (int)			LL (boundary)		
$\overline{h}/\underline{h}$	Unif	Tri	Epa	Unif	Tri	Epa
1.0	1.96	1.96	1.96	1.96	1.95	1.96
1.2	2.24	2.01	2.03	2.23	2.03	2.05
1.4	2.33	2.05	2.08	2.33	2.08	2.11
1.6	2.40	2.09	2.12	2.39	2.12	2.15
1.8	2.45	2.11	2.15	2.44	2.16	2.19
2	2.48	2.14	2.17	2.48	2.18	2.22
3	2.60	2.22	2.27	2.60	2.27	2.32
4	2.66	2.26	2.31	2.66	2.32	2.37
5	2.70	2.30	2.35	2.71	2.35	2.41
6	2.73	2.32	2.37	2.73	2.37	2.43
7	2.75	2.34	2.39	2.76	2.39	2.45
8	2.77	2.35	2.41	2.78	2.41	2.47
9	2.79	2.37	2.42	2.79	2.43	2.48
10	2.80	2.38	2.44	2.81	2.44	2.50
20	2.89	2.45	2.51	2.89	2.52	2.58
50	2.97	2.53	2.59	2.98	2.60	2.66
100	3.02	2.57	2.64	3.02	2.65	2.71

Table 1: Critical values $c_{0.95}(\overline{h}/\underline{h}, k)$ for level-5% tests 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. "NW / LL (int)" refers to Nadaraya-Watson (local constant) regression in the interior or at a boundary, as well as local linear regression in the interior. "LL (boundary)" refers to local linear regression at a boundary (including regression discontinuity designs).

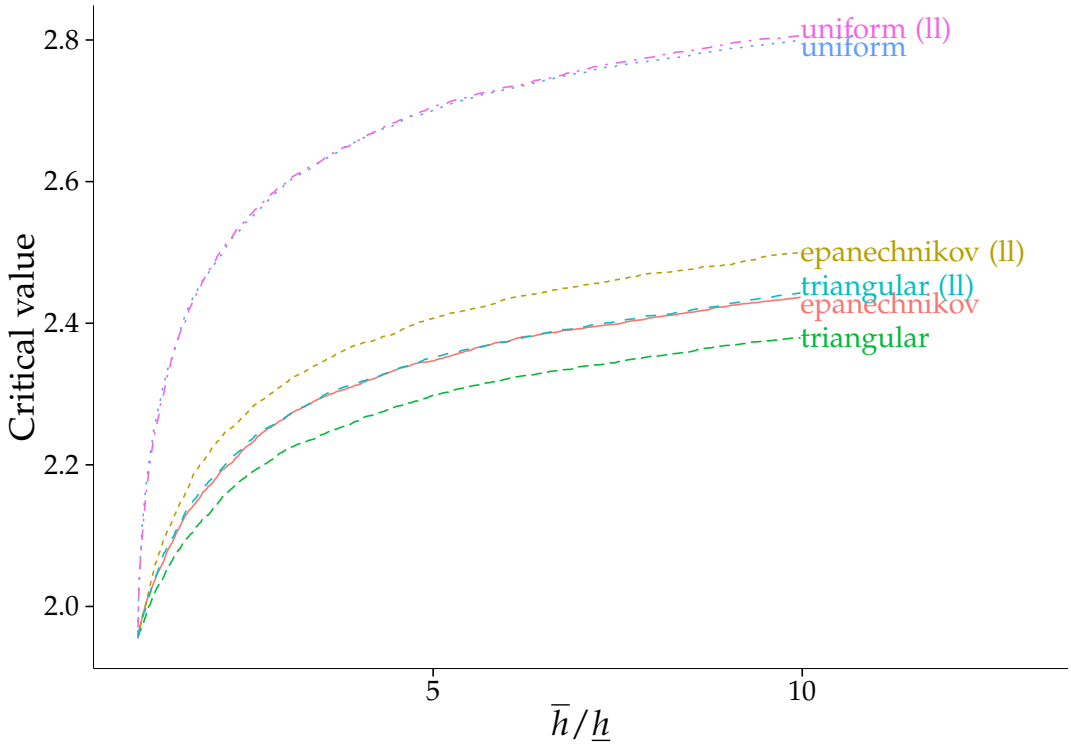


Figure 1: Two-sided 95% critical values for different kernels. "uniform", "triangular" and "epanechnikov" refer to Nadaraya-Watson (local constant) regression in the interior or at a boundary as well as local linear regression in the interior. "uniform (ll)", "triangular (ll)" and "epanechnikov (ll)" refer to local linear regression at a boundary.

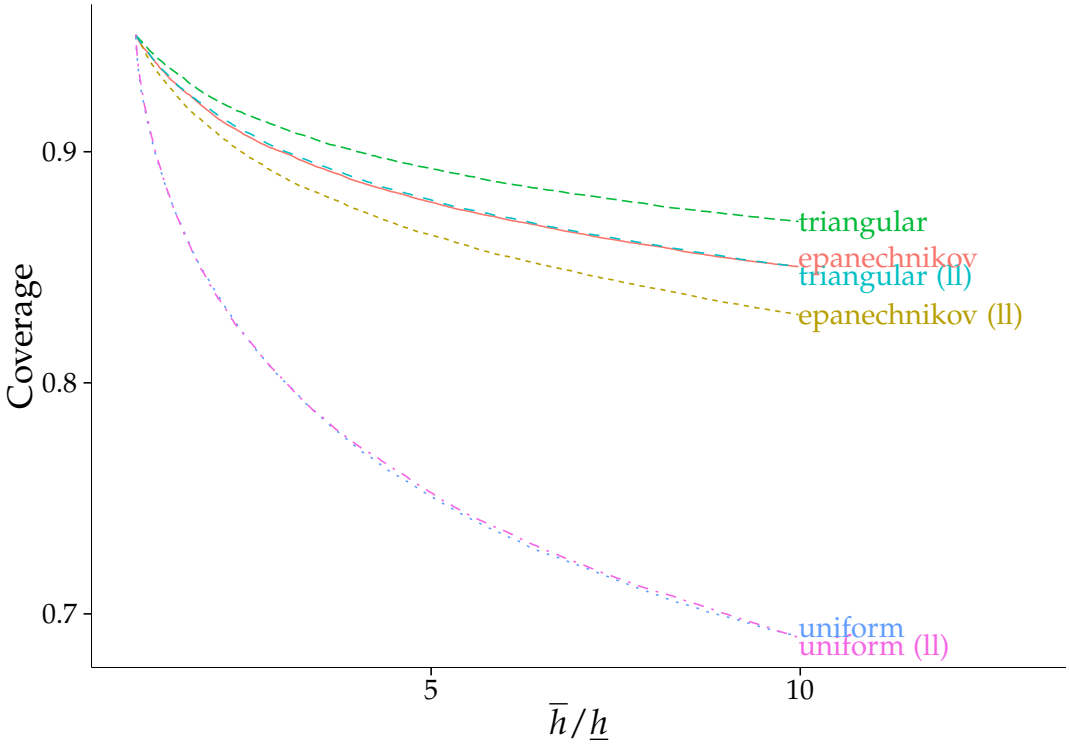
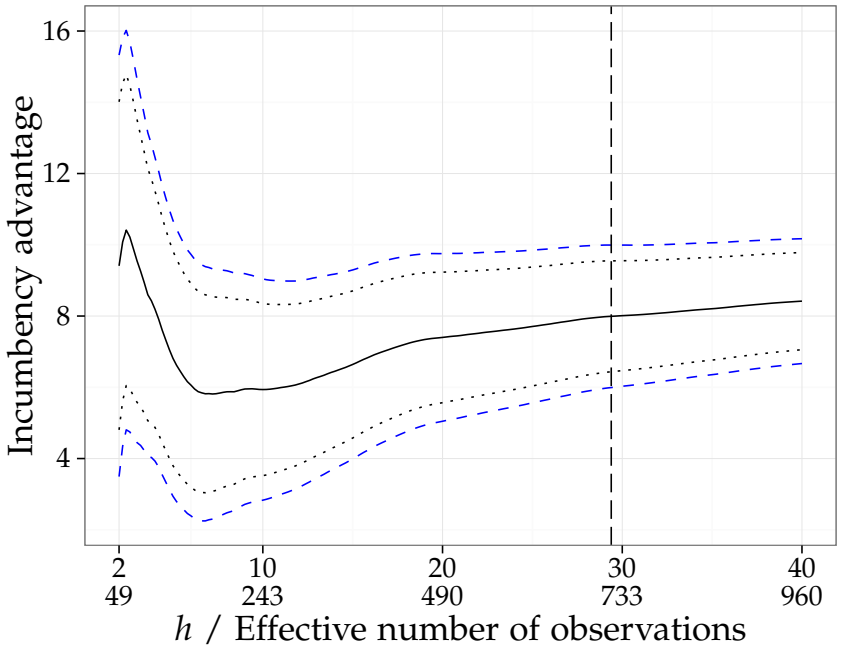


Figure 2: Coverage of unadjusted 95% confidence bands (i.e. using critical values equal to 1.96) for different kernels. "uniform", "triangular" and "epanechnikov" refer to Nadaraya-Watson (local constant) regression in the interior or at a boundary as well as local linear regression in the interior. "uniform (ll)", "triangular (ll)" and "epanechnikov (ll)" refer to local linear regression at a boundary.

(a) Local linear regression



(b) Local quadratic regression

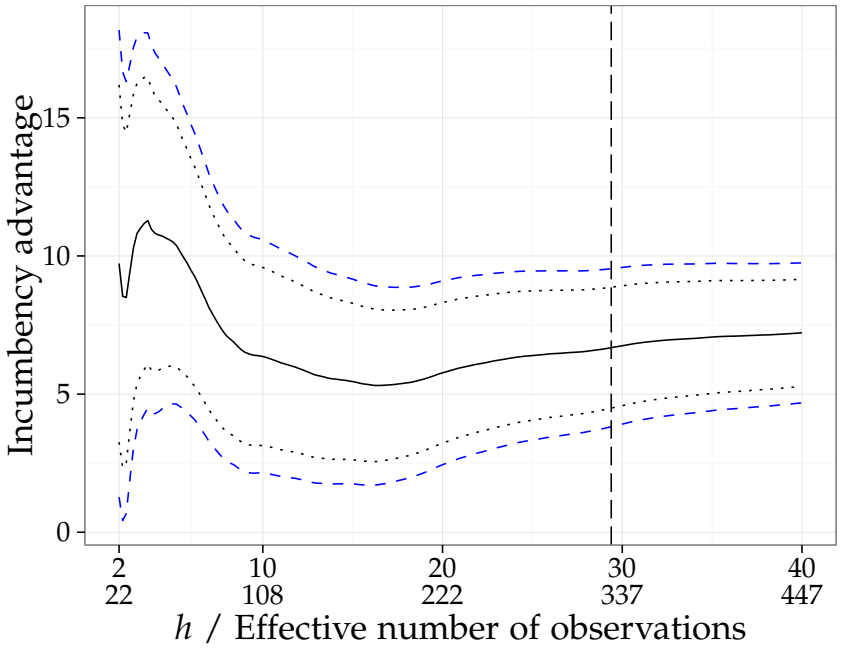


Figure 3: Effect of incumbency on percentage vote share in the next election. Data are from Lee (2008). Local linear (panel (a)) and local quadratic (panel (b)) regression with triangular kernel. Point estimate $\hat{\theta}(h)$ (solid line), pointwise (dotted), and uniform (dashed) confidence bands 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 2.52 (local linear) and 2.56 (local quadratic). Vertical dashed line corresponds to estimates using Imbens and Kalyanaraman (2012) bandwidth.

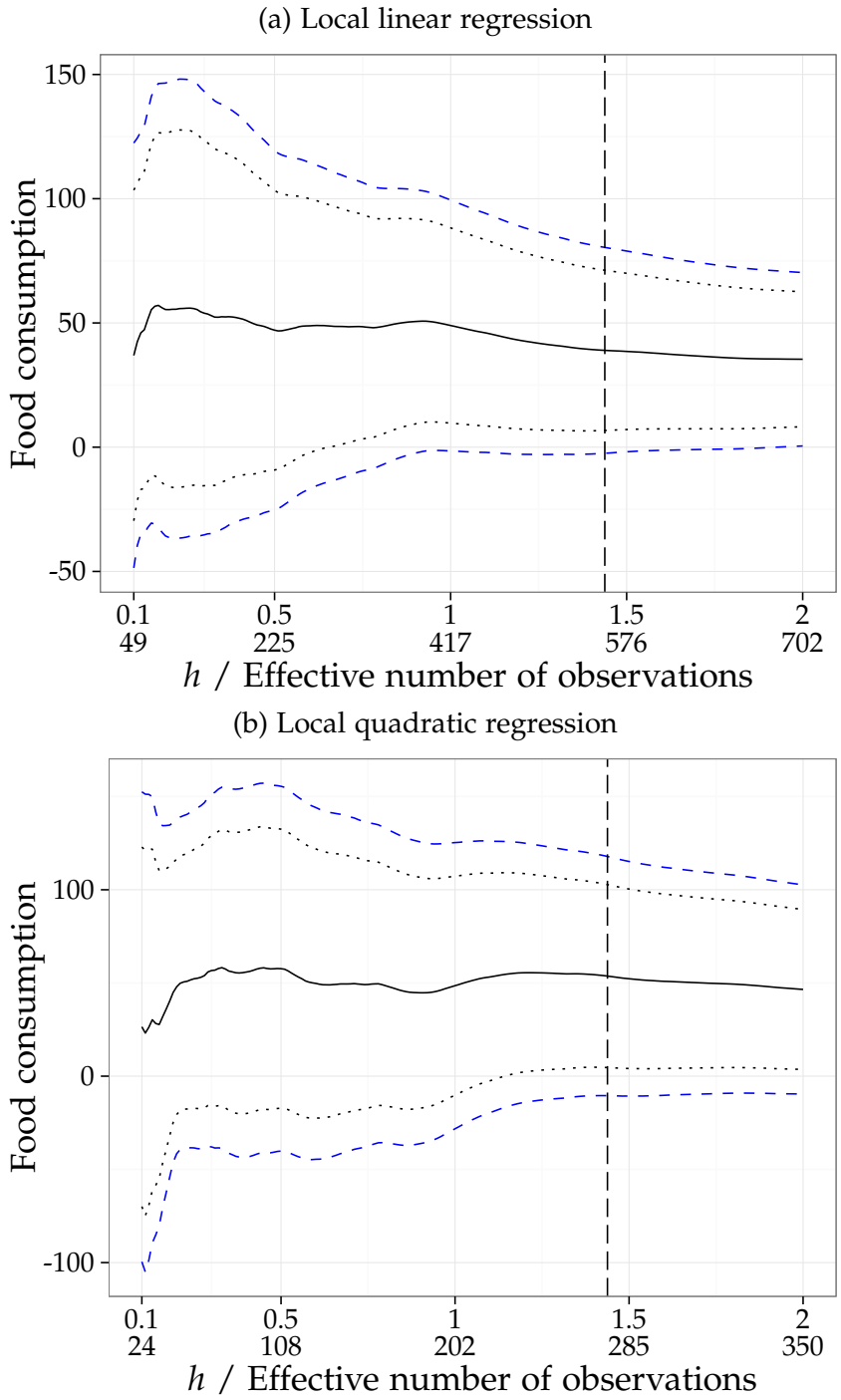


Figure 4: Effect of the Oportunidades cash transfer program on food consumption. Data are from Calonico et al. (2014). Local linear (panel (a)) and local quadratic (panel (b)) regression with triangular kernel. Point estimate $\hat{\theta}(h)$ (solid line), pointwise (dotted), and uniform (dashed) confidence 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 2.52 (local linear) and 2.56 (local quadratic). Vertical dashed line corresponds to estimates using Imbens and Kalyanaraman (2012) bandwidth.

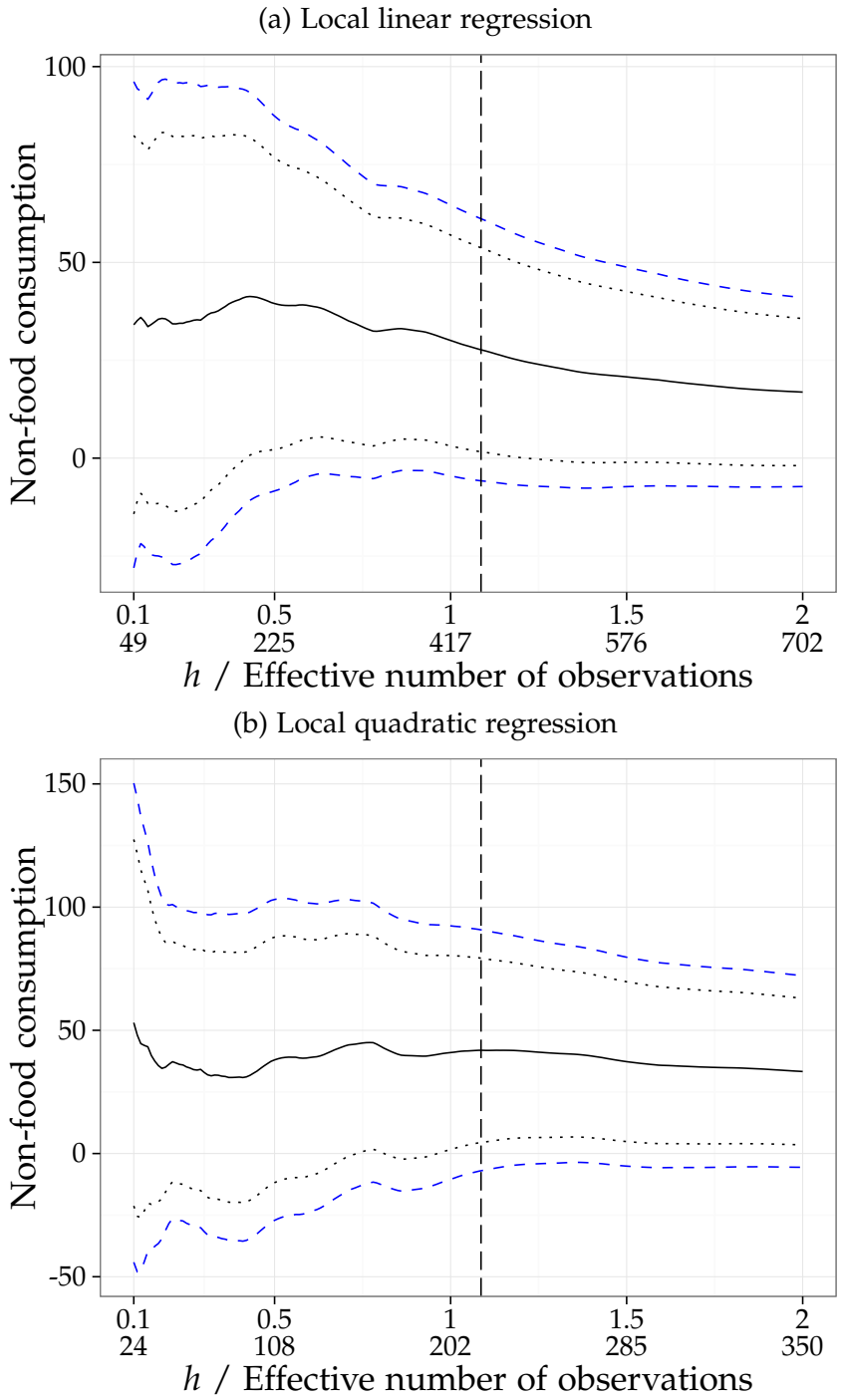


Figure 5: Effect of the Oportunidades cash transfer program on non-food consumption. Data are from Calonico et al. (2014). Local linear (panel (a)) and local quadratic (panel (b)) regression with triangular kernel. Point estimate $\hat{\theta}(h)$ (solid line), pointwise (dotted), and uniform (dashed) 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 2.52 (local linear) and 2.56 (local quadratic). Vertical dashed line corresponds to estimates using Imbens and Kalyanaraman (2012) bandwidth.

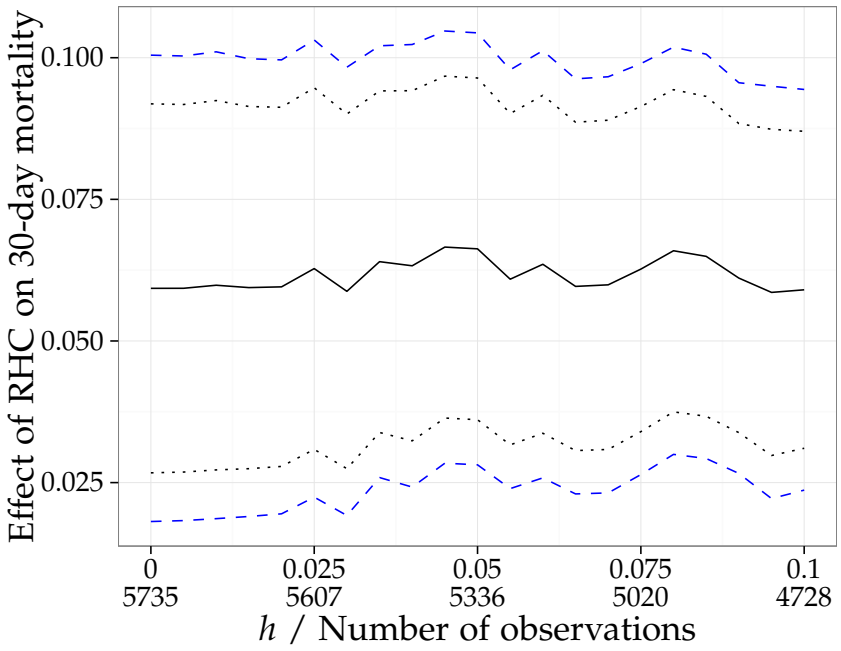


Figure 6: Effect of Right Heart Catheterization on 30-day morality. Data are fromConnors Jr et al. (1996). Point estimate $\hat{\theta}(h)$ (solid line), pointwise (dotted), and uniform (dashed) confidence bands as function of the trimming parameter h, plotted over the range $h \in [0, 0.1]$.