Statistica Sinica **11**(2001), 121-134

BOOTSTRAP CONFIDENCE INTERVALS FOR LOCAL LIKELIHOOD, LOCAL ESTIMATING EQUATIONS AND VARYING COEFFICIENT MODELS

Christian D. Galindo, Hua Liang, Goran Kauermann and Raymond J. Carrol

Texas A&M University and University of Glasgow

Abstract: Four powerful generalizations of the usual local polynomial nonparametric regression methodology are (a) local polynomial methods in generalized linear models; (b) varying coefficient generalized linear models, where the possibly multivariate coefficients in a generalized linear model are estimated nonparametrically; (c) local likelihood methods; and (d) local estimating equations, which generalize nonparametric regression to the estimating equation context. We construct bootstrap confidence intervals for the nonparametrically estimated functions in all four contexts.

Key words and phrases: Bootstrap, estimating equations, local estimating equations, local likelihood, local polynomial regression, nonparametric regression, robust covariance matrix, sandwich estimator.

1. Introduction

Ordinary local polynomial nonparametric regression function estimation (Fan and Gijbels (1996) is a convenient reference) has undergone active development recently. Confidence sets for the nonparametrically estimated function have been developed: Härdle and Marron (1991) use a bootstrap approach while Eubank and Speckman (1993) use asymptotics. Regression function *estimation* has been generalized considerably in the past few years, to generalized linear models (Fan, Heckman and Wand (1995)), to varying coefficient models and local likelihood (Tibshirani and Hastie (1987), Kauermann and Tutz (1999)), and to estimating equations (Carroll, Ruppert and Welsh (1998)). However, to the best of our knowledge, bootstrap–type confidence sets have not been developed for these useful generalizations. The purpose of our article is to provide one such approach, which applies in all the contexts described above.

Our method combines the ideas from the wild–bootstrap of Härdle and Marron (1991) (see Wu (1986) for the parametric version) with the estimating function bootstrap, suggested in a technical report by Kauermann and Tutz, to form a new and asymptotically justified bootstrap procedure. Our bootstrap is different from either: the former is restricted to ordinary kernel regression, while

the latter is restricted to ordinary kernel regression in varying coefficient generalized linear models. Our method improves upon that of the latter by having asymptotically correct coverage.

Local estimating functions are the most general version of these approaches, and we briefly review the ideas and develop some necessary notation. If the data (which might involve responses and covariates) are independent observations $\left(\widetilde{\boldsymbol{\mathcal{Y}}}_1, \widetilde{\boldsymbol{\mathcal{Y}}}_2, \ldots, \widetilde{\boldsymbol{\mathcal{Y}}}_n\right)$, with the $\widetilde{\boldsymbol{\mathcal{Y}}}$'s possibly vector-valued, then a parameter $\boldsymbol{\Theta}$ is estimated by solving the estimating equation

$$0 = \sum_{i=1}^{n} \boldsymbol{\psi}(\widetilde{\boldsymbol{\mathcal{Y}}}_i, \widehat{\boldsymbol{\Theta}}). \tag{1}$$

The function $\boldsymbol{\psi}(\cdot, \cdot)$ is called an estimating function. We allow $\boldsymbol{\Theta}$ to be vector-valued and $\boldsymbol{\psi}$ must have the same dimension as $\boldsymbol{\Theta}$.

In what follows, if $\mathbf{A}$ is $\ell \times q$ and $\mathbf{B}$ is $r \times s$, then $\mathbf{A} \otimes \mathbf{B}$ is the Kronecker product defined as the $\ell r \times qs$ matrix which is formed by multiplying individual elements of $\mathbf{A}$ by $\mathbf{B}$, e.g., if $\mathbf{A}$ is a 2×2 matrix,

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{bmatrix}.$$

As is standard, in nonparametric regression we wish to allow a parameter to depend on a predictor. Thus, we have the parametric problem of estimating a mean, while the nonparametric function analogue allows the mean to depend upon a predictor. In making this transition, we will refer to the function so obtained as $\Theta(x)$.

The local estimating function method for estimating $\boldsymbol{\Theta}(\cdot)$ can involve local polynomials. With superscript (j) denoting a jth derivative with respect to x and with $\mathbf{b}_j = \boldsymbol{\Theta}^{(j)}(x_0)/j!$, the local polynomial of order $p \geq 0$ in a neighborhood of x_0 is $\boldsymbol{\Theta}(x) \approx \sum_{j=0}^{p} \mathbf{b}_j (x - x_0)^j$. The local weight for a value of x near x_0 is denoted by $w_h(x, x_0)$, where h is a tuning constant. For example, the weights $w_h(X, x)$ can be the nearest neighbor weights with span h of loess (Chambers and Hastie (1992)) or kernel weights with kernel density function $K(\cdot)$, bandwidth h and weights $w_h(X, x) = h^{-1}K\{(X - x)/h\} = K_h(X - x)$. Carroll, *et al.* (1998) propose to solve in $(\mathbf{b}_0, \ldots, \mathbf{b}_p)$ the $q \times (p + 1)$ equations

$$0 = \sum_{i=1}^{n} w_h(X_i, x_0)[\mathbf{G}_p (X_i - x_0) \otimes \boldsymbol{\psi}\{\widetilde{\boldsymbol{\mathcal{Y}}}_i, \sum_{j=0}^{p} \mathbf{b}_j (X_i - x_0)^j\}], \tag{2}$$

where $\mathbf{G}_p^t(v) = (1, v, v^2, \ldots, v^p)$. The final estimates are $\widehat{\boldsymbol{\Theta}}(x_0) = \widehat{\mathbf{b}}_0$. Usually, only a scalar function of $\boldsymbol{\Theta}(\cdot)$ is of interest.

A few examples illustrate the generality of the approach.

1. In ordinary nonparametric regression, the response is $\widetilde{\boldsymbol{\mathcal{Y}}} = Y$ and $\boldsymbol{\psi}(\widetilde{\boldsymbol{\mathcal{Y}}}, \mathbf{v}) = Y - v$.
2. In generalized linear models, the mean function is $\mu(x)$, the variance is proportional to $V(x)$, the response is $\widetilde{\boldsymbol{\mathcal{Y}}} = Y$, and $\boldsymbol{\psi}(\widetilde{\boldsymbol{\mathcal{Y}}}, \boldsymbol{\Theta}) = \{Y - \mu(\Theta)\}\mu^{(1)}(\Theta)/V(\Theta)$, where $\mu^{(1)}(\Theta) = (\partial/\partial\Theta)\mu(\Theta)$.
3. In varying coefficient generalized linear models, $\widetilde{\boldsymbol{\mathcal{Y}}} = (Y, Z)$, $\boldsymbol{\Theta} = (\theta_0, \boldsymbol{\Theta}_1^t)^t$ and the mean function is $\mu(\theta_0 + \boldsymbol{\Theta}_1^t\mathbf{Z})$, where $\mathbf{Z}$ is some design matrix. The estimating function is

$$\boldsymbol{\psi}(\widetilde{\boldsymbol{\mathcal{Y}}}, \boldsymbol{\Theta}) = \Big[\{Y - \mu(\theta_0 + \boldsymbol{\Theta}_1^t\mathbf{Z})\}\mu^{(1)}(\theta_0 + \boldsymbol{\Theta}_1^t\mathbf{Z})/V(\theta_0 + \boldsymbol{\Theta}_1^t\mathbf{Z})\Big]\,(1, \mathbf{Z})^t.$$

Let $\boldsymbol{\chi}(\widetilde{\boldsymbol{y}}, \mathbf{v}) = (\partial/\partial\mathbf{v}^t)\boldsymbol{\psi}(\widetilde{\boldsymbol{y}}, \mathbf{v})$. Then, for constants $c(n, p)$ defined below, the asymptotic covariance matrix of $(\widehat{\mathbf{b}}_0, \ldots, \widehat{\mathbf{b}}_p)$ is estimated by $\{\mathbf{B}_h(x_0)\}^{-1}\mathbf{C}_h(x_0)\{\mathbf{B}_h^t(x_0)\}^{-1}$, where

$$\mathbf{C}_h(x_0) = c(n, p)\sum_{i=1}^{n} w_h^2(X_i, x_0)\Big[\Big\{\mathbf{G}_p(X_i - x_0)\mathbf{G}_p^t(X_i - x_0)\Big\} \otimes (\widehat{\boldsymbol{\psi}}_i\widehat{\boldsymbol{\psi}}_i^t)\Big]; \quad (3)$$

$$\mathbf{B}_h(x_0) = \sum_{i=1}^{n} w_h(X_i, x_0)\Big[\Big\{\mathbf{G}_p(X_i - x_0)\mathbf{G}_p^t(X_i - x_0)\Big\} \otimes \widehat{\boldsymbol{\chi}}_i\Big], \quad (4)$$

where $\widehat{\boldsymbol{\psi}}_i = \boldsymbol{\psi}\{\widetilde{\boldsymbol{\mathcal{Y}}}_i, \widehat{\boldsymbol{\Theta}}(X_i)\}$ and analogously for $\widehat{\boldsymbol{\chi}}_i$. The covariance estimate is a nonparametric version of the sandwich estimate (often known as the robust covariance matrix estimator). An asymptotic argument justifying these formula is given by Carroll, *et al.* (1998). In some cases, e.g., local likelihood, it is known that $E(\boldsymbol{\psi}_i\boldsymbol{\psi}_i^t|X = x) = E(\boldsymbol{\chi}_i|X = x)$, in which case one would replace $\widehat{\boldsymbol{\psi}}_i\widehat{\boldsymbol{\psi}}_i^t$ by $\widehat{\boldsymbol{\chi}}_i$ in (3).

The usual sandwich formula (Huber (1967), White (1982), Diggle, Liang and Zeger (1994)) sets $c(n, p) = 1$, and the confidence intervals are formed in the usual way by comparing to a standard normal distribution. In ordinary parametric linear regression, we have found that such a procedure can have poor coverage probabilities, even for samples sizes up to 30. For example, with a sample size of 20 and normally distributed predictors, the coverage is only 92%. When the predictors have a Laplace distribution the coverage falls to 90%. Ordinary bootstrap–t confidence intervals based on this sandwich estimator have coverage probabilities of only 94% and 92% respectively.

In unpublished work, we have found that in simple linear regression, the combination of setting $c(n, p) = n/(n-p-3)$, adjusting the terms $\widehat{\boldsymbol{\psi}}_i$ for leverage, and comparison with the t_{n-p-3} distribution instead of the standard normal distribution, markedly improves coverage — to 95% for normal predictors and 94% for Laplace predictors in the case of a sample size of 20. The net effect is

to replace $\widehat{\boldsymbol{\psi}}_i$ by $c^{1/2}(n,p)\widehat{\boldsymbol{\psi}}_i/\widehat{l}_i^{1/2}$, where $\widehat{l}_i$ is the leverage adjustment. In this paper, we suggest versions of $c(n,p)$ and leverage adjustments, while letting the wild–bootstrap take the place of the t_{n-p-3} distribution.

The outline of this paper is as follows. In Section 2, we describe the algorithm. A small simulation is given in Section 3. Section 4 contains an example from nutritional epidemiology. Concluding remarks are given in Section 5.

2. Algorithm

To motivate the algorithm, we briefly describe the wild–bootstrap of Härdle and Marron (1991) and the estimating function bootstrap suggested in a technical report by Kauermann and Tutz. The combination of the two methods is new.

In ordinary kernel regression of scalar random variables Y on X, the wild–bootstrap works as follows. Let $\widehat{\boldsymbol{\Theta}}_h(x)$ be the kernel regression estimator based on a bandwidth h, and let $\widehat{\boldsymbol{\Theta}}_g(x)$ be based on a larger bandwidth g (see below). The residuals are $\widehat{\epsilon}_i = Y_i - \widehat{\boldsymbol{\Theta}}_h(X_i)$. Härdle and Marron generate a wild–bootstrap distribution of ϵ_i^*, which is defined precisely below but is in general a 2–point distribution with mean zero, variance $\widehat{\epsilon}_i^2$, and third moment $\widehat{\epsilon}_i^3$. Then the bootstrap observations are $Y_i^* = \widehat{\boldsymbol{\Theta}}_g(X_i) + \epsilon_i^*$. The bootstrapped function is $\widehat{\boldsymbol{\Theta}}_h^*(x)$, i.e., ordinary kernel regression with bandwidth h applied to the bootstrap data. Although we can use a distribution with more than two points from which to conduct the bootstrap, simulation results not reported here indicate no appreciable benefit in doing so.

In local estimating equations, the major modification of this algorithm that is necessary is to define an appropriate "residual" to which the wild–bootstrap can be applied. Let $\mathbf{e}_{i,j}$ be a $1 \times j$ vector of zeros with a one in the i^{th} position, and $\mathbf{E}_{i,j,q} = \mathbf{e}_{i,j} \otimes \mathbf{I}_q$. Then we propose that the "residual" in estimating functions is $\{\mathbf{E}_{1,p+1,q}\mathbf{B}_h(X_i)\mathbf{E}_{1,p+1,q}^t\}^{-1}\boldsymbol{\psi}\{\widetilde{\boldsymbol{\mathcal{Y}}}_i, \widehat{\boldsymbol{\Theta}}_h(X_i)\}$.

What Kauermann and Tutz suggest is to fix the first term and bootstrap only the second, in our case by the wild–bootstrap. The algorithm is thus defined as follows.

(a) Let $\widehat{\boldsymbol{\Theta}}_h(\cdot)$ be the local estimating equation estimator based on local polynomials of order p with tuning constant h. Let $\widehat{\boldsymbol{\Theta}}_g(\cdot)$ be the same estimator but with a larger tuning constant g. In kernel regression, one asymptotically justifiable choice for g is $g = h^{(2p+3)/(2p+5)}$ when p is odd and $g = h^{(2p+5)/(2p+7)}$ when p is even. Of course, it is asymptotically justifiable to add constant multipliers in front of these default choices.

(b) Let $\mathbf{B}_h(\cdot)$ be defined by (4).

(c) Let $\boldsymbol{\epsilon}_i^*$ be the wild–bootstrap version of $\widehat{\boldsymbol{\epsilon}}_i = c(n,p,h,X_i)\boldsymbol{\psi}\{\widetilde{\boldsymbol{\mathcal{Y}}}_i, \widehat{\boldsymbol{\Theta}}_h(\mathbf{X}_i)\}$, where $c(n,p,h,X_i)$ is analogous to $c(n,p)$ in (3) and is meant to be an empirical adjustment to improve coverage probability. In our simulations,

we used $c(n,p,h,X_i) = [n_h(X_i)/\{n_h(X_i) - (p+3)\}]^{1/2}$, where $n_h(X_i) = K_h^{-1}(0)\sum_{j=1}^n K_h(X_j - X_i)$, the sum of standardized weights. In our simulations, we had no cases that $n_h(X_i) \le p+3$, but if this were to happen one could either bound $c(n,p,h,X_i)$ by an arbitrary number or follow the suggestion of a referee: $n_h(X_i) = \{\sum_j K_h(X_j - X_i)\}^2/\sum_j K^2(X_j - X_i)$. Let $a = \widehat{\boldsymbol{\epsilon}}_i(1-5^{1/2})/2$, $b = \widehat{\boldsymbol{\epsilon}}_i(1+5^{1/2})/2$, $c = (5+5^{1/2})/10$. Then $\boldsymbol{\epsilon}_i^*$ equals a with probability c and equals b with probability $(1-c)$. This distribution has mean zero and covariance matrix $\widehat{\boldsymbol{\epsilon}}_i\widehat{\boldsymbol{\epsilon}}_i^t$.

(d) Define $\widetilde{\boldsymbol{\mathcal{Y}}}_i^* = \widehat{\boldsymbol{\Theta}}_g(\mathbf{X}_i) + \{\mathbf{E}_{1,p+1,q}\mathbf{B}_h(X_i)\mathbf{E}_{1,p+1,q}^t\}^{-1}\boldsymbol{\epsilon}_i^*$.

(e) The bootstrap function estimate is $\widehat{\boldsymbol{\Theta}}_h^*(\cdot)$, the standard local polynomial regression estimator of order p with tuning constant h when the response is $\widetilde{\boldsymbol{\mathcal{Y}}}_i^*$. *It is important to emphasize that the regression being done is simply a standard local polynomial regression function estimate.* Thus, it is a solution to (2), but with $\psi(y,v) = y - v$.

(f) Finally, apply a confidence interval construction algorithm such as that of Härdle and Marron (1991, p.785), although note that they have a typographical error: in their equation (2.6), the sets $I_{j,k}(\beta_j)$ should be -$I_{j,k}(\beta_j)$, while the term $\widehat{\boldsymbol{\Theta}}_g$ should not be subtracted. Thus, if the upper and lower bounds of a $1-\alpha$ interval for the bootstrap distribution of $\widehat{\boldsymbol{\Theta}}_h^*(\cdot) - \widehat{\boldsymbol{\Theta}}_g$ are a and b, respectively, a pointwise confidence interval is $\widehat{\boldsymbol{\Theta}}_g - b$ to $\widehat{\boldsymbol{\Theta}}_g - a$.

In the appendix, we give a proof for kernel weights that the bootstrap confidence intervals achieve asymptotically their nominal level.

To handle leverage, the estimator $\widehat{\boldsymbol{\psi}}$ can be constructed to have a variance approximating that of $\boldsymbol{\psi}$. For example, when $\boldsymbol{\psi}$ is scalar, (2) can be written as

$$0 = \sum_{i=1}^n w_h(X_i,x_0)\boldsymbol{\psi}\Big\{\widetilde{\boldsymbol{\mathcal{Y}}}_i, \sum_{j=0}^p \mathbf{b}_j\,(X_i - x_0)^j\Big\}\mathbf{G}_p\,(X_i - x_0)\,.$$

Using a Taylor's series expansion, $\widehat{\boldsymbol{\Theta}}_h(x_0) \approx \boldsymbol{\Theta}(x_0) - \mathbf{e}_{1,p+1}\mathbf{B}_h^{-1}(x_0)\mathbf{A}_{h,p}(x_0)\underline{\boldsymbol{\psi}}$, where $\underline{\boldsymbol{\psi}} = (\boldsymbol{\psi}_1,\ldots,\boldsymbol{\psi}_n)^t$, and $\mathbf{A}_{h,p}(x)$ is the $(p+1)\times n$ matrix defined by

$$\mathbf{A}_{h,p}(x) = \begin{bmatrix} w_h(X_1,x) & \cdots & w_h(X_n,x) \\ \vdots & \cdots & \vdots \\ w_h(X_1,x)(X_1-x)^p & \cdots & w_h(X_n,x)(X_n-x)^p \end{bmatrix}.$$

Thus, by defining $\mathbf{S}_h(X_i) = \boldsymbol{\chi}_i\mathbf{e}_{1,p+1}\mathbf{B}_h^{-1}(X_i)\mathbf{A}_{h,p}(X_i)$, we see $\widehat{\boldsymbol{\psi}}_i \approx \{\mathbf{e}_{i,n} - \mathbf{S}_h(X_i)\}\underline{\boldsymbol{\psi}}$. In order to equate the variance of the residual with its estimate, we suggest replacing $\widehat{\boldsymbol{\psi}}_i$ with $c_h(X_i)^{-1/2}\widehat{\boldsymbol{\psi}}_i$, where $c_h(X_i) = \{\mathbf{e}_{i,n} - \mathbf{S}_h(X_i)\}\{\mathbf{e}_{i,n} - \mathbf{S}_h(X_i)\}^t$. Change part (c) of the algorithm to let $\boldsymbol{\epsilon}_i^*$ be the wild–bootstrap version of the modified residual estimate, $\widehat{\boldsymbol{\epsilon}}_{i,new} = [n_h(X_i)/\{n_h(X_i) - (p+3)\}]^{1/2}c_h(X_i)^{-1/2}\widehat{\boldsymbol{\psi}}_i$.

This residual modification can be extended analogously to multivariate $\boldsymbol{\psi}$. In this instance $\widehat{\boldsymbol{\psi}}_i \approx \{\mathbf{E}_{i,n,q} - \mathbf{S}_h(X_i)\}\underline{\boldsymbol{\psi}}$, with $\mathbf{S}_h(X_i) = \boldsymbol{\chi}_i \mathbf{E}_{1,p+1,q} \mathbf{B}_h^{-1}(X_i) \mathbf{A}^*_{h,p,q}(X_i)$, $\mathbf{A}^*_{h,p,q}(X_i) = \mathbf{A}_{h,p}(X_i) \otimes \mathbf{I}_q$, and $\underline{\boldsymbol{\psi}} = (\boldsymbol{\psi}_1^t, \cdots, \boldsymbol{\psi}_n^t)^t$. Now set $\mathbf{C}_h(X_i) = \{\mathbf{E}_{i,n,q} - \mathbf{S}_h(X_i)\}\{\mathbf{E}_{i,n,q} - \mathbf{S}_h(X_i)\}^t$, estimate $\boldsymbol{\psi}_i$ by $\mathbf{C}_h^{-1/2}(X_i)\widehat{\boldsymbol{\psi}}_i$, and once again use $\widehat{\boldsymbol{\epsilon}}_{i,new}$ in place of $\widehat{\boldsymbol{\epsilon}}_i$.

3. Simulation

To check the methods, we performed several small local linear regression simulations. Given X, Y is simulated to follow a normal distribution with unit variance and a mean of either $m_1(X) = 2X + \sin 8X$ or $m_2(X) = (2\pi)^{-1/2} 4(X - 0.5) + 4\exp[-2\{4(X - 0.5)\}^2]$. The latter function, $m_2(\cdot)$, is the one used in the simulations of Härdle and Marron but translated and rescaled to the $(0, 1)$ domain. The distribution of X is set to either Uniform$(0, 1)$ or Normal$(0.5, 0.25)$. We consider the construction of 95% pointwise confidence intervals using both unmodified and modified residuals. All simulations use a fixed bandwidth ($h = 0.175$ and 0.125 for $n = 100$ and 200 respectively), though similar results are achieved using Ruppert's (1997) empirical-bias bandwidth selection method (EBBS). Figure 1 graphs the data along with $m(X)$ and its estimates, based on the bandwidths h and g, for a typical simulation. Note how $\widehat{m}_g(X)$ is over-smoothed compared to $\widehat{m}_h(X)$.

$E(Y|X)$ with estimates based on bandwidths h and g

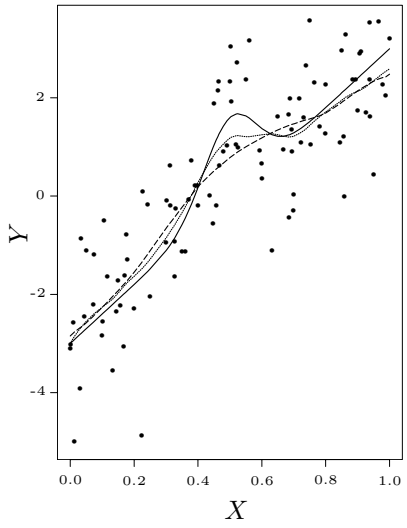

Figure 1. True value of $E(Y|X) = m_2(X) = (2\pi)^{-1/2} 4(X - 0.5) + 4\exp[-2\{4(X - 0.5)\}^2]$, denoted by the solid line, along with the estimates $\widehat{m}_{2h}(X)$ and $\widehat{m}_{2g}(X)$, in dotted and dashed lines respectively.

Denote the unmodified residuals by $\widehat{\boldsymbol{\epsilon}}_i = Y_i - \widehat{m}_h(X_i)$, and the modified residuals by $\widehat{\boldsymbol{\epsilon}}_{i,new} = [n_h(X_i)/\{n_h(X_i) - (p+3)\}]^{1/2} c_h(X_i)^{-1/2} \widehat{\boldsymbol{\epsilon}}_i$. Empirical coverage rates are calculated using 200 simulations, each with 500 bootstraps.

Table 1 summarizes the results of the simulation. We see that the computed intervals fall substantially below the targeted confidence levels when using $\widehat{\boldsymbol{\epsilon}}_i$ to form the bootstrap residuals, even with a sample size of 200 observations. The pointwise intervals at the boundaries cause concern, with low coverage compared to the center.

Table 1. Empirical coverage rate comparison of confidence intervals using unmodified residuals estimates, $\widehat{\boldsymbol{\epsilon}}_i = Y_i - \widehat{m}_h(X_i)$, versus modified residual estimates, $\widehat{\boldsymbol{\epsilon}}_{i,new} = [n_h(X_i)/\{n_h(X_i) - (p+3)\}]^{1/2} c_h(X_i)^{-1/2} \widehat{\boldsymbol{\epsilon}}_i$. Target confidence levels for pointwise intervals are 95%.

Comparison of confidence intervals using unmodified versus modified residual estimates

Sample Size	100		200	
	$\widehat{\boldsymbol{\epsilon}}_i$	$\widehat{\boldsymbol{\epsilon}}_{i,new}$	$\widehat{\boldsymbol{\epsilon}}_i$	$\widehat{\boldsymbol{\epsilon}}_{i,new}$
$X \sim \text{Uniform}(0,1)$				
$m_1(X)$				
$X = 0$	0.835	0.895	0.880	0.925
$X = 0.5$	0.930	0.960	0.960	0.965
$X = 1$	0.835	0.930	0.885	0.935
$m_2(X)$				
$X = 0$	0.815	0.875	0.855	0.920
$X = 0.5$	0.800	0.840	0.925	0.955
$X = 1$	0.840	0.920	0.900	0.950
$X \sim \text{Normal}(0.5, 0.25)$				
$m_1(X)$				
$X = 0$	0.620	0.880	0.730	0.885
$X = 0.5$	0.870	0.925	0.940	0.965
$X = 1$	0.610	0.875	0.725	0.915
$m_2(X)$				
$X = 0$	0.630	0.885	0.730	0.885
$X = 0.5$	0.765	0.905	0.920	0.935
$X = 1$	0.605	0.885	0.745	0.910

Now consider the use of the estimates $\widehat{\boldsymbol{\epsilon}}_{i,new}$, which are greater in magnitude than $\widehat{\boldsymbol{\epsilon}}_i$, allowing for wider intervals and increased coverage rates. The correction factors are largest in areas with the least information, and reduce the effects of imprecision by widening the confidence intervals. When X follows a uniform distribution the estimation of $m(X)$ uses approximately half the observations on

the boundary of the unit interval than at the center. Accordingly, the confidence intervals on the endpoints of X are wider than those at the center, as illustrated in Figure 2 with a set of data from a run different from that of Figure 2. This phenomenon is more extreme when X follows a normal distribution, with even fewer observations at the boundaries when compared to internal regions. Consequently, we would not expect the simulations to perform as well at the endpoints when X is normal than when X is uniform — precisely what occurs in the simulations.

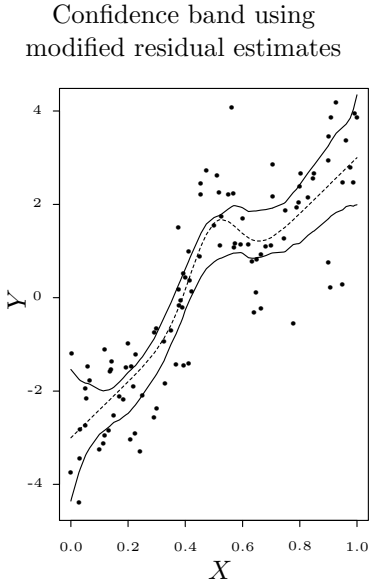

Figure 2. Confidence band constructed using wild–bootstrap method with modified residuals, $\widehat{\boldsymbol{\epsilon}}_{i,new}$. The dashed line represents the true function $m_2(X) = (2\pi)^{-1/2}4(X - 0.5) + 4\exp[-2\{4(X - 0.5)\}^2]$.

Unlike the outcome of bootstrapping off $\widehat{\boldsymbol{\epsilon}}_i$, the modified residual construct reduces the coverage disparity between pointwise confidence intervals at the boundary points and internal points. Also, with a sample size of 100 the confidence intervals are, in most instances, nearer nominal levels than the results for $\widehat{\boldsymbol{\epsilon}}_i$ with a sample size of 200, see Table 1.

We also computed nominal 80% global confidence intervals for the function $m_1(x)$ when X was uniformly distributed, using the Härdle–Marron construction. As in their paper, using the unmodified residuals results in considerable undercoverage, with the actual level being 60% with $n = 200$. Using modified residuals, the actual level with $n = 200$ is 76%, much closer to the nominal level.

4. Example

Here we reconsider an example from nutritional epidemiology discussed in detail by Carroll, *et al.* (1998). The instrument of choice in large nutritional

epidemiology studies is the Food Frequency Questionnaire (FFQ). For proper interpretation of epidemiologic studies that use FFQ's as the basic dietary instrument, one needs to know the relationship between reported intakes from the FFQ and true usual intake, defined operationally below. Such a relationship is ascertained through a substudy, commonly called a calibration study. Carroll, *et al.* focus on the estimation of the correlation between FFQ intake and usual intake, using the variable Percent of Calories from Fat. FFQ's are thought to often involve a systematic bias (i.e., under- or over-reporting at the level of the individual). The other two instruments that are commonly used are the 24-hour food recall and the multiple-day food record (FR). Each of these FR's is more work-intensive and costly, but is thought to involve considerably less bias than a FFQ. Carroll, *et al.* comment on this and other issues in nutrition data.

For the ith individual ($i = 1, \ldots, n$), let Q_i denote the intake of a nutrient reported on a FFQ. For the jth ($j = 1, \ldots, m$) replicate on the ith person, let F_{ij} denote the intake reported by a FR. Finally, long–term usual intake for the ith person is denoted by T_i. A simple model relating these three is a standard linear errors-in-variables model

$$\begin{aligned} Q_i &= \beta_0 + \beta_1 T_i + \epsilon_i; \\ F_{ij} &= T_i + U_{ij}; \quad j = 1, \ldots, m. \end{aligned} \tag{5}$$

The U_{ij}'s are the within–individual variation in FR's. All random errors, i.e., ϵ's and U's, are uncorrelated for purposes of this paper.

The Nurses' Health Study (Rosner, Willett and Spiegelman (1989)), hereafter denoted by NHS, has a calibration study of $n = 168$ women, all of whom completed a single FFQ and four multiple-day food diaries ($m = 4$ in our notation). Our interest here is in estimating the quantity ρ_{QT}, the correlation between the FFQ and usual intake. Carroll, *et al.* note that there appears to be a relationship between ρ_{QT} and $X =$ age, and it is this relationship which we seek to estimate.

Nutrition data with repeated measurements typically have the feature of time trends in total amounts and sometimes in percentages, so that, for example, one might expect reported caloric intake (energy) to decline over time. To take this into account, we ratio adjusted all measurements so that the mean of each FR equals the first.

As described previously, i denotes the individual, Q_i and T_i are the nutrient intakes as reported on the FFQ and usual intake, respectively, and F_{ij} is the jth replicated FR for the ith individual. The mean of the replicated FR's is $\overline{F}_i$. The unknown parameters in the problem are characterized as $\mathbf{\Theta} = (\theta_1, \ldots, \theta_6)$, where $\theta_1 = E(Q)$, $\theta_2 = E(F) = E(T)$, $\theta_3 = E(Q^2)$, $\theta_4 = E(QF) = E(QT)$, $\theta_5 = \text{var}(U)$ and $\theta_6 = E(T^2)$. Note that for any two replicates F_{ij} and F_{ik},

$j \neq k$, $\theta_6 = E(F_{ij}F_{ik})$. Letting $\widetilde{\boldsymbol{\mathcal{Y}}}_i = (Q_i, F_{i1}, \ldots, F_{im})$ be the observed data, the usual method of moments estimating function is

$$\boldsymbol{\psi}(\widetilde{\boldsymbol{\mathcal{Y}}}_i, \boldsymbol{\Theta}) = \begin{bmatrix} Q_i \\ \overline{F}_i \\ Q_i^2 \\ Q_i\overline{F}_i \\ (m-1)^{-1}\sum_{j=1}^m (F_{ij} - \overline{F}_i)^2 \\ \{m(m-1)\}^{-1}\sum_{j=1}^m \sum_{k\neq j}^m F_{ij}F_{ik} \end{bmatrix} - \boldsymbol{\Theta}. \tag{6}$$

Numerically, the solution to (2) is easily obtained by direct local least squares on each component of $\boldsymbol{\Theta}$, e.g., by regressing Q_i on X_i to obtain $\widehat{\theta}_1$. The main parameter of interest is the correlation between Q and T, $\rho_{QT}(x_0) = \{\theta_4(x_0) - \theta_1(x_0)\theta_2(x_0)\}[\{\theta_3(x_0) - \theta_1^2(x_0)\}\{\theta_6(x_0) - \theta_2^2(x_0)\}]^{-1/2}$.

For our illustration, we use kernel weighting with the Epanechnikov kernel $K(x) = (3/4)(1 - x^2)I(|x| \leq 1)$. The methods of Carroll, *et al.* suggest that the bandwidth is well–approximated by $h = 15$ for $x \leq 35$ and $x > 55$, $h = 22$ for $40 \leq x \leq 53$ and with smooth interpolation in between.

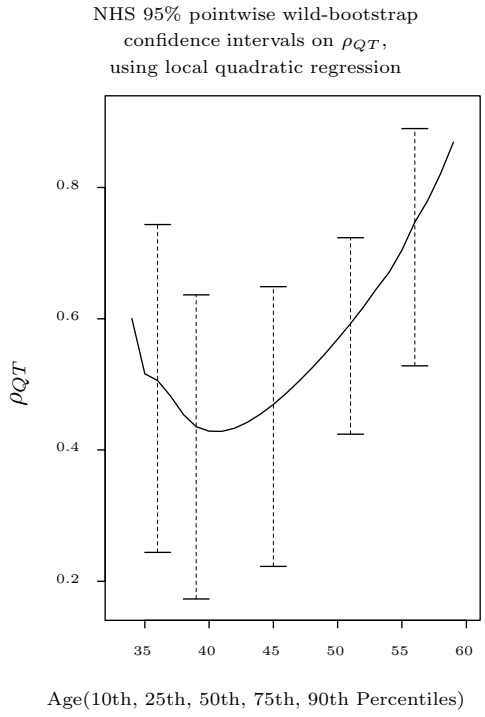

Figure 3. Nurses' Health Study percent of calories from fat, estimating ρ_{QT} using local quadratic regression, along with 95% wild–bootstrap pointwise confidence intervals on the 10th, 25th, 50th, 75th, and 90th percentiles of age.

Figure 3 displays the estimated curve $\widehat{\rho}_{QT}(age)$ for NHS Percent of Calories from Fat, using local quadratic regression, along with 95% wild–bootstrap pointwise confidence intervals at the 10th, 25th, 50th, 75th, and 90th sample percentiles of age. An important characteristic of the graph is that correlations are lower, in a practical sense, for those individuals under 50 years of age than for those older than 50. To assess statistical significance, the data are split into two populations on the basis of age (below 50 versus 50 or older) and $\widehat{\rho}_{QT}$ is computed for both groups. The correlations are found to be significantly different at level 0.02.

Table 2. Nurses' Health Study percent of calories from fat, confidence intervals and corresponding lengths on ρ_{QT}, using local mean, linear, and quadratic regression. The sandwich intervals are computed by $\widehat{\rho}_{QT,h}(x) \pm t_{\alpha/2,df}\widehat{\sigma}(x)$, with $df = n - (p+1)$ and $\widehat{\sigma}(x)$ determined from the sandwich estimate of variance.

Confidence intervals on ρ_{QT} in NHS
using wild–bootstrap and sandwich methods

Percentile		**10**	**25**	**50**	**75**	**90**
Mean	**Bootstrap**	(0.284,0.640)	(0.368,0.652)	(0.399,0.671)	(0.424,0.682)	(0.467,0.742)
		0.355	0.284	0.271	0.257	0.275
	Sandwich	(0.293,0.644)	(0.372,0.648)	(0.420,0.670)	(0.446,0.689)	(0.494,0.759)
		0.351	0.276	0.250	0.243	0.265
Linear	**Bootstrap**	(0.161,0.704)	(0.238,0.639)	(0.416,0.662)	(0.500,0.746)	(0.561,0.897)
		0.543	0.401	0.246	0.245	0.335
	Sandwich	(0.256,0.751)	(0.263,0.660)	(0.410,0.668)	(0.499,0.753)	(0.582,0.908)
		0.494	0.397	0.257	0.254	0.326
Quadratic	**Bootstrap**	(0.244,0.743)	(0.173,0.636)	(0.222,0.648)	(0.424,0.723)	(0.528,0.889)
		0.499	0.463	0.426	0.299	0.361
	Sandwich	(0.265,0.745)	(0.203,0.667)	(0.265,0.673)	(0.443,0.742)	(0.585,0.908)
		0.479	0.464	0.408	0.299	0.323

Table 2 presents the endpoints and lengths of confidence intervals constructed using the wild-bootstrap method, as well as with the sandwich formula estimate of standard error presented in Section 1. Though the two methods do not consistently share common endpoints (bootstrap intervals are not constructed to be symmetric about the estimate), lengths are roughly the same. This (approximate) equivalence should not be of much surprise as both the bootstrap method and sandwich formula asymptotically estimate the same measure of spread.

5. Discussion

We have provided confidence sets and pointwise confidence intervals for nonparametric function estimates based on local estimating equations. The methods are a generalization of the wild–bootstrap of Härdle and Marron (1991), and are based on local polynomial methods. The coverage probabilities of the wild–bootstrap can be far from nominal in moderately sized problems, and to correct for this we suggest two finite–sample improvements. The first essentially attempts to adjust the local estimating equation-based "residual" to have the correct variance: this is just studentization in ordinary nonparametric regression. The second correction is to adjust the coverage probabilities formed by sandwich-type methods, such as the wild–bootstrap, to be more nearly nominal. In the linear regression case, the corrections had a substantial impact on coverage probabilities.

Acknowledgement

Carroll's research was supported by a grant from the National Cancer Institute (CA–57030), and through the Texas A&M Center for Environmental and Rural Health by a grant from the National Institute of Environmental Health Sciences (P30-ES09106).

Appendix

In this section, we provide the asymptotic justification of our method for kernel weights. In effect, the result follows by noting a known asymptotic expansion equivalent to (4.2) of the proof of Härdle and Marron, From this, the details follow essentially line–by–line.

Recall that the kernel function is $K(\cdot)$, $K_h(v) = h^{-1}K(v/h)$, and $\mathbf{G}_p(v) = (1, v, \ldots, v^p)^t$. Let $f_X(\cdot)$ be the density function of X.

The basic idea for construction of a wild–bootstrap confidence interval is fairly simple and follows the outline in Härdle and Marron (1991, p.785). Suppose that $\widehat{\boldsymbol{\eta}}_1$ is an estimate of a parameter $\boldsymbol{\eta}$, and that for a sequence of constants $c(n) \to \infty$, $c(n)(\widehat{\boldsymbol{\eta}}_1 - \boldsymbol{\eta}) \to \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Suppose further that there is a possibly different estimator $\widehat{\boldsymbol{\eta}}_2$ with the property that under bootstrap sampling, $c(n)(\widehat{\boldsymbol{\eta}}^* - \widehat{\boldsymbol{\eta}}_2) \to \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, see Härdle and Marron for more technical details. Let $\widehat{\boldsymbol{\eta}}_2 + \boldsymbol{\mathcal{C}}_n$ be the resulting $(1-\alpha)100\%$ bootstrap confidence set. Then $\widehat{\boldsymbol{\eta}}_1 - \boldsymbol{\mathcal{C}}_n$ is an asymptotic $(1-\alpha)100\%$ confidence set for $\boldsymbol{\eta}$. To see this, note that asymptotically $\boldsymbol{\mathcal{C}}_n \approx (\boldsymbol{\mu} + \mathbf{V})/c(n)$ where $\mathbf{V}$ is a set with probability $1-\alpha$ under the normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Hence the coverage probability of the set $\widehat{\boldsymbol{\eta}}_1 - \boldsymbol{\mathcal{C}}_n$ is

$$\text{pr}\{\boldsymbol{\eta} \in \widehat{\boldsymbol{\eta}}_1 - (\boldsymbol{\mu} + \mathbf{V})/c(n)\} = \text{pr}[c(n)\{\widehat{\boldsymbol{\eta}}_1 - \boldsymbol{\eta} - \boldsymbol{\mu}/c(n)\} \in \mathbf{V}] \to 1 - \alpha,$$

as claimed.

This argument shows that what is required is that the limit distribution of $c(n)\{\widehat{\boldsymbol{\Theta}}_h(x) - \boldsymbol{\Theta}(x)\}$ in the sample and the bootstrap limit distribution of $c(n)\{\widehat{\boldsymbol{\Theta}}^*_h(x) - \widehat{\boldsymbol{\Theta}}_g(x)\}$ be the same, where $c(n)$ is proportional to $n^{(p+1)/(2p+3)}$ if p is odd and is proportional to $n^{(p+2)/(2p+5)}$ if p is even. For simplicity we consider only the case that p is odd.

Let $h = \mathrm{d}n^{-1/(2p+3)}$ and $c(n) = n^{(p+1)/(2p+3)}$. Recall that $\boldsymbol{\psi}(\cdot)$ is a vector of length q. Let $s_K(j) = \int z^j K(z)dz$ and $\gamma_K(j) = \int z^j K^2(z)dz$. Let $\mathbf{D}_p(s)$ be the $(p+1)\times(p+1)$ matrix with (j,k)th element equal to $s_K(j+k-2)$, $\mathbf{D}_p(\gamma)$ be the $(p+1)\times(p+1)$ matrix with (j,k)th element equal to $\gamma_K(j+k-2)$, and $\mathbf{D}_s(L) = \{s_K(L), s_K(L+1), \ldots, s_K(L+p)\}^t$. Define $\mathbf{R}(x) = E[\boldsymbol{\chi}\{\widetilde{\boldsymbol{\mathcal{Y}}}, \boldsymbol{\Theta}(X)\}|X = x]$ and $\mathbf{C}(x) = E[\boldsymbol{\psi}\{\widetilde{\boldsymbol{\mathcal{Y}}}, \boldsymbol{\Theta}(X)\}\boldsymbol{\psi}^t\{\widetilde{\boldsymbol{\mathcal{Y}}}, \boldsymbol{\Theta}(X)\}|X = x]$.

We now turn to the details of the proof. In our context, there is an equivalent to equation (4.2) of Härdle and Marron. Indeed, for p odd, Carroll, *et al.* (1998) show that

$$\begin{aligned}\widehat{\boldsymbol{\Theta}}(x) - \boldsymbol{\Theta}(x) \approx{}& \frac{\{1+o_p(1)\}\, h^{p+1}\mathbf{E}_{1,p+1,q}\left\{\mathbf{D}_p^{-1}(s)\mathbf{D}_s(p+1)\otimes\boldsymbol{\Theta}^{(p+1)}(x)\right\}}{(p+1)!}\\ &-\mathbf{E}_{1,p+1,q}\{f_X(x)\}^{-1}\left\{\mathbf{D}_p^{-1}(s)\otimes\mathbf{R}^{-1}(x)\right\} n^{-1}\sum_{i=1}^{n} K_h(X_i - x)\\ &\times\left[\mathbf{G}_p\{(X_i-x)/h\}\otimes\boldsymbol{\psi}\{\widetilde{\boldsymbol{\mathcal{Y}}}_i, \boldsymbol{\Theta}(X_i)\}\right]\{1+o_p(1)\}. \qquad (7)\end{aligned}$$

Using the argument of Härdle and Marron essentially without change, their Lemma 1 holds. In particular, the appropriate asymptotic distribution is

$$\begin{aligned}&c(n)\left\{\widehat{\boldsymbol{\Theta}}(x) - \boldsymbol{\Theta}(x)\right\} \to \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ where} \qquad (8)\\ \boldsymbol{\mu} &= d^{p+1}\mathbf{E}_{1,p+1,q}\left\{\mathbf{D}_p^{-1}(s)\mathbf{D}_s(p+1)\otimes\boldsymbol{\Theta}^{(p+1)}(x)\right\}/(p+1)!;\\ \boldsymbol{\Sigma} &= d\{f_X(x)\}^{-1}\mathbf{E}_{1,p+1,q}\left\{\mathbf{D}_p^{-1}(s)\otimes\mathbf{R}^{-1}(x)\right\}\{\mathbf{D}_p(\gamma)\otimes\mathbf{C}(x)\}\\ &\qquad\times\left\{\mathbf{D}_p^{-1}(s)\otimes\mathbf{R}^{-1}(x)\right\}^t\mathbf{E}_{1,p+1,q}^t\\ &= d\{f_X(x)\}^{-1}\mathbf{E}_{1,p+1,q}\left\{\mathbf{D}_p^{-1}(s)\mathbf{D}_p(\gamma)\mathbf{D}_p^{-1}(s)\otimes\mathbf{R}^{-1}(x)\mathbf{C}(x)\mathbf{R}^{-t}(x)\right\}\mathbf{E}_{1,p+1,q}^t.\end{aligned}$$

It is now important to remember exactly how we constructed the bootstrap estimate, namely via *ordinary* local polynomial kernel regression. Thus, Lemma 2 of Härdle and Marron also follows routinely, and the only question is whether the limit distribution is the same as calculated above. This easily follows by an application of (7). Because the bootstrap estimate is a ordinary local polynomial estimator, $\mathbf{R}(x)$ is replaced in (7) by the $q\times q$ identity matrix, $\boldsymbol{\Theta}(\cdot)$ is replaced by $\widehat{\boldsymbol{\Theta}}_g(\cdot)$ and the limit distribution is easily seen to equal (8), as required.

References

Carroll, R. J., Ruppert, D. and Welsh, A. H. (1998). Local estimating equations. *J. Amer. Statist. Assoc.* **93**, 214-227.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical Models in S.* Wadsworth, Pacific Grove.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Clarendon Press, Oxford.

Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88**, 1287-1301.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modeling and Its Applications.* Chapman and Hall, London.

Fan, J., Heckman, N. E. and Wand, M. P. (1995). Local polynomial kernel regression for generalized linear models and quasi-likelihood functions. *J. Amer. Statist. Assoc.* **90**, 141-150.

Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19**, 778-796.

Huber, P. J. (1967). The behavior of maximum likelihood estimation under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1** (Edited by L. M. LeCam and J. Neyman), 221-233, University of California Press.

Kauermann, G. and Tutz, G. (1999). On model diagnostics and bootstrapping in varying coefficient models. *Biometrika* **86**, 119-128.

Rosner, B., Willett, W. C. and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statist. Medicine* **8**, 1051-1070.

Ruppert, D. (1997). Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. *J. Amer. Statist. Assoc.* **92**, 1049-1062.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *J. Amer. Statist. Assoc.* **82**, 559-567.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1-25.

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis (with discussion). *Ann. Statist.* **14**, 1261-1295.

Department of Statistics, Texas A&M University, College Station TX 77843-3143, U.S.A.

E-mail: cgalindo@yahoo.com

E-mail: carroll@stat.tamu.edu

Department of Statistics and Robertson Centre, Boyd Orr Building, University of Glasgow, Glasgow G12 8QQ, U.K.

E-mail: goeran@stats.gla.ac.uk

(Received September 1998; accepted June 2000)