The Annals of Statistics
1991, Vol. 19, No. 2, 778–796

BOOTSTRAP SIMULTANEOUS ERROR BARS FOR NONPARAMETRIC REGRESSION[1]

BY W. HÄRDLE AND J. S. MARRON

Université Catholique de Louvain and Universität Bonn and Universität Bonn

Simultaneous error bars are constructed for nonparametric kernel estimates of regression functions. The method is based on the bootstrap, where resampling is done from a suitably estimated residual distribution. The error bars are seen to give asymptotically correct coverage probabilities uniformly over any number of gridpoints. Applications to an economic problem are given and comparison to both pointwise and Bonferroni-type bars is presented through a simulation study.

1. Motivation. Regression smoothing is an effective method for estimation of mean curves in a flexible nonparametric way. Since this technique makes no structural assumptions on the underlying curve, it is very important to have a device for understanding when observed features are significant. A question often asked in this context is whether or not an observed peak or valley is actually a feature of the underlying regression function or is only an artifact of the observational noise. For such issues, confidence intervals should be used that are simultaneous (i.e., uniform over location) in nature. This paper proposes and analyzes a method of obtaining any number of simultaneous error bars at a grid of points. The method is simple to implement and does not rely on the evaluation of quantities which appear in asymptotic distributions. The construction is based on a residual resampling technique which models the conditional error distribution and also takes the bias properly into account (at least asymptotically).

For an understanding of these ideas, consider Figure 1. Figure 1a shows a scatter plot of the expenditure for potatoes as a function of income for the year 1973, from the Family Expenditure Survey (1968–1983). Figure 1b shows a nonparametric regression estimate which was obtained by smoothing the point cloud, using the kernel algorithm described in Section 2. As a means of understanding the variability in the kernel smooth, Figure 1b also shows error bars, i.e., vertical confidence intervals constructed by the bootstrap method proposed in Section 2. These bars are estimated simultaneous 80% confidence intervals. Note that the error bars are longer on the right-hand side, which reflects the fact that there are fewer observations there and hence more uncertainty in the curve estimate. The error bars are asymmetric in particular

Received March 1989; revised March 1990.
[1]Research supported by Deutsche Forschungsgemeinschaft, SFB 303.
AMS 1980 *subject classifications*. Primary 62G05; secondary 62G99.
Key words and phrases. Bootstrap, error bars, kernel smoothing, nonparametric regression, variability bound.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*.
www.jstor.org

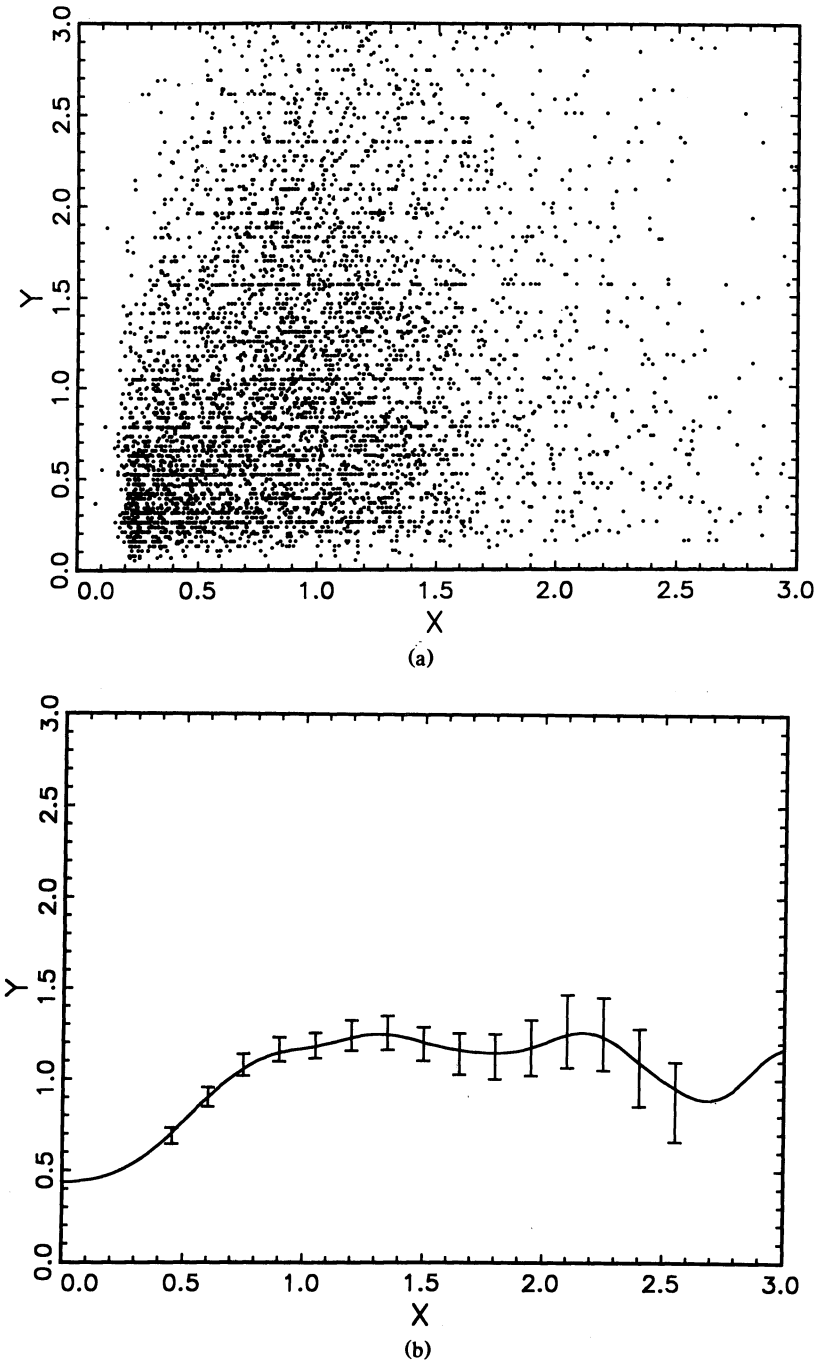

FIG. 1. *Expenditure for potato vs. income* (a) *scatter plot* (b) *regression kernel smooth* (*quartic kernel with band with* $h = 0.3$) *and errors bars.*

at points with high curvature which reflects the correct centering of the bars by a bias term.

Bierens and Pott-Buter (1987) derived variability bands with pointwise coverage probability for a related question in demand theory. Clearly there is a need for effective simultaneous error bars in all applications of nonparametric regression. Hall and Titterington (1988) constructed a confidence band for calibration of radio carbon dating. Knafl, Sacks and Ylvisaker (1985) derived uniform variability bands under the assumption of a Gaussian error structure.

The use of bootstrap methods for assessing variability bands in nonparametric regression was to our knowledge first suggested by McDonald (1982). There are several ways of bootstrapping in the context of nonparametric smoothing. The interactive method used by McDonald was based on resampling from the empirical distribution of the pairs of observations. This approach has also been investigated by Dikta (1988) who showed that, up to a bias term, a type of pointwise bootstrap confidence interval is asymptotically correct. If the predictor variables are fixed nonrandom values, resampling should be done from estimated residuals as has been argued by Bickel and Freedman (1981) in the setting of linear regression. Härdle and Bowman (1988) applied this resampling scheme to the nonparametric regression procedure, also in the case of random predictor variables on estimated residuals. This form of bootstrapping preserves the error structure in the data and guarantees that the bootstrap observations have errors with mean zero. There are two main advantages to this approach. First, it correctly accounts for the bias and hence does not require additional estimation of bias or the use of a suboptimal (undersmoothed) curve estimator. Second, no assumption of homoscedasticity is required; the method automatically adapts to different residual variances at different locations.

The resampled data is smoothed to give an approximation to the simultaneous distribution of the estimator at a grid of points. This distribution can either be used directly to obtain simultaneous error bars, or a simple Bonferroni approach can be used. We also study methods for generating bars which are based on groups of gridpoints. This approach provides a general framework, which includes the direct and Bonferroni methods as extremes.

In Section 2 we give a technical introduction to our method and present theorems which demonstrate the asymptotic validity of the bootstrap simultaneous errors bars. In Section 3 simulations and the previous application are discussed. We describe this economic example in more detail and do a comparison of different grids of error bars through simulation. The simulations indicate that handling the bias is the most difficult aspect of this problem, especially when the regression function has substantial curvature. The analysis of Section 3 provides a quantification of this difficulty. For this reason, in the examples we considered, 80% error bars had actual coverage as poor as 50–65%. In Section 4 we give proofs of the theorems in Section 2.

2. Bootstrap error bars. Stochastic design nonparametric regression is based on observations $\{(X_i, Y_i)\}_{i=1}^n \in \mathbb{R}^{d+1}$ and the goal is to estimate $m(x) = E(Y|X = x)$: $\mathbb{R}^d \to \mathbb{R}$. The form of the kernel regression estimator, developed

by Nadaraya (1964) and Watson (1964) is

$$\hat{m}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{K_h(x - X_i)Y_i}{\hat{f}_h(x)}, \tag{2.1}$$

where

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \tag{2.2}$$

and where $K_h(u) = h^{-d}K(u/h)$ is a kernel weight function with bandwidth h. All results of this paper are stated in terms of this estimator, although the essential ideas clearly extend to other types of kernel estimators such as those of Gasser and Müller (1984) and also to other regression estimators, such as spline methods, as discussed in Eubank (1988).

The choice of the bandwidth is crucial to the performance of the estimator. An asymptotic analysis of this choice and discussion of various data based bandwidth selectors may be found in Chapters 4 and 5 of Härdle (1989). The results of the present paper are formulated in such a way as to allow this type of objective bandwidth choice to be employed.

One approach to the problem of finding simultaneous error bars would be to work with limiting normal distributions of the estimator at the grid points. However, the joint distribution of the estimator at these gridpoints has substantial positive correlation, which makes the derivation of joint normal theory confidence intervals nontrivial. In fact, they essentially should be done by simulation methods. Since simulation methods are needed anyway, it seems better to use a more direct approach through bootstrapping, as opposed to relying on the normal approximation and also to facing the problems of parameter estimation that such an approach entails.

While bootstrap methods are well-known tools for assessing variability, more care must be taken to properly account for the type of bias encountered in nonparametric curve estimation. In particular, the naive bootstrap approach of resampling from the pairs $\{(X_i, Y_i)\colon i = 1, \ldots, n\}$ is inappropriate because the bootstrap bias will be 0. Our approach to this problem is to first use the estimated residual

$$\hat{\varepsilon}_i = Y_i - \hat{m}_h(X_i). \tag{2.3}$$

The essential idea is to resample from the estimated residuals, which are the differences between the observations and the pilot estimate and then use this data to construct an estimator whose distribution will approximate the distribution of the original estimator.

To better retain the conditional distributional characteristics of the estimate, we do not resample from the entire set of residuals, as in Härdle and Bowman (1988). One possibility would be to resample from a set of residuals determined by a window function, but this has the disadvantage of requiring choice of the window width. To avoid this we use the idea of *wild bootstrapping*, as proposed in Härdle and Mammen (1989) [but see Rosenblueth (1975) for access to related literature], where each bootstrap residual is drawn from the two-point distribution which has mean zero, variance equal to the square

of the residual and third moment equal to the cube of the residual. In particular define a new random variable ε_i^* having a two-point distribution $\hat{G}_i$, where $\hat{G}_i = \gamma\delta_a + (1-\gamma)\delta_b$ is defined through the three parameters a, b, γ, and where δ_a, δ_b denote point measures at a, b, respectively. Some algebra reveals that the parameters a, b, γ at each location X_i are given by $a = \hat{\varepsilon}_i(1-\sqrt{5})/2$, $b = \hat{\varepsilon}_i(1+\sqrt{5})/2$ and $\gamma = (5+\sqrt{5})/10$. These parameters ensure that $E\varepsilon^* = 0$, $E\varepsilon^{*2} = \hat{\varepsilon}_i^2$ and $E\varepsilon^{*3} = \hat{\varepsilon}_i^3$. In a certain sense the resampling distribution $\hat{G}_i$ can be thought of as attempting to reconstruct the distribution of each residual through the use of one single observation. Therefore it is called the wild bootstrap. It is actually the cumulative effect of all these residuals that is used in the generation of the simultaneous error bars. The above formulation of the wild bootstrap, based on a two-point distribution, is only one possible approach. Other distributions could be considered as well and an interesting question for further work is finding whether some will give better performance. See Section 7 of Wu (1986) for some closely related ideas in linear regression.

After resampling, new observations

$$Y_i^* = \hat{m}_g(X_i) + \varepsilon_i^* \tag{2.4}$$

are defined, where $\hat{m}_g(x)$ is a kernel estimator with bandwidth g taken to be larger than h (a heuristic explanation of why it is essential to oversmooth g is given later). Then the kernel smoother (2.1) is applied to the bootstrapped data $\{(X_i, Y_i^*)\}_{i=1}^n$ using bandwidth h. Let $\hat{m}_h^*(x)$ denote this kernel smooth. A number of replications of $\hat{m}_h^*(x)$ can be used as the basis for simultaneous error bars because the distribution of $\hat{m}_h(x) - m(x)$ is approximated by the distribution of $\hat{m}_h^*(x) - \hat{m}_g(x)$, as Theorem 1 shows.

Here and in the following, to help keep the various probability structures straight, we use the symbol $Y|X$ to denote the conditional distribution of $Y_1, \ldots, Y_n | X_1, \ldots, X_n$ and the symbol $*$ to denote the bootstrap distribution of $Y_1^*, \ldots, Y_n^* | (X_1, Y_1), \ldots, (X_n, Y_n)$.

For an intuitive understanding of why the bandwidth g used in the construction of the bootstrap residuals should be oversmoothed, consider the means of $\hat{m}_h(x) - m(x)$ under the $Y|X$-distribution and $\hat{m}_h^*(x) - \hat{m}_g(x)$ under the $*$-distribution in the simple situation when the marginal density $f(x)$ is constant in a neighborhood of x. Asymptotic analysis as in Rosenblatt (1969) shows that

$$E^{Y|X}\big(\hat{m}_h(x) - m(x)\big) \approx h^2\left(\int u^2 K/2\right) m''(x).$$

$$E^*\big(\hat{m}_h^*(x) - \hat{m}_g(x)\big) \approx h^2\left(\int u^2 K/2\right) \hat{m}_g''(x).$$

Hence for these two distributions to have the same bias, we need $\hat{m}_g''(x) \to m''(x)$. This requires choosing g tending to zero at a rate slower than the *optimal bandwidth* h for estimating $m(x)$, see Gasser and Müller (1984).

There are several ways to use the bootstrap approximation to understand the variability in $\hat{m}_h(x)$. We prefer a finite set of error bars instead of a continuous band because for a reasonably dense collection (as in Figure 1b), there is little information lost and the bar approach is much easier to compute and also to analyze. The simplest is to calculate pointwise $1 - \alpha$ confidence intervals, but these will then not be simultaneous in nature. A naive way of extending pointwise intervals to M simultaneous confidence intervals is by applying the Bonferroni method, which is to correct the significance level by the number of locations at which the error bars are to be constructed. This involves first finding M pointwise intervals with confidence coefficient $1 - \alpha/M$. Then by the Bonferroni inequality, the collection of these intervals will have simultaneous confidence coefficient at least $1 - \alpha$. A drawback to the Bonferroni approach is that the resulting intervals will quite often be too long. The reason is that this method does not make use of the substantial positive correlation of the curve estimates at nearby points.

A more direct approach to finding simultaneous error bars is to consider the simultaneous coverage on pointwise error bars and then adjust the pointwise level to give a simultaneous coverage probability of $1 - \alpha$. Note that there are also many other ways to obtain simultaneous error bars, but this has the compelling feature of assigning equal size (in the confidence interval sense) to each bar.

A general framework, which includes both the Bonferroni and direct methods, can be formulated by thinking in terms of groups of grid points. First partition the set of locations where error bars are to be computed into M groups. Suppose the groups are indexed by $j = 1, \ldots, M$ and the locations within each group are denoted by $x_{j,k}$, $k = 1, \ldots, N_j$. The groups should be chosen so that for each j, the $x_{j,k}$ values in each group are within $2h$ of each other. The reason for this is that when the x values are further than $2h$ apart, the estimates are independent and independent theory simultaneous error bars are quite close to those derived from Bonferroni theory (this can be seen, for example, by calculating the lengths of independent theory and Bonferroni theory intervals for standard normal random variables, which turn out to be typically within about 3% of each other). In the one-dimensional case this is easily accomplished by dividing the x-axis into intervals of length roughly $2h$. The asymptotics given later are based on the assumption that the number of x's in each group does not change with n. More precisely, the set of grid points $x_{j,k}$, $k = 1, \ldots, N_j$ has the same asymptotic relative location c_k (not depending on n) to some reference point $x_{j,0}$ in each group j. Therefore define

$$x_{j,k} = c_k h + x_{j,0}, \qquad k = 1, \ldots, N_j. \tag{2.5}$$

In the multidimensional case, the simplest formulation is to have each group lying in a hypercube with length $2h$. Now within each group j we use the bootstrap replications to approximate the joint distribution of

$$\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x}) = \left\{\hat{m}_h(x_{j,k}) - m(x_{j,k}) : k = 1, \ldots, N_j\right\}.$$

Next we state a theorem which shows that the bootstrap works for the set of locations within each group. For notational convenience we suppress the dependence on j. Technical assumptions are:

ASSUMPTION 1. $m(x)$, $f(x)$ and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$ are twice continuously differentiable.

ASSUMPTION 2. The kernel function K is symmetric and nonnegative, $c_K = \int K^2 < \infty$ and $d_K = \int u^2 K(u)\,du < \infty$.

ASSUMPTION 3. $\sup_x E(\varepsilon^3|X = x) < \infty$.

ASSUMPTION 4. $f(x_0) \geq \eta > 0$.

Under Assumptions 1 and 2, reasonable choice of h will be in the set

$$H_n = \left[\underline{c}n^{-1/(4+d)}, \bar{c}n^{-1/(4+d)}\right], \qquad 0 < \underline{c} < \bar{c} < \infty.$$

For this choice of bandwidth, the kernel smoother $\hat{m}_h(\underset{\sim}{x})$ is asymptotically optimal, see Section 5.1 of Härdle (1989). This assumption is not restrictive because, for $\underline{c}$ and $\bar{c}$ reasonably small and large, respectively, it will be satisfied with probability tending to 1 if h is chosen by cross-validation, for example, see Härdle, Hall and Marron (1988). The exact specification of the rate of convergence of g is less important for the validity of the following theorem, although it must tend to zero at a rate slower than h. Hence it is assumed that g is chosen from the set

$$G_n = [n^{-1/(4+d)+\delta}, n^{-\delta}], \qquad \delta > 0.$$

A fine tuning of the choice of bandwidth g is presented in Theorem 3.

THEOREM 1. *Given the previous assumptions, we have along almost all sample sequences and for all* $\underset{\sim}{z} \in \mathbb{R}^N$,

$$\sup_{h\in H_n} \sup_{g\in G_n} \left| P^{Y|X}\left\{\sqrt{nh^d}\left[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})\right] < \underset{\sim}{z}\right\} - P^*\left\{\sqrt{nh^d}\left[\hat{m}_h^*(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})\right] < \underset{\sim}{z}\right\}\right| \to 0.$$

Note that our assumption on the speed of the bandwidth h ensures that each of the previous probabilities has a nontrivial limit. In fact, the proof of the theorem comes from showing that both $\sqrt{nh^d}[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})]$ and $\sqrt{nh^d}[\hat{m}_h^*(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})]$ have the same limiting normal distribution. The reason that uniform convergence (in h and g) in the previous result is important is that it ensures that the result still holds when h or g are replaced by random data driven bandwidths. For each group j this joint distribution is used to obtain simultaneous $1 - \alpha/M$ error bars that are simultaneous over $k = 1, \ldots, N_j$ as follows. Let $\beta > 0$ denote a generic size for individual confidence

intervals. Our goal is to choose β so that the resulting simultaneous size is $1 - \alpha/M$. For each $x_{j,k}$, $k = 1, \ldots, N_j$, define the interval $I_{j,k}(\beta)$ to have endpoints which are the $\beta/2$ and the $1 - \beta/2$ quantiles of the $(\hat{m}_h^*(x_{j,k}) - \hat{m}_g(x_{j,k}))$ distribution. Then define α_β to be the empirical *simultaneous* size of the β confidence intervals, i.e., the proportion of curves which lie outside at least one of the intervals in the group j. Next find the value of β, denoted by β_j, which makes $\alpha_{\beta_j} = \alpha/M$. The resulting β_j intervals within each group j will then have confidence coefficient $1 - \alpha/M$. Hence by the Bonferroni bound, the entire collection of intervals $I_{j,k}(\beta_j)$, $k = 1, \ldots, N_j$, $j = 1, \ldots, M$ will simultaneously contain at least $1 - \alpha$ of the distribution of $\hat{m}_h^*(x_{j,k})$ about $\hat{m}_g(x_{j,k})$. Thus the intervals $I_{j,k}(\beta_j) - \hat{m}_g(x_{j,k}) + \hat{m}_h(x_{j,k})$ will be simultaneous confidence intervals with confidence coefficient at least $1 - \alpha$. The result of this process is summarized as:

THEOREM 2. *Define M groups of locations $x_{j,k}$, $k = 1, \ldots, N_j$, $j = 1, \ldots, M$, where simultaneous error bars are to be established. Compute uniform confidence intervals for each group. Correct the significance level across groups by the Bonferroni method. Then the bootstrap error bars establish asymptotic simultaneous confidence intervals, i.e.,*

$$\lim_{n \to \infty} P\{m(x_{j,k}) \in I_{j,k}(\beta_j) - \hat{m}_g(x_{j,k}) + \hat{m}_h(x_{j,k}), \quad k = 1, \ldots, N_j, j = 1, \ldots, M\} \geq 1 - \alpha. \tag{2.6}$$

As a practical method for finding β_j for each group j, we suggest the following halving approach (also called a bisection search). In particular, first try $\beta = \alpha/2M$ and calculate α_β. If the result is more than α/M, then try $\beta = \alpha/4M$, otherwise next try $\beta = 3\alpha/4M$. Continue this halving approach unit neighboring (since only finitely many bootstrap replications are made, there is only a finite grid of possible β's available) values β_* and β^* are found so that $\alpha_{\beta_*} < \alpha/M < \alpha_{\beta^*}$. Finally, take a weighted average of the β_* and the β^* intervals where the weights are $(\alpha_{\beta^*} - \alpha/M)/(\alpha_{\beta^*} - \alpha_{\beta_*})$ and $(\alpha/M - \alpha_{\beta_*})/(\alpha_{\beta^*} - \alpha_{\beta_*})$, respectively.

Note that Theorem 2 contains, as a special case, the asymptotic validity of both the Bonferroni and the direct simultaneous error bars. Bonferroni is the special case $N_1 = \cdots = N_M = 1$ and the direct method is where $M = 1$.

The previous theorems require that M, the number of neighborhoods, remain constant with respect to n. The reason is that otherwise, the Bonferroni method of combining across neighborhoods, will require the significance level for each neighborhood to tend to zero. This means we could no longer apply Theorem 1, because it is formulated in terms of fixed $\underset{\sim}{z}$. An interesting direction for further work would be to investigate a suitable analogue of Theorem 1, which would allow M to grow. The neighborhood approach should be very useful here because only M need grow, not N.

The next issue is how to fine tune the choice of the pilot bandwidth g. While it is true that the bootstrap works (in the sense of giving asymptotically correct

coverage probabilities) with a rather crude choice of g, it is intuitively clear that specification of g will play a role in how well it works for finite samples. Since the main role of the pilot smooth is to provide a correct adjustment for the bias, we use the goal of bias estimation as a criterion. We think theoretical analysis of the previous type will be more straightforward than allowing the N_j to increase, which provides further motivation for considering this general grouping framework.

In particular, recall that the bias in the estimation of $m(x)$ by $\hat{m}_h(x)$ is given by

$$b_h(x) = E^{Y|X}\hat{m}_h(x) - m(x).$$

The bootstrap bias of the estimator constructed from the resampled data is

$$\begin{aligned}\hat{b}_{h,g}(x) &= E^*[\hat{m}_h^*(x)] - \hat{m}_g(x) \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{K_h(x - X_i)\hat{m}_g(X_i)}{\hat{f}_h(x)} - \hat{m}_g(x).\end{aligned}$$

The following theorem gives an asymptotic representation of the mean square error for the problem of estimating $b_h(x)$ by $\hat{b}_{h,g}(x)$. It is then straightforward to find g to minimize this representation. Such a choice of g will make the means of the $Y|X$ and * distributions close to each other.

For notational simplicity, we state this result explicitly only for the case $d = 1$. Extension to general d is straightforward, but messy, because the derivatives need to be replaced by sums of partial derivatives. In addition to the technical assumptions required for Theorem 1, we also need:

ASSUMPTION 5. m and f are four times continuously differentiable.

ASSUMPTION 6. K is twice continuously differentiable.

THEOREM 3. *Under Assumptions* 1–6, *along almost all sample sequences,*

$$E\left[\left(\hat{b}_{h,g}(x) - b_h(x)\right)^2 \middle| X_1, \ldots, X_n\right] \sim h^4\left[C_1 n^{-1} g^{-5} + C_2 g^4\right], \tag{2.7}$$

in the sense that the ratio tends in probability to 1, *where*

$$C_1 = \int \frac{(K'')^2((1/2)d_K)^2\sigma^2(x)}{f(x)},$$

$$C_2 = \frac{((1/2)d_K)^4\left[(mf)^{(4)} - (mf'')''\right](x)^2}{f(x)^2}.$$

An immediate consequence of Theorem 3 is that the rate of convergence for $d = 1$ of g should be $n^{-1/9}$. This makes precise the previous intuition which indicated that g should be slightly oversmoothed. In addition, under these assumptions, reasonable choices of h will be of the order $n^{-1/5}$. Hence, (2.7)

shows once again that g should tend to zero more slowly than h. Note that unlike the previous results, Theorem 3 is not stated uniformly over h. The reason is that we are only trying to give some indication of how the pilot bandwidth g should be selected. Note also that Theorem 3 applies only to the mean of the distributions, when a better choice of g would probably take into account other distributional aspects as well. For example, some preliminary calculations along this line show that the effect of g on the variances is of the same order as the effect on the mean. We do not choose to pursue this further, because deeper analysis appears quite complicated and seems too tangential to the points we are trying to make in this paper.

All of the results in this paper have been stated in terms of the so-called stochastic design model where the regressors X are thought of as realizations of random variables. Since these results are all conditional on $X_1, \ldots, X_n$, our ideas carry over immediately to the case where the X's are fixed and chosen by the experimenter.

In the case of binary regression [dose-response curves, Cox (1970), page 8], where the response variable Y takes on only the values 0 or 1, there are more natural ways of obtaining bootstrap confidence intervals than those described here. A direct application of our method would give bootstrapped data Y^* which take on values different from 0 and 1. A seemingly more natural approach would be to bootstrap from a Bernoulli distribution with parameter $\hat{m}_g(X_i)$.

3. Simulations and application. In this section we consider three main points. The first is investigation of how much practical difference there is between pointwise, simultaneous and Bonferroni confidence intervals. Second, we compute the coverage probabilities of the bootstrap confidence intervals, introduced in Section 2, in several simulation settings. Third, we give further details concerning the example considered in Section 1.

To study the practical difference between the various types of error bars, we consider the distribution of $\hat{m}_h(x) - m(x)$ at a grid of x values for some specific examples. We chose the underlying curve to be $m(x) = x + 4e^{-2x^2}/\sqrt{2\pi}$. To see what this looks like, consider Figure 2. The solid curve in each part of Figure 2 is this $m(x)$. This form is both convenient to work with when calculating various constants, and also is challenging for the methodology, because the hump is an interesting feature to be detected.

We chose the marginal distribution of X to be $N(0, 1)$ and took the conditional distribution of $Y|X$ to be $N(m(X), \sigma^2)$, for $\sigma = 0.3, 0.6, 1, 1.5$. For each of these four distributions, 200 observations were generated.

To study the differences between the various error bars, for each setting, 500 pseudodata sets were generated. Then we calculated kernel estimates, at the points $x = -2, -1.8, -1.6, \ldots, 1.8, 2$, using a standard normal density as kernel. The bandwidth was chosen to be h_0 as previously discussed. Figure 2 shows, for the $\sigma = 1$ distribution, $m(x)$ overlayed with error bars whose endpoints are various types of quantiles of the distribution of $\hat{m}_h(x)$. The centers of the error bars are at the means of these distributions and show

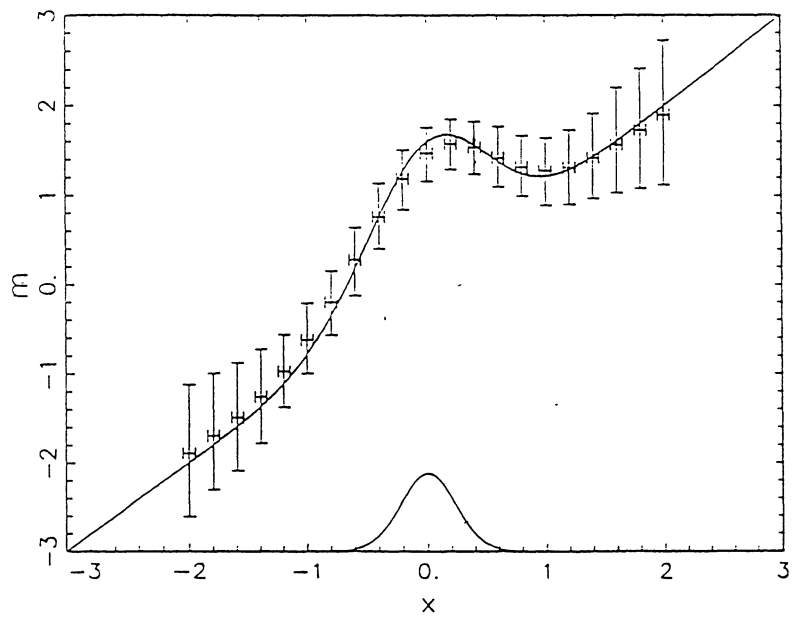

FIG. 2. *Overlay of* $m(x)$ *with empirical (from* 500 *simulation runs) quantiles of* $\hat{m}_{h_0}(x)$ *distribution. Centers of bars are means of distributions. Error bars are* 80% *simultaneous.*

clearly the bias that is inherent to nonparametric regression estimation. Note in particular how substantial bias is caused by both the curvature of $m(x)$ near the hump and by the curvature of $f(x)$, near $x = -2, 2$. The bars in Figure 2 are simultaneous bars.

For easy comparison of the lengths of these intervals with the other types, consider Figure 3. This shows, for the same x values, the lengths of the four types of bars. Of course these bars are all shorter near the center, which reflects the fact that there is more data there, so the estimates are more accurate. As expected, the lengths increase from pointwise, to actual simultaneous, to neighborhood, to Bonferroni. Also note that, as stated in Section 2, the difference between the actual simultaneous bars and the neighborhood simultaneous bars is really quite small, while the pointwise are a lot narrower. The one perhaps surprising feature is that the Bonferroni bars are not very much wider than the neighborhood bars.

To see how the bootstrap methodology proposed in Section 2 performed for the simulation settings considered here, we calculated estimates of the simultaneous coverage probabilities for 21 equally spaced error bars on $[-1, 1]$. These estimates were calculated by applying the methodology to 500 psuedo-data sets, for each of the various settings. For each data set we used 500 bootstrap replications. The pilot bandwidth g was taken to minimize a global version of the asymptotic representation given in (2.7), where the quantities that depend on x were replaced by their integral over $[-1, 1]$. The bootstrap

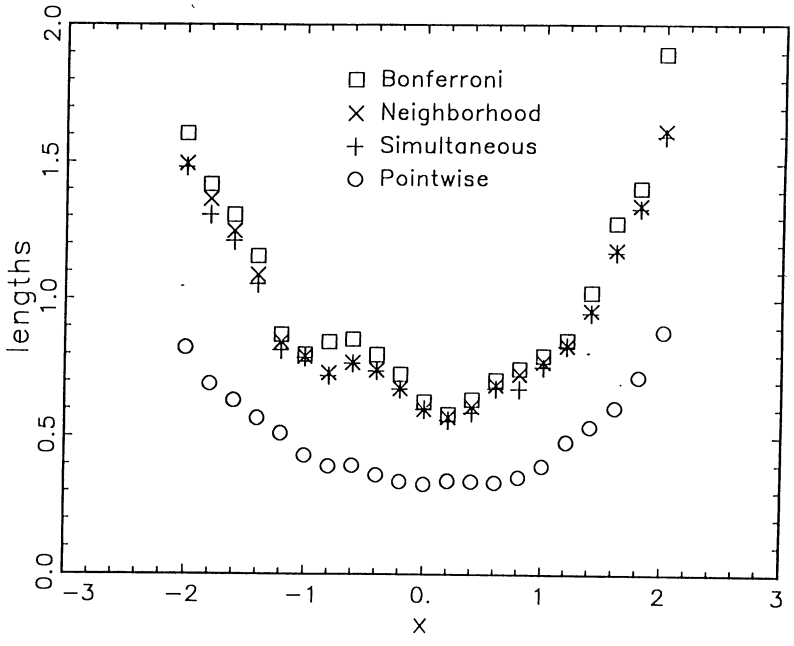

FIG. 3. *Lengths of the bars in Figure* 3, *x locations are the same.*

distribution was then used to derive the four types of error bars: pointwise, actual simultaneous, neighborhood simultaneous and Bonferroni. Then for each type of bar, the estimated simultaneous coverage probability is the proportion of times that the 500 bars cover the true curve $m(x)$ at each x value. The estimates are given in Table 1. To give an idea of the Monte Carlo variability in these estimates, also included are the radii of approximate 95% confidence intervals, of the form $1.96\sqrt{\hat{p}(1-\hat{p})}\,/\sqrt{500}$, where $\hat{p}$ is the estimated probability. Such confidence intervals are of course rather poor for $\hat{p}$

TABLE 1
Estimated (from 500 *simulation runs) coverage probabilities for bootstrap error bars*

	Pointwise	**Simultaneous**	**Neighborhood**	**Bonferroni**
$\sigma = 0.3, h = h_0$	0.03 ± 0.02	0.52 ± 0.04	0.55 ± 0.04	0.65 ± 0.04
$\sigma = 0.6, h = h_0$	0.09 ± 0.02	0.55 ± 0.04	0.59 ± 0.04	0.69 ± 0.04
$\sigma = 1.0, h = h_0$	0.10 ± 0.03	0.59 ± 0.04	0.63 ± 0.04	0.74 ± 0.04
$\sigma = 1.5, h = h_0$	0.16 ± 0.03	0.56 ± 0.04	0.65 ± 0.04	0.79 ± 0.04
$\sigma = 1.0, h = h_0/2$	0.04 ± 0.02	0.57 ± 0.04	0.60 ± 0.04	0.65 ± 0.04
$\sigma = 1.0, h = h_0$	0.10 ± 0.03	0.59 ± 0.04	0.63 ± 0.04	0.74 ± 0.04
$\sigma = 1.0, h = 2 * h_0$	0.01 ± 0.01	0.10 ± 0.03	0.16 ± 0.03	0.33 ± 0.04

close to 0, but in most cases suffice to give a decent idea of the variability involved.

This table looks somewhat disappointing since the observed coverage probabilities are all significantly below the desired value of 80%. Careful investigation revealed that this was due to problems with the estimated bias. More precisely it was caused by a systematic underadjustment in our bias correction (i.e., bias in the estimated bias adjustment). In Figure 4 the difference between the solid curve $m(x)$ and the dashed curve $E\hat{m}_h(x)$ is the true bias for our simulation setting in the case $\sigma = 0.3$, $h = h_0$. This bias is estimated for each data set by the difference between $\hat{m}_g(x)$ and $E^*\hat{m}_h^*(x)$. The bias in this estimation process is then the difference between the curve made of dots and dashes $E\hat{m}_g(x)$ and the dotted curve $E(E^*\hat{m}_h^*(x))$. Observe that because $E\hat{m}_g(x)$ has less curvature than $m(x)$, the estimated bias will typically be smaller than the actual bias. The effect does not look very large, but simultaneous coverage turns out to be a very sensitive quantity. Note that this also explains why the $h = 2 * h_0$ line of Table 1 has much smaller coverage probabilities than the others, since such a large h value means more bias than in the other settings. Of course this bias effect goes away asymptotically, but in the example considered here, Figure 4 shows that it is not negligible (and we believe this problem will exist quite often). Experiments with different values

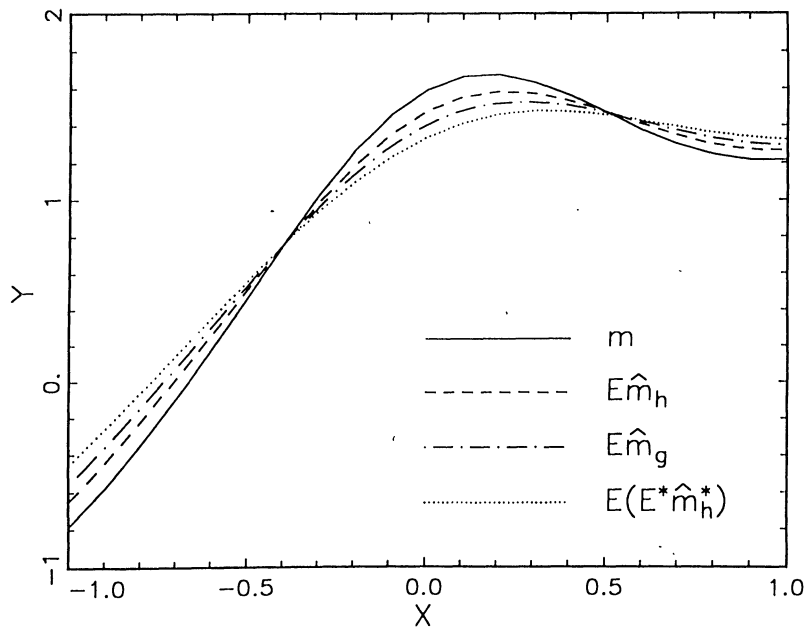

FIG. 4. *Comparison of true bias* $(E\hat{m}_h - m)$ *with expected estimated bias* $(E(E^*\hat{m}_g^*) - E\hat{m}_g)$ *for* $\sigma = 0.3$, $h = h_0$.

TABLE 2
Estimated (from 500 simulation runs) coverage probabilities for bootstrap error bars with bias correction

	Pointwise	Simultaneous	Neighborhood	Bonferroni
$\sigma = 0.3, h = h_0$	0.09 ± 0.02	0.85 ± 0.03	0.87 ± 0.03	0.94 ± 0.02
$\sigma = 0.6, h = h_0$	0.15 ± 0.03	0.83 ± 0.03	0.86 ± 0.03	0.94 ± 0.02
$\sigma = 1.0, h = h_0$	0.20 ± 0.03	0.83 ± 0.03	0.88 ± 0.03	0.94 ± 0.02
$\sigma = 1.5, h = h_0$	0.24 ± 0.04	0.82 ± 0.03	0.87 ± 0.03	0.94 ± 0.02
$\sigma = 1.0, h = h_0/2$	0.05 ± 0.02	0.87 ± 0.03	0.89 ± 0.03	0.93 ± 0.02
$\sigma = 1.0, h = h_0$	0.20 ± 0.03	0.83 ± 0.03	0.88 ± 0.03	0.94 ± 0.02
$\sigma = 1.0, h = 2 * h_0$	0.37 ± 0.04	0.79 ± 0.04	0.86 ± 0.03	0.95 ± 0.02

of g failed to alleviate this problem. An approach to the problem motivated by Figure 4 is to replace h by $c \cdot h$ for some $c > 1$ in the bias estimate. Determination of c and further analysis is beyond the scope of this paper.

To further verify that the problem here was with the bias, as indicated in Figure 4, and not with the wild bootstrap technique, we reran the simulations with the following bias adjustment. The bootstrap residuals ε_i^* were replaced by unbiased residuals ε_i^{**}, which were resampled as previously indicated, except that $\hat{\varepsilon}_i$ was replaced by $Y_i - m(x_i)$. Then the bootstrap data Y_i^* was replaced by unbiased data $Y_i^{**} = m(x_i) + \varepsilon_i^{**}$. Table 2 shows the resulting coverage probabilities.

Observe that now most of the coverage probabilities for the simultaneous bars are essentially 80%, with those that are off being slightly larger. This indicates that if the previously discussed bias problem did not exist, then the bootstrap methodology proposed here would give very slightly conservative performance (i.e., error bars too wide) for the example we have considered. Note that as expected from the previous analysis, the neighborhood bars exhibit coverage probabilities which are slightly bigger than the simultaneous (not a significant difference in most cases), but the Bonferroni are quite a bit larger. Also as expected, the coverage probabilities for the pointwise bars are far too small.

In the example on demand theory treated in Figure 1, the functional form of this so-called Engel curve is of specific interest for theoretical economists. In particular the concavity of the curve at about two times the mean income ($x = 2.0$, as these data have been normalized by dividing by their mean) has important implications regarding the law of demand, see Hildenbrand and Hildenbrand (1986). The error bars for this potato/income example were constructed using the previous bootstrap method. Figure 1b indicates the nonmonotonicity of this Engel curve and supports other functional forms than those traditionally used, such as linear or working-type forms.

The previously described problems with bias are not a major problem in this example, because if the underadjustment of bias were improved, then our

conclusion of concavity near $x = 2.0$ is in fact strengthened. Also as the sample size in much larger now, it seems reasonable to hope that the asymptotic negligibility of the bias problem is closer to being realized.

4. Proofs.

PROOF OF THEOREM 1. For notational simplicity, the proof is given explicitly only for the case $d = 1$. The theorem is an immediate consequence of the following lemmas.

LEMMA 1. *Along almost all sample sequences,*

$$\sqrt{nh}\,[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})] \to N(B,V),$$

uniformly in h and g, in the sense that for all $\underset{\sim}{z} \in \mathbb{R}^N$,

$$\sup_{h\in H_n}\sup_{g\in G_n}\left|P^{Y|X}\left\{\sqrt{nh}\,[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})] < \underset{\sim}{z}\right\} - \Phi_{B,V}(\underset{\sim}{z})\right| \to 0,$$

where $\Phi_{B,V}$ denotes the normal cumulative distribution with mean B and covariance V and where

$$B = d_K\left\{m''(\underset{\sim}{x}) + 2m'(\underset{\sim}{x})\frac{f'(\underset{\sim}{x})}{f(\underset{\sim}{x})}\right\},$$

$$V = (v_{kl}), \qquad v_{kl} = \frac{K^{(2)}(c_k - c_l)\sigma^2(x_0)}{f(x_0)}$$

for $K^{(2)}$ the convolution of K with itself.

LEMMA 2. *Along almost all sample sequences,*

$$\sqrt{nh}\,[\hat{m}_h^*(\underset{\sim}{x}) - \hat{m}_g(\underset{\sim}{x})] \to N(B,V),$$

*uniformly in h and g, in the same sense as in Lemma 1 (except that the $Y|X$ distribution is replaced by the * distribution).*

PROOF OF LEMMA 1. The Cramér–Wold device is used in this proof. We will show that for all $\underset{\sim}{t} \in \mathbb{R}^N$ and all $z \in \mathbb{R}$,

$$\left|P^{Y|X}\left\{\underset{\sim}{t}^T\left(\sqrt{nh}\,[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})]\right) < z\right\} - \Phi\left(\left(z - \underset{\sim}{t}^T B\right)\Big/\sqrt{\underset{\sim}{t}^T V \underset{\sim}{t}}\right)\right| \to 0, \tag{4.1}$$

uniformly over $h \in H_n$, where Φ denotes the univariate standard normal c.d.f. To obtain uniformity over h requires some modification of the Cramér–Wold

device. In particular, Theorems 7.6 and 7.7 of Billingsley (1968) need to be extended in a straightforward fashion. To establish this, following Härdle and Marron (1985), we first make the linear approximation

$$\sqrt{nh}\,[\hat{m}_h(\underset{\sim}{x}) - m(\underset{\sim}{x})] = L_n + o_p(L_n), \tag{4.2}$$

where

$$L_n = \sqrt{nh}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(\underset{\sim}{x} - X_i)[Y_i - m(\underset{\sim}{x})]}{f(\underset{\sim}{x})}\right\}.$$

The term $o_p(L_n)$ is of lower order uniformly over H_n by (5.1) of Härdle and Marron (1985) and by Lemma 1 of that paper. Now write

$$L_n = V_n + B_n,$$

where

$$V_n = \sqrt{nh}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(\underset{\sim}{x} - X_i)\varepsilon_i}{f(\underset{\sim}{x})}\right\}$$

and $\varepsilon_i = Y_i - m(X_i)$,

$$B_n = \sqrt{nh}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{K_h(\underset{\sim}{x} - X_i)[m(X_i) - m(\underset{\sim}{x})]}{f(\underset{\sim}{x})}\right\}.$$

The proof of Lemma 1 follows from

$$\underset{\sim}{t}^T V_n \to N(0, \underset{\sim}{t}^T V \underset{\sim}{t}), \tag{4.3}$$

$$\underset{\sim}{t}^T B_n \to \underset{\sim}{t}^T B, \tag{4.4}$$

uniformly over $h \in H_n$.

To prove (4.1), we use Esseen's inequality for arbitrary independent random variables given, for example, on page 111 of Petrov (1975). For this purpose define $W_{hi}(\underset{\sim}{x}) = n^{-1/2}h^{1/2}K_h(\underset{\sim}{x} - X_i)/f(\underset{\sim}{x})$,

$$S_{2n} = \sum_{i=1}^{n} \operatorname{Var}\left(\underset{\sim}{t}TW_{hi}(\underset{\sim}{x})\varepsilon_i \middle| X_1, \ldots, X_n\right)$$

and

$$S_{3n} = \sum_{i=1}^{n} E\left(\left|\underset{\sim}{t}^T W_{hi}(\underset{\sim}{x})\varepsilon_i\right|^3 \middle| X_1, \ldots, X_n\right).$$

The Esseen inequality completes the verification of (4.3), when we show that

$$\sup_h S_{3n}/S_{2n}^{3/2} = o(1) \quad \text{a.s.}$$

To evaluate S_{2n}, note that $E^X S_{2n} = \underset{\sim}{t}^T V_{1n} \underset{\sim}{t}$, where the (k, l) element of V_{1n} is

by the assumption on x_k

$$\int K_h(x_k - u) K_h(x_l - u) f(u) \sigma^2(u)\, du / (f(x_k) f(x_l))$$

$$= h^{-1} K^{(2)}(c_l - c_k) \sigma^2(x_0) / f(x_0) + O(h^2).$$

Since $S_{2n} \to ES_{2n}$, a.s. by Theorem 1 of Feller [(1970), page 238] we have that

$$S_{2n} = h^{-1} K^{(2)}(c_l - c_k) \sigma^2(x_0) / f(x_0) + o(1) \quad \text{a.s.}$$

Uniformity over h is obtained by a suitable strengthening of the previous theorem. In the same manner the term S_{3n} can be evaluated to see that

$$\sup_h n^{1/2} h^{1/2} S_{3n} = O(1) \quad \text{a.s.}$$

Thus the statement (4.3) follows.

For the proof of (4.4), see the bias evaluation in Collomb (1981) or Härdle (1989). □

The proof of Proof of Lemma 2 is similar in spirit to that of Lemma 1, but is slightly more complicated because more terms arise.

PROOF OF THEOREM 3. The proof of (2.7) uses methods related to those in the proof of Theorem 1, so only the main steps are explicitly given. The first step is to decompose into variance and squared bias components,

$$E\left[\left(\hat{b}_{h,g}(x) - b_h(x)\right)^2 \middle| X_1, \ldots, X_n\right] = \mathscr{V}_n + \mathscr{B}_n^2, \tag{4.5}$$

where

$$\mathscr{V}_n = \operatorname{Var}\left(\hat{b}_{h,g}(x) \middle| X_1, \ldots, X_n\right),$$

$$\mathscr{B}_n = E\left(\hat{b}_{h,g}(x) - b_h(x) \middle| X_1, \ldots, X_n\right).$$

Using the same linearization technique as at (4.2) together with

$$\mathscr{B}_n = \mathscr{B}_{n1} + o(\mathscr{B}_{n1}),$$

where

$$\mathscr{B}_{n1} = \left[\int K_g(x - t) \mathscr{U}_h(t)\, dt - \mathscr{U}_h(x)\right] \Big/ f(x)$$

for

$$\mathscr{U}_h(x) = \int K_h(x - s)[m(s) - m(x)] f(s)\, ds.$$

Now by first integrating by substitution, then differentiating and finally Taylor expanding and collecting terms,

$$\mathscr{U}_h'(x) = h^2\left(\tfrac{1}{2} d_K\right)\left[(mf)^{(4)} - (mf'')''\right](x) + o(h^2).$$

Hence, by another substitution and Taylor expansion,

$$\mathscr{B}_{n2} = g^2h^2\left(\tfrac{1}{2}d_K\right)^2\left[(mf)^{(4)} - (mf'')''\right](x) + o(g^2h^2).$$

Thus, along almost all sample sequences,

$$\mathscr{B}_n^2 = C_2g^4h^4 + o(g^4h^4) \tag{4.6}$$

for C_2 as defined in the statement of Theorem 3.

Calculations in a similar spirit show that

$$\mathscr{V}_n = n^{-1}h^4g^{-5}C_1 + o(n^{-1}h^4g^{-5}),$$

where C_1 is defined in the statement of Theorem 3. This, together with (4.5) and (4.6) completes the proof of Theorem 3. □

REFERENCES

BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

BIERENS, H. J. and POTT–BUTER, H. A. (1987). Specification of household expenditure functions and equivalence scales by nonparametric regression. Technical Report 1987–44, Free Univ., Amsterdam.

BILLINGSLEY, P. (1986). *Convergence of Probability Measures*. Wiley, New York.

COLLOMB, G. (1981). Estimation nonparamétrique de la régression: Revue bibliographique. *Internat. Statist. Rev.* **49** 75–93.

COX, D. R. (1970). *Analysis of Binary Data*. Chapman and Hall, New York.

DIKTA, G. (1988). Approximation of nearest neighbor regression function estimators. Technical Report, Univ. Giessen.

EUBANK, R. (1988). *Spline Smoothing and Nonparametric Regression*. Dekker, New York.

Family Expenditure Survey, Annual Base Tapes (1968–1983) Department of Employment, Statistics Division, Her Majesty's Stationary Office, London 1968–1983. The data utilized in this book were made available by the ESRC Data Archive at the University of Essex.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

GASSER, T. and MÜLLER, H. G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

HALL, P. and TITTERINGTON, M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.* **27** 228–254.

HÄRDLE, W. (1989). *Applied Nonparametric Regression*. Econometric Society Monograph Series. Cambridge Univ. Press.

HÄRDLE, W. and BOWMAN, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102–110.

HÄRDLE, W. and MAMMEN, E. (1989). Comparing nonparametric versus parametric regression fits. Discussion paper A–177, Univ. Bonn.

HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? (with discussion). *J. Amer. Statist. Assoc.* **83** 86–101.

HILDENBRAND, K. and HILDENBRAND, W. (1986). On the mean income effect: A data analysis of the U.K. family expenditure survey. In *Contributions to Mathematical Economics* (W. Hildenbrand and A. Mas-Colell, eds.) 247–268. North-Holland, Amsterdam.

KNAFL, G., SACKS, J. and YLVISAKER, D. (1985). Confidence bands for regression functions. *J. Amer. Statist. Assoc.* **80** 683–691.

McDONALD, J. A. (1982). *Projection Pursuit Regression with the Orion I Workstation*, a 20 min 16 mm color sound film, available for loan from Jerome H. Friedman, Computation Research Group, Bin 88, SLAC, P.O. Box 4349, Stanford, Calif. 94305.

NADARAYA, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **10** 186–190.

PETROV, V. (1975). *Sums of Independent Random Variables*. Springer, New York.

ROSENBLATT, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis II* 25–31. Academic, New York.

ROSENBLUETH, E. (1975). Point estimates for probability moments. *Proc. Nat. Acad. Sci. U.S.A.* **72** 3812–3814.

WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.

WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1343.

CORE
34 VOIE DU ROMAN PAYS
1348 LOUVAIN-LA-NEUVE
BELGIUM

RECHTS- UND
STAATSWISSENSCHAFTLICHE FAKULTÄT
UNIVERSITÄT BONN
ADENAUERALLEE 24-26
5300 BONN 1
GERMANY