Estimating linear functionals of the error distribution in nonparametric regression

Ursula U. Müller
Universität Bremen

Anton Schick [*]
Binghamton University

Wolfgang Wefelmeyer
Universität Siegen

Abstract

This paper addresses estimation of linear functionals of the error distribution in nonparametric regression models. It derives an i.i.d. representation for the empirical estimator based on residuals, using undersmoothed estimators for the regression curve. Asymptotic efficiency of the estimator is proved. Estimation of the error variance is discussed in detail. In this case, undersmoothing is not necessary.

AMS 2000 subject classifications. Primary 62G05, 62G08, 62G20.

Key words and Phrases. Locally polynomial smoother, residual, empirical estimator, nonparametric regression estimator, leave-one-out estimator, variance estimator, influence function, gradient, efficient estimator.

1. Introduction

Consider a regression model $Y = r(X) + \varepsilon$, where the error ε has mean zero, finite variance, and an otherwise unknown distribution, and the covariate X is random and independent of ε. One is primarily interested in estimating the regression function r, but it is also of interest to estimate features of the distribution of the error. If we have independent observations $(X_1, Y_1), \dots, (X_n, Y_n)$ from the regression model, we can estimate first r by $\hat{r}$, say, and then the errors ε_i by the residuals $\hat{\varepsilon}_i = Y_i - \hat{r}(X_i)$. The expectation $Eh(\varepsilon)$ of some function h can then be estimated by the empirical estimator $\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i)$ based on the residuals. Most of the literature is on estimating the distribution function $F(t) = P(\varepsilon \le t)$, i.e. $h(z) = \mathbf{1}[z \le t]$, or the variance $E\varepsilon^2$, i.e. $h(z) = z^2$.

Estimation of the residuals is particularly simple if the regression function $r = r_\vartheta$ is known up to a finite-dimensional parameter ϑ. Then one usually has $n^{1/2}$-consistent estimators $\hat{\vartheta}$

[*] Supported in part by NSF Grant DMS 0072174

for ϑ and can estimate r by $r_{\hat{\vartheta}}$ and ε_i by $\varepsilon_i(\hat{\vartheta}) = Y_i - r_{\hat{\vartheta}}(X_i)$. Weak convergence of the empirical process based on the residuals of parametric regression models is studied by Koul (1970), Loynes (1980), and Shorack (1984), among others; see also the monograph Koul (2002). We refer to Ghoudi and Rémillard (1998) for general results on empirical processes based on "pseudo-observations". Analogous results exist for ergodic autoregressive models with parametric autoregression function; see Boldin (1982, 1983), Koul and Sen (1991), and Koul and Ossiander (1996) for linear autoregression, Boldin (1989) for moving average processes, Kreiss (1991) for general linear processes, Bai (1994) for ARMA models, Koul (1996) for nonlinear time series, and Boldin (1998) for ARCH models. Extensions to explosive autoregressive models are in Koul and Leventhal (1989).

Here we are concerned with the *nonparametric* regression model $Y = r(X) + \varepsilon$, with r unknown (up to smoothness). Then the problem arises that an estimator $\hat{r}$ for r will not be $n^{1/2}$-consistent any more, and hence the residuals $\hat{\varepsilon}_i = Y_i - \hat{r}(X_i)$ will differ from the true errors ε_i by more than the order $n^{-1/2}$. Nevertheless, the empirical estimator $\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i)$ based on the residuals will still be $n^{1/2}$-consistent for $Eh(\varepsilon)$ under appropriate conditions. This is an instance of the plug-in phenomenon: Smooth functionals of function estimators may have parametric rates. (There is a large literature on plug-in, especially for nonlinear functionals of densities and regression functions; we refer to Goldstein and Messer 1995, Birgé and Massart 1995, Eggermont and LaRiccia 1999, and Efromovich and Samarov 2000.) Specifically, in Section 2 we give conditions for the i.i.d. representation

$$n^{-1/2}\sum_{i=1}^n h(\hat{\varepsilon}_i) = n^{-1/2}\sum_{i=1}^n (h(\varepsilon_i) - E[h'(\varepsilon)]\varepsilon_i) + o_p(1). \tag{1.1}$$

Simple sufficient conditions would be that the regression function r has a bounded second derivative, the function h has a bounded derivative, and the covariate X is bounded with a density that is continuous and positive on its support. A discussion of our assumptions is in Section 3 where we construct explicit estimators using undersmoothed local polynomial smoothers. We assume that h is smooth. This excludes the distribution function $F(t) = P(\varepsilon \le t)$. Under stronger assumptions on the error and covariate distributions, a functional central limit theorem for the empirical distribution function based on the residuals from a heteroscedastic regression model is proved in Akritas and Van Keilegom (2001). For symmetric error distribution see also Koshevnik (1996).

We use representation (1.1) to prove, in Section 5, that $\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i)$ is asymptotically efficient if the error density f, the regression function r, and possibly the covariate density g, are unknown. Representation (1.1) implies that $\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i)$ is asymptotically normal, with variance $\tau_*^2 = E[(h(\varepsilon) - Eh(\varepsilon) - E[h'(\varepsilon)]\varepsilon)^2]$. Perhaps surprisingly, this variance can be considerably *smaller* than the asymptotic variance $\tau^2 = E[(h(\varepsilon) - Eh(\varepsilon))^2]$ of the empirical estimator $\frac{1}{n}\sum_{i=1}^n h(\varepsilon_i)$ based on the true errors. Suppose, for example, that the errors are normal with mean zero and variance σ^2. Consider estimating the third moment. Then $h(z) = z^3$. The empirical estimator $\frac{1}{n}\sum_{i=1}^n \varepsilon_i^3$ has asymptotic variance $15\sigma^6$. On the other hand, we have $Eh'(\varepsilon) = 3\sigma^2$; hence the asymptotic variance of $\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^3$ is $(15-9)\sigma^6 = 6\sigma^6$. This is a considerable variance *reduction.*

The paradox is explained by noting that the empirical estimator $\frac{1}{n}\sum_{i=1}^n h(\varepsilon_i)$ based on the *true* errors makes no use of the information that the errors have mean zero. We can use this information to introduce modified empirical estimators

$$H_n(c) = \frac{1}{n}\sum_{i=1}^n (h(\varepsilon_i) - c\varepsilon_i).$$

Their variance is minimized by

$$c = \int zh(z)dF(z)/\sigma^2. \tag{1.2}$$

This constant depends on F and must be estimated, e.g. by a ratio of empirical estimators,

$$\hat{c}_n = \frac{1}{n}\sum_{i=1}^n \varepsilon_i h(\varepsilon_i)\Big/\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2.$$

The resulting plug-in version $H_n(\hat{c}_n)$ (which still uses the actual errors) is never worse than $\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i)$. We prove this at the end of Section 5. There we also compare $\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i)$ and $H_n(\hat{c}_n)$ in terms of the information in knowing the regression function and in knowing that the errors have mean zero. It is quite obvious that $\frac{1}{n}\sum_{i=1}^n h(\varepsilon_i)$ does not use that the errors have mean zero. But how does our estimator $\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i)$ exploit this information? This can be seen from condition (2.6): The estimator for the regression function is constructed such that $\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i = o_p(n^{-1/2})$. This means that we have a better "estimator" for zero than $\frac{1}{n}\sum_{i=1}^n \varepsilon_i$, which is $O_p(n^{-1/2})$. The faster rate $\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i = o_p(n^{-1/2})$ might be surprising, but we should remind the reader that for least squares estimates one has even $\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i = 0$. In view of this faster rate, replacement of the actual errors in $H_n(\hat{c}_n)$ by residuals leads to an estimate equivalent to $\frac{1}{n}\sum_{i=1}^n h(\hat{\varepsilon}_i)$.

It is easy to check that $\tau_*^2 = \tau^2$ if and only if $E[h'(\varepsilon)] = 0$ or $\sigma^2 E[h'(\varepsilon)] = 2E[h(\varepsilon)\varepsilon]$, and that $\tau_*^2 < \tau^2$ if and only if $0 < \sigma^2 E[h'(\varepsilon)] < 2E[h(\varepsilon)\varepsilon]$ or $0 > \sigma^2 E[h'(\varepsilon)] > 2E[h(\varepsilon)\varepsilon]$. For normal errors one can show that $\sigma^2 E[h'(\varepsilon)] = E[h(\varepsilon)\varepsilon]$ under the assumptions needed for the representation. Then the first inequality holds if $E[h(\varepsilon)\varepsilon] < 0$, and the second holds if $E[h(\varepsilon)\varepsilon] > 0$. Thus, if the errors happen to be normal, then $\tau_*^2 \le \tau^2$ so that the empirical estimator based on the residuals has an asymptotic variance that is never bigger than that of the empirical estimator based on the actual errors. However, one cannot always expect that $\tau_*^2 \le \tau^2$ for other densities. For example, take $h(x) = x^2\mathrm{sgn}(x) - 3x$ and $f(x) = \exp(-|x|)/2$ for a case where $\tau_*^2 > \tau^2$.

It is instructive to compare our result with corresponding results for *parametric* regression $Y_i = r_\vartheta(X_i) + \varepsilon_i$. For simplicity we take ϑ one-dimensional. A straightforward Taylor expansion gives

$$n^{-1/2} \sum_{i=1}^n h(\varepsilon_i(\hat\vartheta)) = n^{-1/2} \sum_{i=1}^n h(\varepsilon_i) - n^{1/2}(\hat\vartheta - \vartheta) Eh'(\varepsilon) E\dot r_\vartheta(X) + o_p(1),$$

with $\dot r_\vartheta(x)$ the derivative of $r_\vartheta(x)$ with respect to ϑ. For linear regression and the empirical distribution function, i.e. $h(z) = \mathbf{1}[z \le t]$, see e.g. Koul (1969, 1970, 1987) and also Shorack and Wellner (1986, Section 4.6). We refer to Mammen (1996) for such a stochastic expansion in linear models of increasing dimension. Nonlinear (and heteroscedastic) autoregression and smooth functions h are considered in Schick and Wefelmeyer (2002). In general, the empirical estimator based on the residuals is *not* efficient, even if an efficient estimator $\hat\vartheta$ for ϑ is used. Again the reason is that, unlike $\hat r$, the estimator $r_{\hat\vartheta}$ for r_ϑ does not use the information that the errors have mean zero. Efficient modifications are in Wefelmeyer (1994) and Schick and Wefelmeyer (2002) for linear and nonlinear autoregression, respectively.

The function $h(z) = z^2$ is of particular interest. Then $Eh(\varepsilon) = E\varepsilon^2$ is the error variance $\sigma^2 = \int z^2 dF(z)$. This function is a degenerate special case: It is the only function h for which $Eh'(\varepsilon)$ vanishes for all error distributions with zero mean. This means that the first-order term in the stochastic expansion (1.1) vanishes. In particular, the empirical estimator $\frac{1}{n}\sum_{i=1}^n \hat\varepsilon_i^2$ is adaptive with respect to the regression function in the sense that it is asymptotically equivalent to $\frac{1}{n}\sum_{i=1}^n \varepsilon_i^2$, the best estimator based on the true errors, i.e. for known regression function. We show in Section 4 that in this simple case we get by with weaker assumptions and can avoid undersmoothing the regression estimator. If the errors have finite fourth moments, then $\frac{1}{n}\sum_{i=1}^n \hat\varepsilon_i^2$ is asymptotically normal with mean zero and variance $\int z^4 dF(z) - \sigma^4$. The asymptotic variance of this estimator has been calculated before; see Buckley, Eagleson and Silverman (1988) for normal errors, and Hall and Marron (1990). There are many papers on simpler, difference-based, estimators with larger asymptotic variances; see Hall, Kay and Titterington (1990) and the references there. For comparisons of different estimators see Carter and Eagleson (1992) and Dette, Munk and Wagner (1998, 1999). An efficient version of a difference-based estimator is in Müller, Schick and Wefelmeyer (2001).

Let us briefly mention possible applications and extensions. We have already discussed the degenerate case of the error variance. Our result in Section 2 applies also to other moments and absolute moments. It leads in particular to efficient estimators for skewness $E(\varepsilon^3)/\sigma^3$ and kurtosis $E(\varepsilon^4)/\sigma^4$. We can also use it to estimate the characteristic function $E[\exp(it\varepsilon)]$ and other such transformations of the error distribution. This can for example be used to test normality of the errors. Minimum contrast functionals are defined as minimizers in t of expectations $Eh_t(\varepsilon)$; they can be estimated by minimizers of $\frac{1}{n}\sum_{i=1}^n h_t(\hat\varepsilon_i)$. Here one would want a version of our result that is uniform over a class of functions h. The distribution of such estimators can also be studied by writing them as integrals with respect to the empirical distribution function, but results on the latter require stronger conditions.

2. Result

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be independent observations from the nonparametric regression model $Y = r(x) + \varepsilon$, with an error ε that has mean zero and finite variance, and is independent of the random covariate X. Our goal is to estimate $E[h(\varepsilon)]$. We consider the estimator

$$\hat{H}_n = \frac{1}{n} \sum_{i=1}^{n} h(\hat{\varepsilon}_i),$$

based on the residuals $\hat{\varepsilon}_i = Y_i - \hat{r}_{ni}$ where $\hat{r}_{ni}$ estimates $r(X_i)$.

We make the following assumptions on h and the estimators $\hat{r}_{ni}$. Set

$$\mu(s) = \int h(z - s) f(z)\, dz, \quad s \in \mathbf{R}.$$

Assumption 1. There are positive numbers $\alpha \leq 1$, c, C_1, C_2 such that

$$\int (h(z+t+s) - h(z+t))^2 f(z)\, dz \leq C_1 |s|^{1+\alpha}, \quad |s|, |t| \leq c, \tag{2.1}$$

and μ is differentiable at 0 with

$$|\mu(s) - \mu(0) - \mu'(0)s| \leq C_2 |s|^{1+\alpha}, \quad |s| \leq c. \tag{2.2}$$

We write $\mathbf{X} = (X_1, \dots, X_n)$, $\mathbf{Y} = (Y_1, \dots, Y_n)$ and $\mathbf{Y}_{-i} = (Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n)$ for the vector with Y_i left out. Set

$$\hat{r}_{nij} = E(\hat{r}_{ni} | \mathbf{X}, \mathbf{Y}_{-j}).$$

Assumption 2. The estimator $\hat{r}_{ni}$ is a leave-one-out estimator in the sense that it does not depend on Y_i, so that $\hat{r}_{nii} = \hat{r}_{ni}$. Moreover, for α as in Assumption 1,

$$\max_{i=1,\dots,n} |\hat{r}_{ni} - r(X_i)| = o_p(1), \tag{2.3}$$

$$\frac{1}{n} \sum_{i=1}^{n} |\hat{r}_{ni} - r(X_i)|^{1+\alpha} = o_p(n^{-1/2}), \tag{2.4}$$

$$\frac{1}{n} \sum\sum_{i \neq j} E(|\hat{r}_{nij} - \hat{r}_{ni}|^{1+\alpha} | \mathbf{X}) = o_p(1), \tag{2.5}$$

$$n^{-1/2} \sum_{i=1}^{n} (\hat{r}_{ni} - r(X_i)) - n^{-1/2} \sum_{i=1}^{n} \varepsilon_i = o_p(1). \tag{2.6}$$

The leave-one-out estimator is chosen for technical convenience. The last condition (2.6) is equivalent to $\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i = o_p(n^{-1/2})$, a condition we already stressed in the Introduction. Assumption (2.5) is similar to an assumption used in Condition R of Schick (1993). The proof of the following theorem relies also on his arguments. A detailed discussion of Assumptions 1 and 2 is in the next section.

Theorem 1. *Suppose Assumptions 1 and 2 hold. Then*

$$\hat{H}_n = \frac{1}{n}\sum_{i=1}^{n}(h(\varepsilon_i) + \mu'(0)\varepsilon_i) + o_p(n^{-1/2}).$$

Consequently, $n^{1/2}(\hat{H}_n - Eh(\varepsilon))$ *is asymptotically normal with mean zero and variance*

$$\tau_*^2 = E[(h(\varepsilon) - Eh(\varepsilon) + \mu'(0)\varepsilon)^2].$$

3. Discussion of the assumptions

Let us first look at Assumption 1. If h is Lipschitz, then assumption (2.1) holds with $\alpha = 1$. If h is absolutely continuous and

$$\int \sup_{|t|\le 2c} |h'(z+t)|^2 f(z)\,dz < \infty,$$

then assumption (2.1) holds with $\alpha = 1$. If h is twice differentiable with an f-integrable derivative h', and if

$$\int \sup_{|s|\le c} |h''(z-s)| f(z)\,dz < \infty,$$

then assumption (2.2) holds with $\alpha = 1$ and $\mu'(0) = -\int h'(z)f(z)\,dz$. Differentiability of h is not required if f is sufficiently smooth. For example, if f is twice continuously differentiable with integrable derivatives, then assumption (2.2) holds with $\alpha = 1$ for every bounded h.

If $\int |z|^{2m} f(z)\,dz$ is finite for an integer $m > 1$, then the function $h(z) = z^m$ fulfills assumption (2.1) with $\alpha = 1$. In this case, assumption (2.2) holds as well with $\alpha = 1$, and $\mu'(0) = -m\int z^{m-1} f(z)\,dz$. Similarly, Assumption 1 holds for $h(z) = |z|^m$ with $\alpha = 1$ and $\mu'(0) = -m\int |z|^{m-1}\mathrm{sgn}(z) f(z)\,dz$.

Let us now address Assumption 2. Using the moment inequality, a sufficient condition for assumptions (2.4) and (2.5) is

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{r}_{ni} - r(X_i))^2 = o_p(n^{-1/(1+\alpha)}), \tag{3.1}$$

$$\frac{1}{n}\sum\sum_{i\neq j} E((\hat{r}_{nij} - \hat{r}_{ni})^2|\mathbf{X}) = o_p(n^{-1/(1+\alpha)}). \tag{3.2}$$

Now consider linear smoothers

$$\hat{r}_{ni} = \sum_{j=1}^{n} A_{nij} Y_j = \sum_{j=1}^{n} A_{nij} r(X_j) + \sum_{j=1}^{n} A_{nij}\varepsilon_j, \tag{3.3}$$

where A_{nij} depends on the covariates only. For $\hat{r}_{ni}$ to be a leave-one-out estimator, we must have $A_{nii} = 0$. Sufficient conditions for assumptions (2.3) to (2.6) are

$$\max_{i=1,\dots,n} \Big| \sum_{j=1}^{n} A_{nij}\varepsilon_j \Big| \;=\; o_p(1), \tag{3.4}$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{nij}^2 \;=\; o_p(n^{-1/(1+\alpha)}), \tag{3.5}$$

$$\frac{1}{n}\sum_{j=1}^{n}\Big(1 - \sum_{i=1}^{n} A_{nij}\Big)^2 \;=\; o_p(1), \tag{3.6}$$

$$\max_{i=1,\dots,n} \Big| \sum_{j=1}^{n} A_{nij} r(X_j) - r(X_i) \Big| \;=\; o_p(n^{-1/2}). \tag{3.7}$$

Indeed, relations (3.4) and (3.7) imply (2.3). Relations (3.5) and (3.7) imply (3.1) in view of

$$\frac{1}{n}\sum_{i=1}^{n} E((\hat{r}_{ni} - r(X_i))^2 | \mathbf{X}) = \sigma^2 \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{nij}^2 + \frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n} A_{nij} r(X_j) - r(X_i)\Big)^2.$$

Since $\hat{r}_{nij} - \hat{r}_{ni} = -A_{nij}\varepsilon_j$, we immediately see that (3.5) implies (3.2). In view of (3.7), relation (2.6) is equivalent to

$$n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{nij}\varepsilon_j - n^{-1/2}\sum_{j=1}^{n}\varepsilon_j = n^{-1/2}\sum_{j=1}^{n}\varepsilon_j\Big(\sum_{i=1}^{n} A_{nij} - 1\Big) = o_p(1).$$

But the conditional second moment given $\mathbf{X}$ of this expression equals σ^2 times the left-hand side of (3.6).

Assumption (3.7) controls the maximal bias and requires smoothness of r. Let L be a non-negative integer, and let b_n denote a bandwidth. Assume that r is L-times differentiable with L-th derivative Hölder with exponent β, and that

$$\textstyle\sum_{j=1}^{n} A_{nij} = 1, \tag{3.8}$$

$$\textstyle\sum_{j=1}^{n} A_{nij}(X_j - X_i)^l = 0, \quad l = 1, \dots, L, \quad \text{if } L > 0, \tag{3.9}$$

$$A_{nij} = 0 \quad \text{if } |X_j - X_i| > b_n, \tag{3.10}$$

$$\textstyle\max_{i=1,\dots,n} \sum_{j=1}^{n} |A_{nij}| = O_p(1). \tag{3.11}$$

Under (3.8) and (3.9) we can write

$$\sum_{j=1}^{n} A_{nij} r(X_j) - r(X_i) = \sum_{j=1}^{n} A_{nij}\big(r(X_j) - r(X_i) - \sum_{l=1}^{L}(X_j - X_i)^l r^{(l)}(X_i)\big).$$

Now use (3.10) and the Hölder continuity of $r^{(L)}$ to bound the left-hand side of (3.7) by a constant times $b_n^{L+\beta}\max_{i=1,\dots,n}\sum_{j=1}^n |A_{nij}|$. Under (3.11), this is of order $O_p(b_n^{L+\beta})$. Hence condition (3.7) holds if $n^{1/2}b_n^{L+\beta}\to 0$. Under standard assumptions, local polynomial smoothers fulfill (3.8) to (3.11) if the bandwidth b_n does not converge to zero too fast. Under the above assumptions on r, the optimal bandwidth is proportional to $n^{-1/(2L+2\beta+1)}$. For this choice, we do not have $n^{1/2}b_n^{L+\beta}\to 0$. This means that we must undersmooth this type of estimator.

To be specific, we introduce explicit weights A_{nij} and give assumptions on g and r such that Assumption 2 holds. The weights will be those of a leave-one-out local polynomial smoother of degree L. To define them, introduce

$$\Lambda_{ni}(\beta_0,\dots,\beta_L)=\sum_{\substack{j=1\\ j\neq i}}^n\Big(Y_j-\sum_{l=0}^L\beta_l\Big(\frac{X_j-X_i}{b_n}\Big)^l\Big)^2 w\Big(\frac{X_j-X_i}{b_n}\Big),\quad \beta_0,\dots,\beta_L\in\mathbf{R},$$

for some symmetric density w with support $[-1,1]$ and bounded derivative. Denote the minimizer of $\Lambda_{ni}(\beta_0,\dots,\beta_L)$ by $(\hat\beta_{ni0},\dots,\hat\beta_{niL})$. We estimate $r(X_i)$ by $\hat r_{ni}=\hat\beta_{ni0}$. For $l=0,1,2,\dots$ and $i=1,\dots,n$ let

$$w_{nl}(x)=w\Big(\frac{x}{b_n}\Big)\frac{x^l}{b_n^{l+1}}\quad\text{and}\quad p_{nil}(x)=\frac{1}{n-1}\sum_{\substack{j=1\\ j\neq i}}^n w_{nl}(X_j-x).$$

The estimator $\hat r_{ni}=\hat\beta_{ni0}$ is a linear smoother with weights

$$A_{nij}=\frac{1}{n-1}\sum_{l=0}^L q_{nil}w_{nl}(X_j-X_i),\quad i\neq j,$$

where $(q_{ni0},\dots,q_{niL})$ is the first row of the inverse of the $(L+1)\times(L+1)$ matrix $M_{ni}=[p_{ni,l+k}(X_i)]_{l,k=0,\dots,L}$. If $L=0$, this is the usual leave-one-out kernel estimator:

$$\hat r_{ni}=\sum_{\substack{j=1\\ j\neq i}}^n w\Big(\frac{X_i-X_j}{b_n}\Big)Y_j\Big/\sum_{\substack{j=1\\ j\neq i}}^n w\Big(\frac{X_i-X_j}{b_n}\Big).$$

Theorem 2. *Suppose g is continuous and positive on its support $[0,1]$. Assume that r is L times differentiable on $[0,1]$ and that its L-th derivative is Hölder with positive exponent β. Suppose the bandwidth b_n fulfills $n^{1/2}b_n^{L+\beta}\to 0$, $b_n n^{1/3}\to\infty$, and $b_n n^{\alpha/(1+\alpha)}\to\infty$. Then Assumption 2 holds for the leave-one-out local polynomial smoother $\hat r_{ni}=\hat\beta_{ni0}$ of degree L described above. Hence, if h satisfies Assumption 1, we have again*

$$\hat H_n=\frac{1}{n}\sum_{i=1}^n(h(\varepsilon_i)+\mu'(0)\varepsilon_i)+o_p(n^{-1/2}).$$

The assumptions on the bandwidth in the above theorem can only be met if $2L + 2\beta$ is sufficiently large. If $\alpha \geq 1/2$, we need $2L + 2\beta > 3$ and can then take b_n proportional to $n^{-\gamma}$ with γ in the open interval $(1/(2L + 2\beta), 1/3)$. If $\alpha < 1/2$, we need $2L + 2\beta > (\alpha + 1)/\alpha$ and can then take b_n proportional to $n^{-\gamma}$ for some γ in the open interval $(1/(2L+2\beta), \alpha/(1+\alpha))$. In particular, if r is twice continuously differentiable on $[0, 1]$ and $\alpha \geq 1/2$, then we can take $b_n = n^{-\gamma}$ with $\gamma \in (1/4, 1/3)$. The proof of Theorem 2 is sketched in Section 6.

4. Estimating the error variance

An important special case is the estimation of the error variance $\sigma^2 = \int z^2 f(z)\,dz$. We have already seen that in this case Theorem 1 shows that the estimator $\hat{H}_n = \frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2$ based on the residuals is asymptotically equivalent to the empirical estimator based on the true errors,

$$\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 + o_p(n^{-1/2}). \tag{4.1}$$

For this result, a simpler, more direct, argument can be given. Moreover, this argument avoids the undersmoothing of the regression estimator needed in Section 3. Write

$$\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^n \varepsilon_i^2 - \frac{2}{n}\sum_{i=1}^n \varepsilon_i(\hat{r}_{ni} - r(X_i)) + \frac{1}{n}\sum_{i=1}^n (\hat{r}_{ni} - r(X_i))^2.$$

One sees that (4.1) holds if

$$V_1 = \frac{1}{n}\sum_{i=1}^n (\hat{r}_{ni} - r(X_i))^2 = o_p(n^{-1/2}), \tag{4.2}$$

$$V_2 = \frac{1}{n}\sum_{i=1}^n \varepsilon_i(\hat{r}_{ni} - r(X_i)) = o_p(n^{-1/2}). \tag{4.3}$$

The argument used for (6.3), but now with $D_{ni} = \varepsilon_i(\hat{r}_{ni} - r(X_i))$, gives

$$nE(V_2^2|\mathbf{X}) \leq \sigma^2 \frac{1}{n}\sum_{i=1}^n E((\hat{r}_{ni} - r(X_i))^2|\mathbf{X}) + \sigma^2 \frac{1}{n}\sum\sum_{i \neq j} E((\hat{r}_{nij} - \hat{r}_{ni})^2|\mathbf{X}).$$

Thus we have the following result.

Theorem 3. *Suppose $\hat{r}_{ni}$ does not depend on Y_i, and*

$$\frac{1}{n}\sum_{i=1}^n E((\hat{r}_{ni} - r(X_i))^2|\mathbf{X}) = o_p(n^{-1/2}), \tag{4.4}$$

$$\frac{1}{n}\sum\sum_{i \neq j} E((\hat{r}_{nij} - \hat{r}_{ni})^2|\mathbf{X}) = o_p(1). \tag{4.5}$$

Then (4.1) holds:

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + o_p(n^{-1/2}).$$

If $\hat{r}_{ni}$ is the linear smoother (3.3) again, with $A_{nii} = 0$, then sufficient conditions for (4.4) and (4.5) are

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}A_{nij}^2 = o_p(n^{-1/2}), \tag{4.6}$$

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\sum_{j=1}^{n}A_{nij}r(X_j) - r(X_i)\Big)^2 = o_p(n^{-1/2}). \tag{4.7}$$

If conditions (3.8), (3.10) and (3.11) hold and r is Hölder with exponent β, then the left-hand side of (4.7) is of order $O_p(b_n^{2\beta})$. Since the left-hand side of (4.6) is typically of order $O_p(n^{-1}b_n^{-1})$, one needs $n^{1/2}b_n^{2\beta} \to 0$ and $n^{1/2}b_n \to \infty$ for (4.6) and (4.7) to hold. This of course requires $\beta > 1/2$.

Corollary 1. *Suppose g and r are as in Theorem 2 with $L + \beta > 1/2$ and $\hat{r}_{ni}$ is the leave-one-out polynomial estimator of degree L with bandwidth b_n such that $n^{1/2}b_n^{L+\beta} \to 0$ and $n^{1/2}b_n \to \infty$. Then (4.1) holds:*

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 + o_p(n^{-1/2}).$$

The requirements on the bandwidth are satisfied for the optimal bandwidth which in this case is proportional to $n^{-1/(2L+2\beta+1)}$. So no undersmoothing is needed here.

5. Efficiency

Now we show that $\hat{H}_n$ is efficient for $\int h(z)f(z)\,dz$ under the additional assumption that f has finite Fisher information for location:

Assumption 3. The error density f is absolutely continuous with almost everywhere derivative f', and $J = \int \ell(z)^2 f(z)\,dz < \infty$, where $\ell(z) = -f'(z)/f(z)$.

We recall that the nonparametric regression model (1.1) is locally asymptotically normal in the following sense. Let F and G denote the distribution functions associated with the densities f and g. Introduce a local model at (r, f, g) by perturbing (r, f, g) as follows. For $u \in L_2(G)$ set

$$r_{nu}(z) = r(z) + n^{-1/2}u(z).$$

Set

$$\begin{aligned} L_{2,0}(F) &= \{v \in L_2(F) : \int v(z)\, dF(z) = 0\}, \\ V &= \{v \in L_{2,0}(F) : \int zv(z)\, dF(z) = 0\}. \end{aligned}$$

Note that the projections of h and ℓ onto V are

$$h_0(z) = h(z) - \int hdF - z\int xh(x)dF(x)/\sigma^2 \quad \text{and} \quad \ell_0(z) = \ell(z) - z/\sigma^2.$$

Write

$$J_0 = \int \ell_0(z)dF(z) = J - 1/\sigma^2.$$

For $v \in V$ choose densities f_{nv} with $\int zf_{nv}(z)dz = 0$ which are Hellinger differentiable at f with derivative v,

$$\int \Big(n^{1/2}\big(f_{nv}(z)^{1/2} - f(z)^{1/2}\big) - \frac{1}{2}\, v(z)f(z)^{1/2}\Big)^2 dz \to 0.$$

For $w \in L_{2,0}(G)$ choose densities g_{nw} which are Hellinger differentiable at g with derivative w,

$$\int \Big(n^{1/2}\big(g_{nw}(x)^{1/2} - g(x)^{1/2}\big) - \frac{1}{2}\, w(x)g(x)^{1/2}\Big)^2 dx \to 0.$$

Now let $p_{rfg}(x, y) = f(y - r(x))g(x)$ denote the density of an observation (X_i, Y_i) from our model (1.1). It follows from arguments in Hájek and Šidák (1967, pages 210–214), see also Schick (1993, Lemma 3.4), that $p_{r_{nu}f_{nv}g_{nw}}$ is Hellinger differentiable at (r, f, g),

$$\begin{aligned} \int \Big(&n^{1/2}(p_{r_{nu}f_{nv}g_{nw}}(x, y)^{1/2} - p_{rfg}(x, y)^{1/2}) \\ &- \frac{1}{2}\, \big(\ell(y - r(z))u(z) + v(y - r(z)) + w(z)\big)p_{rfg}(x, y)^{1/2}\Big)^2 dydz \to 0. \end{aligned}$$

Hence we have the stochastic expansion

$$\begin{aligned} \sum_{i=1}^n \log \frac{p_{r_{nu}f_{nv}g_{nw}}}{p_{rfg}}(X_i, Y_i) &= n^{-1/2}\sum_{i=1}^n \big(\ell(\varepsilon_i)u(X_i) + v(\varepsilon_i) + w(X_i)\big) \\ &\quad -\frac{1}{2}\, E\big[\big(\ell(\varepsilon)u(X) + v(\varepsilon) + w(X)\big)^2\big] + o_p(1), \end{aligned} \tag{5.1}$$

and $n^{-1/2} \sum_{i=1}^n \big(\ell(\varepsilon_i)u(X_i) + v(\varepsilon_i) + w(X_i)\big)$ is asymptotically normal by the central limit theorem. This is local asymptotic normality (LAN). We call the norm of (u, v, w) in (5.1) the *LAN norm.* Since ε and X are independent with $E\ell(\varepsilon) = Ev(\varepsilon) = 0$, the LAN norm can be written

$$\begin{aligned} \|(u, v, w)\|^2_{\text{LAN}} &= E\big[\big(\ell(\varepsilon)u(Z) + v(\varepsilon) + w(Z)\big)^2\big] \\ &= J\int u^2dG + 2\int \ell_0 vdF \int udG + \int v^2dF + \int w^2dG. \end{aligned}$$

Here we have replaced ℓ by its projection ℓ_0 onto V. The corresponding *LAN inner product* is

$$\begin{aligned}((u,v,w),(u_1,v_1,w_1))_{\mathrm{LAN}} &= J\int uu_1dG + \int \ell_0 v dF \int u_1 dG + \int \ell_0 v_1 dF \int u dG \\ &\quad + \int vv_1 dF + \int ww_1 dG.\end{aligned}$$

The *natural inner product* of the local model is

$$((u,v,w),(u_1,v_1,w_1)) = \int uu_1dG + \int vv_1dF + \int ww_1dG.$$

Consider a real-valued functional χ of (r,f,g). Suppose χ is *differentiable* at (r,f,g) with *natural gradient* $(u_*,v_*,w_*) \in L_2(G) \times V \times L_{2,0}(G)$,

$$\begin{aligned} n^{1/2}(\chi(r_{nu},f_{nv},g_{nw}) - \chi(r,f,g)) &\to ((u,v,w),(u_*,v_*,w_*)) \\ &\qquad \text{for all } (u,v,w) \in L_2(G) \times V \times L_{2,0}(G).\end{aligned}$$

The *LAN gradient* $(u_\chi,v_\chi,w_\chi) \in L_2(G) \times V \times L_{2,0}(G)$ of χ at (r,f,g) is the gradient of χ expressed in terms of the LAN inner product, and determined by

$$\begin{aligned} ((u,v,w),(u_*,v_*,w_*)) &= ((u,v,w),(u_\chi,v_\chi,w_\chi))_{\mathrm{LAN}} \\ &\qquad \text{for all } (u,v,w) \in L_2(G) \times V \times L_{2,0}(G). \end{aligned} \tag{5.2}$$

Call an estimator $\hat\chi$ *regular* for χ at (r,f,g) with *limit* L if L is a random variable such that

$$\begin{aligned} n^{1/2}(\hat\chi - \chi(r_{nu},f_{nv},g_{nw})) &\Rightarrow L \quad \text{under } P_{r_{nu}f_{nv}g_{nw}} \\ &\qquad \text{for all } (u,v,w) \in L_2(G) \times V \times L_{2,0}(G).\end{aligned}$$

The convolution theorem (see Bickel et al., 1998, Section 3.3) says that L is the convolution of a normal random variable with mean zero and variance $\|(u_\chi,v_\chi,w_\chi)\|^2_{\mathrm{LAN}}$, and some other random variable. This justifies calling $\hat\chi$ *efficient* for χ at (r,f,g) if $\hat\chi$ is regular and asymptotically normal with mean zero and variance $\|(u_\chi,v_\chi,w_\chi)\|^2_{\mathrm{LAN}}$. It also follows from the convolution theorem that $\hat\chi$ is efficient if and only if

$$n^{1/2}(\hat\chi - \chi(r,f,g)) = n^{-1/2}\sum_{i=1}^n \big(\ell(\varepsilon_i)u_\chi(X_i) + v_\chi(\varepsilon_i) + w_\chi(X_i)\big) + o_p(1). \tag{5.3}$$

The LAN gradient (u_χ,v_χ,w_χ) can be written in terms of the natural gradient (u_*,v_*,w_*) of χ as follows. Set $u,v=0$ in (5.2) to obtain

$$\int ww_*dG = \int ww_\chi dG \quad \text{for all } w \in L_{2,0}(G),$$

and hence $w_\chi = w_*$. Set $u,w=0$ in (5.2) to obtain

$$\int vv_*dF = \int \ell_0 v dF \int u_\chi dG + \int vv_\chi dF \quad \text{for all } v \in V,$$

and hence

$$v_\chi = v_* - \ell_0 \int u_\chi dG.$$

To calculate u_χ, it will be convenient to write $L_2(G)$ as an orthogonal sum of functions with mean zero and of constants, $L_2(G) = L_{2,0}(G) \oplus [1]$, or $u = u - \int udG + \int udG$. Set $v, w = 0$ in (5.2) and use $J - J_0 = \sigma^{-2}$ to obtain

$$\begin{aligned} \int uu_* dG &= \int \Big(u - \int udG\Big)\Big(u_* - \int u_* dG\Big) dG + \int udG \int u_* dG \\ &= J \int \Big(u - \int udG\Big)\Big(u_\chi - \int u_\chi dG\Big) dG + \sigma^{-2} \int udG \int u_\chi dG \\ &\quad + \int \ell_0 v_* dF \int udG \quad \text{for all } u \in L_2(G). \end{aligned}$$

This implies

$$\begin{aligned} u_\chi - \int u_\chi dG &= \frac{1}{J}\Big(u_* - \int u_* dG\Big), \\ \int u_\chi dG &= \sigma^2 \Big(\int u_* dG - \int \ell_0 v_* dF \Big). \end{aligned}$$

Hence we have the following result.

Proposition 1. *If the functional χ has natural gradient (u_*, v_*, w_*), then the LAN gradient of χ is (u_χ, v_χ, w_*) with*

$$\begin{aligned} u_\chi &= \frac{1}{J}\Big(u_* - \int u_* dG\Big) + \sigma^2 \Big(\int u_* dG - \int \ell_0 v_* dF \Big), \\ v_\chi &= v_* - \ell_0 \sigma^2 \Big(\int u_* dG - \int \ell_0 v_* dF \Big). \end{aligned}$$

We are interested in the functional $\chi(r, f, g) = Eh(\varepsilon) = \int h(z) dF(z)$. We have

$$n^{1/2} \Big(\int h(z) dF_{nv}(z) - \int h(z) dF(z) \Big) \to \int h(z) v(z) dF(z) = \int h_0(z) v(z) dF(z).$$

Hence $Eh(\varepsilon)$ is differentiable with natural gradient $(u_*, v_*, w_*) = (0, h_0, 0)$. From Proposition 1 we obtain the following result.

Corollary 2. *The LAN gradient of $Eh(\varepsilon)$ is $(u_h, v_h, 0)$ with*

$$u_h = -\sigma^2 \int \ell_0 h_0 dF \quad \textit{and} \quad v_h = h_0 + \ell_0 \sigma^2 \int \ell_0 h_0 dF.$$

Hence by (5.3) an efficient estimator $\hat{H}_n$ of $Eh(\varepsilon)$ is characterized by

$$
\begin{aligned}
(5.4) \qquad n^{1/2}(\hat{H}_n - Eh(\varepsilon)) &= n^{-1/2}\sum_{i=1}^{n}\big(\ell(\varepsilon_i)u_h(Z_i) + v_h(\varepsilon_i) + w_h(Z_i)\big) + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}\Big(h_0(\varepsilon_i) - (\ell(\varepsilon_i) - \ell_0(\varepsilon_i))\sigma^2 \int \ell_0 h_0 dF\Big) + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}\Big(h_0(\varepsilon_i) - \varepsilon_i \int \ell_0 h_0 dF\ \Big) + o_p(1)\\
&= n^{-1/2}\sum_{i=1}^{n}\Big(h(\varepsilon_i) - \int hdF - \varepsilon_i \int \ell hdF\Big) + o_p(1).
\end{aligned}
$$

For the last equation we have used $\ell(z) - \ell_0(z) = z/\sigma^2$ and $\int \ell_0 h_0 dF = \int \ell_0 hdF = \int \ell hdF - \sigma^{-2}\int zh(z)dF(z)$.

Finite Fisher information for location implies continuous Hellinger differentiability for the location model. Hence by assumption (2.1) and Lemma 7.2 in Ibragimov and Hasminskii (1981, p. 67), we have

$$\mu'(0) = -\int \ell hdF.$$

Comparing the characterization (5.4) with the i.i.d. representation in Theorem 1, we arrive at the following result:

Theorem 4. *Suppose Assumptions 1 to 3 hold. Then the estimator $\frac{1}{n}\sum_{i=1}^{n} h(\hat{\varepsilon}_i)$ introduced in Section 2 is efficient.*

The function $h_\chi(z) = h(z) - \int hdF - z\int \ell hdF$ in (5.4) is called the *efficient influence function* for estimators of $Eh(\varepsilon)$. Suppose now that we know the regression function r. Then we can observe the errors, and our problem reduces to estimating $Eh(\varepsilon)$ from i.i.d. observations $\varepsilon_1, \dots, \varepsilon_n$. The efficient influence function, say h_0, is obtained as above, now keeping r fixed: $h_0(z) = h(z) - \int hdF - cz$, with c defined as in (1.2). This shows that the improved empirical estimator $H_n(\hat{c}_n)$ introduced at the end of Section 2 is efficient if r is known. This result is due to Levit (1975). We have the orthogonal representation

$$h_\chi(z) = h_0(z) - z\int \ell_0 hdF.$$

Hence the variance increase of our estimator $\hat{H}_n$ over $H_n(\hat{c}_n)$ is $\sigma^2(\int \ell_0 hdF)^2$.

The regression model is called *adaptive* with respect to r if we can estimate $Eh(\varepsilon)$ for all h as well not knowing r as knowing r. This is the case only if $\ell_0 = 0$, i.e. if $\ell(z)$ is proportional to z, i.e. for normal error distribution.

6. Proofs

Proof of Theorem 1. Write

$$\Delta_{ni} = \hat{r}_{ni} - r(X_i) = \varepsilon_i - \hat{\varepsilon}_i.$$

By assumption (2.3), there is a sequence $c_n \downarrow 0$ such that $P(\max_{i=1,\dots,n} |\Delta_{ni}| > c_n) \to 0$. Without loss of generality we may assume that $c_n \le c$, where c is the constant appearing in Assumption 1. Now set

$$\hat{H}_n^* = \frac{1}{n}\sum_{i=1}^n h(\varepsilon_i - \Delta_{ni}^*) \quad \text{with } \Delta_{ni}^* = (-c_n) \vee \Delta_{ni} \wedge c_n.$$

Since $P\{\Delta_{ni} \ne \Delta_{ni}^* \text{ for some } i = 1, \dots, n\} \le P\{\max_{i=1,\dots,n} |\Delta_{ni}| > c_n\} \to 0$, we have

$$\hat{H}_n^* = \hat{H}_n + o_p(n^{-1/2}). \tag{6.1}$$

Also, by assumption (2.6),

$$\frac{1}{n}\sum_{i=1}^n \Delta_{ni}^* = \frac{1}{n}\sum_{i=1}^n \varepsilon_i + o_p(n^{-1/2}). \tag{6.2}$$

Introduce

$$\begin{aligned} D_{ni} &= h(\varepsilon_i - \Delta_{ni}^*) - h(\varepsilon_i) - \int (h(z - \Delta_{ni}^*) - h(z)) f(z)\, dz \\ &= h(\varepsilon_i - \Delta_{ni}^*) - h(\varepsilon_i) - \mu(\Delta_{ni}^*) + \mu(0). \end{aligned}$$

In view of (6.1) and (6.2), it suffices to show that

$$T_1 = n^{-1/2}\sum_{i=1}^n D_{ni} = o_p(1), \tag{6.3}$$

$$T_2 = n^{-1/2}\sum_{i=1}^n (\mu(\Delta_{ni}^*) - \mu(0) - \mu'(0)\Delta_{ni}^*) = o_p(1). \tag{6.4}$$

It follows from assumptions (2.2) and (2.4) that

$$|T_2| \le C_2 n^{-1/2}\sum_{i=1}^n |\Delta_{ni}^*|^{1+\alpha} = o_p(1).$$

To verify (6.3), it suffices to show that

$$E(T_1^2|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n E(D_{ni}^2|\mathbf{X}) + \frac{1}{n}\sum\sum_{i\ne j} E(D_{ni}D_{nj}|\mathbf{X}) = o_p(1). \tag{6.5}$$

By assumption (2.1),

$$\begin{aligned}
E(D_{ni}^2|\mathbf{X},\mathbf{Y}_{-i}) &= \int (h(z-\Delta_{ni}^*)-h(z)-\mu(\Delta_{ni}^*)+\mu(0))^2 f(z)\,dz \\
&\le \int (h(z-\Delta_{ni}^*)-h(z))^2 f(z)\,dz \\
&\le C_1|\Delta_{ni}^*|^{1+\alpha} \le C_1 c_n^{1+\alpha}.
\end{aligned}$$

Thus

$$\frac{1}{n}\sum_{i=1}^n E(D_{ni}^2|\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n E(E(D_{ni}^2|\mathbf{X},\mathbf{Y}_{-i})|\mathbf{X}) \le C_1 c_n^{1+\alpha} = o_p(1).$$

To deal with the cross-product terms, we introduce

$$\begin{aligned}
\Delta_{nij} &= (-c_n)\vee E(\Delta_{ni}|\mathbf{X},\mathbf{Y}_{-j})\wedge c_n, \\
D_{nij} &= h(\varepsilon_i-\Delta_{nij})-h(\varepsilon_i)-\mu(\Delta_{nij})+\mu(0).
\end{aligned}$$

The key is the identity

$$E(D_{ni}D_{nj}|\mathbf{X}) = E((D_{ni}-D_{nij})(D_{nj}-D_{nji})|\mathbf{X}), \quad i\ne j. \tag{6.6}$$

To see this, note that

$$E(D_{ni}|\mathbf{X},\mathbf{Y}_{-i}) = \int (h(z-\Delta_{ni}^*)-h(z)-\mu(\Delta_{ni}^*)+\mu(0))f(z)\,dz = 0.$$

Also, since D_{nji} does not depend on ε_i,

$$E(D_{ni}D_{nji}|\mathbf{X}) = E(E(D_{ni}D_{nji}|\mathbf{X},\mathbf{Y}_{-i})|\mathbf{X}) = E(D_{nji}E(D_{ni}|\mathbf{X},\mathbf{Y}_{-i})|\mathbf{X}) = 0, \quad i\ne j.$$

Similarly, one verifies $E(D_{nij}D_{nj}|\mathbf{X}) = E(D_{nij}D_{nji}|\mathbf{X}) = 0$. This proves (6.6). Thus we get

$$\begin{aligned}
\Big|\frac{1}{n}\sum\sum_{i\ne j} E(D_{ni}D_{nj}|\mathbf{X})\Big| &= \Big|\frac{1}{n}\sum\sum_{i\ne j} E((D_{ni}-D_{nij})(D_{nj}-D_{nji})|\mathbf{X})\Big| \\
&\le \frac{1}{n}\sum\sum_{i\ne j} E((D_{ni}-D_{nij})^2|\mathbf{X}) \\
&\le \frac{1}{n}\sum\sum_{i\ne j} E\Big(\int (h(z-\Delta_{nij})-h(z-\Delta_{ni}^*))^2 f(z)\,dz\Big|\mathbf{X}\Big).
\end{aligned}$$

Now use assumption (2.1) together with assumption (2.3) to bound the last term by

$$\frac{1}{n}\sum\sum_{i\ne j} C_1 E(|\Delta_{nij}-\Delta_{ni}^*|^{1+\alpha}|\mathbf{X}).$$

This is $o_p(1)$ by assumption (2.5) because $|\Delta_{nij}-\Delta_{ni}^*| \le |\hat r_{nij}-\hat r_{ni}|$ as the map $t\mapsto(-c)\vee t\wedge c$ is Lipschitz with Lipschitz constant 1 for each $c>0$. This completes the proof of (6.5) and hence of (6.3).

Proof of Theorem 2. We shall only sketch the proof and leave the details to the reader. The key are the following properties.

$$\max_{0\le l\le L}\max_{1\le i\le n}|q_{nil}| = O_p(1) \quad \text{and} \quad \max_{0\le l\le 2L}\max_{1\le i\le n}|p_{nil}(X_i)| = O_p(1). \tag{6.7}$$

They follow from the corresponding results for the usual polynomial smoothers in which the role of p_{nil} is played by p_{nl} defined by

$$p_{nl}(x) = \frac{1}{n}\sum_{j=1}^{n} w_{nl}(X_j - x), \quad x \in [0,1],$$

and the fact that $|p_{nil}(x) - p_{nl}(x)| \le \|w\|_\infty n^{-1}b_n^{-1}$ for all x, l and i. Indeed, one has $\sup_{0\le x\le 1}|p_{nl}(x) - E(p_{nl}(x))| = o_p(1)$ for all $l = 0,\dots,2L$ by a standard argument and in view of $b_n n^{1/3} \to \infty$, and one can show that the eigenvalues of the positive definite matrices $[E(p_{n,k+l}(x))]_{l,k=0,\dots,L}$, $0 \le x \le 1$, belong to a compact subset of $(0,\infty)$. It is now easy to check that (3.8) to (3.11) hold, implying (3.7) in view of $n^{1/2}b_n^{L+\beta} \to 0$. From (6.7) we also obtain

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} A_{nij}^2 = O_p(n^{-1}b_n^{-1}).$$

Thus (3.5) follows from $b_n n^{\alpha/(1+\alpha)} \to \infty$. Since the errors have finite second moment, we have $\max_{1\le i\le n}|\varepsilon_i| = o_p(n^{1/2})$. In view of this and (6.7), condition (3.4) follows from

$$\sup_{0\le x\le 1}\left|\frac{1}{n}\sum_{j=1}^{n} w_{nl}(X_j - x)\varepsilon_j\right| = o_p(1), \quad l = 0,\dots,L.$$

But these are verified by standard methods under the rate assumption $n^{1/3}b_n \to \infty$: First replace ε_j by $\varepsilon_{nj} = \varepsilon_j\mathbf{1}[|\varepsilon_j| < n^{1/2}] - E[\varepsilon\mathbf{1}[|\varepsilon| < n^{1/2}]]$ and then apply the Hoeffding inequality. Finally, (3.6) is obtained by direct calculations utilizing that $\sup_{b_n\le x\le 1-b_n}|p_{nl}(x) - g(x)\int t^l w(t)\,dt| = o_p(1)$.

Acknowledgment. We thank an Associate Editor and two referees for suggestions that improved the exposition considerably.

References

Akritas, M. G. and Van Keilegom, I. (2001). Non-parametric estimation of the residual distribution. *Scand. J. Statist.* 28, 549–567.

Bai, J. (1994). Weak convergence of the sequential empirical processes of residuals in ARMA models. *Ann. Statist.* 22, 2051–2061.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models.* Springer, New York.

Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* 23, 11–29.

Boldin, M. V. (1982). Estimation of the distribution of noise in an autoregression scheme. *Theory Probab. Appl.* 27, 866–871.

Boldin, M. V. (1983). Testing hypotheses in autoregression schemes by the Kolmogorov and ω^2 criteria. *Soviet Math. Dokl.* 28, 550–553.

Boldin, M. V. (1989). On testing hypotheses in the sliding average scheme by the Kolmogorov–Smirnov and ω^2 tests. *Theory Probab. Appl.* 34, 699–704.

Boldin, M. V. (1998). On residual empirical distribution functions in ARCH models with applications to testing and estimation. *Mitt. Math. Sem. Giessen* 235, 49–66.

Buckley, M. J., Eagleson, G. K. and Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika* 75, 189–199.

Carter, C. K. and Eagleson, G. K. (1992). A comparison of variance estimators in nonparametric regression. *J. Roy. Statist. Soc. Ser. B. (Methodological)* 54, 773–780.

Dette, H., Munk, A. and Wagner, T. (1998). Estimating the variance in nonparametric regression — what is a reasonable choice? *J. Roy. Statist. Soc. Ser. B. (Methodological)* 60, 751–764.

Dette, H., Munk, A. and Wagner, T. (1999). A review of variance estimators with extensions to multivariate nonparametric regression models. In: S. Ghosh, Ed., *Multivariate Analysis, Design of Experiments, and Survey Sampling*, Statistics: Textbooks and Monographs 159, Dekker, New York, 469–498.

Efromovich, S. and Samarov, A. (2000). Adaptive estimation of the integral of squared regression derivatives. *Scand. J. Statist.* 27, 335–351.

Eggermont, P. P. B. and LaRiccia, V. N. (1999). Best asymptotic normality of the kernel density entropy estimator for smooth densities. *IEEE Trans. Inform. Theory* 45, 1321–1326.

Ghoudi, K. and Rémillard, B. (1998). Empirical processes based on pseudo-observations. In: B. Szyszkowicz, Ed., *Asymptotic Methods in Probability and Statistics*, North-Holland, Amsterdam, 171–197.

Goldstein, L. and Messer, K. (1992). Optimal plug-in estimators for nonparametric functional estimation. *Ann. Statist.* 20, 1306–1328.

Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests.* Academic Press, New York.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika* 77, 521–528.

Hall, P. and Marron, J. S. (1990). On variance estimation in nonparametric regression. *Biometrika* 77, 415–419.

Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation. Asymptotic Theory.* Applications of Mathematics 16, Springer, New York.

Koshevnik, Yu. A. (1996). Semiparametric estimation of a symmetric error distribution from regression models. *Publ. Inst. Statist. Univ. Paris* 40, 77–91.

Koul, H. L. (1969). Asymptotic behavior of Wilcoxon type confidence regions in multiple linear regression. *Ann. Math. Statist.* 40, 1950–1979.

Koul, H. L. (1970). Some convergence theorems for ranks and weighted empirical cumulatives. *Ann. Math. Statist.* 41, 1768–1773.

Koul, H. L. (1987). Tests of goodness-of-fit in linear regression. *Colloq. Math. Soc. János Bolyai* 45, 279–315.

Koul, H. L. (1996). Asymptotics of some estimators and sequential residual empiricals in nonlinear time series. *Ann. Statist.* 24, 380–404.

Koul, H. L. (2002). *Weighted Empiricals and Linear Models.* Lecture Notes in Statistics 166, Springer, New York.

Koul, H. L. and Levental, S. (1989). Weak convergence of the residual empirical process in explosive autoregression. *Ann. Statist.* 17, 1784–1794.

Koul, H. L. and Ossiander, M. (1994). Weak convergence of randomly weighted dependent residual empiricals with applications to autoregression. *Ann. Statist.* 22, 540–562.

Koul, H. L. and Sen, P. K. (1991). Weak convergence of a weighted residual empirical process in autoregression. *Statist. Decisions* 9, 235–262.

Kreiss, J.-P. (1991). Estimation of the distribution function of noise in stationary processes. *Metrika* 38, 285–297.

Levit, B. Y. (1975). Conditional estimation of linear functionals. *Problems Inform. Transmission* 11, 39–54.

Loynes, R. M. (1980). The empirical distribution function of residuals from generalised regression. *Ann. Statist.* 8, 285–299.

Mammen, E. (1996). Empirical process of residuals for high-dimensional linear models. *Ann. Statist.* 24, 307–335.

Müller, U. U., Schick, A. and Wefelmeyer W. (2001). Estimating the error variance in nonparametric regression by a covariate-matched U-statistic. Technical Report, Department of Mathematical Sciences, Binghamton University. http://www.math.binghamton.edu/anton/preprint.html.

Schick, A. (1993). On efficient estimation in regression models. *Ann. Statist.* 21, 1486–1521. Correction and addendum: 23 (1995) 1862–1863.

Schick, A. and Wefelmeyer W. (2002). Estimating the innovation distribution in nonlinear autoregressive models. To appear in: *Ann. Inst. Statist. Math.*

Shorack, G. R. (1984). Empirical and rank processes of observations and residuals. *Canad. J. Statist.* 12, 319–332.

Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley Series in Probability and Mathematical Statistics, Wiley, New York.

Wefelmeyer, W. (1994). An efficient estimator for the expectation of a bounded function under the residual distribution of an autoregressive process. *Ann. Inst. Statist. Math.* 46, 309–315.

Ursula U. Müller
Fachbereich 3: Mathematik und Informatik
Universität Bremen
Postfach 330 440
28334 Bremen, Germany
`uschi@math.uni-bremen.de`
`http://www.math.uni-bremen.de/∼uschi/`

Anton Schick
Department of Mathematical Sciences
Binghamtom University
Binghamton, NY 13902-6000, USA
`anton@math.binghamton.edu`
`http://math.binghamton.edu/anton/index.html`

Wolfgang Wefelmeyer
Fachbereich 6 Mathematik
Universität Siegen
Walter-Flex-Str. 3
57068 Siegen, Germany
`wefelmeyer@mathematik.uni-siegen.de`
`http://www.math.uni-siegen.de/statistik/wefelmeyer.html`