The Annals of Statistics
1987, Vol. 15, No. 2, 610–625

ESTIMATION OF HETEROSCEDASTICITY IN REGRESSION ANALYSIS

BY HANS-GEORG MÜLLER AND ULRICH STADTMÜLLER

University of Marburg and University of Ulm

Consider the regression model $Y_i = g(t_i) + \varepsilon_i$, $1 \le i \le n$, with nonrandom design variables (t_i) and measurements (Y_i) for the unknown regression function $g(\cdot)$. We assume that the data are heteroscedastic, i.e., $E(\varepsilon_i^2) = \sigma_i^2 \not\equiv$ const. and investigate how to estimate σ_i^2. If $\sigma_i^2 = \sigma^2(t_i)$ with a smooth function $\sigma^2(\cdot)$, initial estimators $\tilde{\sigma}_i^2$ can be improved by kernel smoothers and the resulting class of estimators is shown to be uniformly consistent. These estimates can be used to improve the estimation of the regression function g itself in parametric and nonparametric models. Further applications are suggested.

1. Introduction. We consider the heteroscedastic regression model

$$Y_{i,n} = g(t_{i,n}) + \varepsilon_{i,n}, \qquad 1 \le i \le n, \tag{1.1}$$

where we omit indices n whenever feasible and assume that the design $\{t_1, \ldots, t_n\}$ is fixed. The r.v.'s (Y_i) are measurements of the unknown regression function g: $[0,1] \to \mathbb{R}$, contaminated with errors (ε_i), which are assumed to have expectation zero and are independently but not identically distributed. There are two main approaches to estimate the regression function g: the parametric approach, which assumes that g follows a parametric (linear or nonlinear) model and the nonparametric approach [Priestley and Chao (1972) and Reinsch (1967)], which only assumes that g is "smooth." The problem of heteroscedasticity, i.e., of nonconstant error variance, is generally recognized by applied statisticians dealing with regression methods who usually judge by looking at residual plots whether the model should be heteroscedastic or not and what the approximate behavior of the error variance is [Anscombe and Tukey (1963)]. However, it is well known that such subjective judging by eye can be quite misleading.

It is the aim of this paper to develop objective methods for estimating the variances of the errors under minimal assumptions on the regression model (1.1). As a first step, we consider initial variance estimates which are squared weighted sums of m (a fixed integer ≥ 2) observations neighboring a fixed point where the variance is to be estimated. Since these initial variance estimates are not consistent, we smooth them with a kernel estimate. The resulting final estimator of the local error variance is shown to be uniformly consistent.

In order to establish results for the proposed estimator, we consider the initial variance estimates as coming from a nonparametric regression model. In order to analyze this model, we have to investigate kernel estimators for m-dependent data. We derive a uniform consistency result for general linear estimators in the

Received July 1985; revised June 1986.

AMS 1980 *subject classifications.* Primary 62G05; secondary 62J02.

Key words and phrases. Local variance, kernel estimators, rates of uniform convergence, nonparametric regression, parametric regression, bandwidth variation in kernel estimators, convergence of weighted averages of m-dependent random variables, weighted least squares.

case of m-dependent data (Lemma 5.2), which is then specialized to the case of kernel estimators (Theorem 5.1). An application of the latter result then yields strong uniform rates for our local variance estimator (Theorem 3.1).

Among the possible applications, we discuss in more detail the estimation of optimal weights in weighted linear regression and the estimation of optimal local bandwidths in nonparametric kernel regression. For the first problem, we obtain as a solution an asymptotically efficient weighted least-squares estimate (Theorem 4.1), extending a result of Carroll (1982). For the second problem, we show that by substituting our local variance estimate and an estimate of a higher derivative of the regression function into the asymptotic formula for the optimal local bandwidth, the resulting variable bandwidth kernel estimators perform better w.r.t. integrated mean squared error (IMSE) than constant bandwidth kernel estimators supplied with the asymptotically optimal bandwidth (Theorem 4.2).

Estimation of the local variance has been considered in the context of simple linear regression with the aim of estimating optimal weights for weighted least squares by Fuller and Rao (1978) and by Carroll (1982). The estimate of Fuller and Rao, based on local squared residuals, was improved by Carroll who proposed to smooth neighboring squared residuals with the kernel method and derived rates of convergence for this estimate. Carroll's approach was shown to be superior in a simulation study by Matloff, Rose and Tai (1984). An estimate for a constant variance in a nonparametric regression model, which will be discussed in Section 2, was proposed by Rice (1984).

The organization of the paper is as follows:

In Section 2, the initial variance estimators are discussed, including some optimality considerations. Section 3 contains an investigation of the final kernel-smoothed local variance estimators. The results needed here on uniform convergence rates for weighted averages of m-dependent random variables, which are of interest in their own right, are compiled in Section 5. In Section 4, applications of the variance estimates to various statistical problems are proposed.

2. Initial local variance estimates. For the error variables we make the following:

ASSUMPTION A. The error variables (ε_i) are independent, $E(\varepsilon_i) = 0$, and there exist a constant $\gamma \in (0, 1]$ and a function $\sigma^2(t) \in \mathrm{Lip}_\gamma([0, 1])$ such that $E(\varepsilon_i^2) = \sigma^2(t_i)$, $1 \le i \le n$.

Here, the local variances are assumed to vary smoothly (i.e., to be Lipschitz continuous of order γ). We consider the local variance at t_{ν_0}, an interior point of $[0, 1]$, and write $\sigma_0^2 = \sigma^2(t_{\nu_0})$ and $g_0 = g(t_{\nu_0})$. The proposed class of initial local variance estimates at t_{ν_0} is given by

$$\tilde{\sigma}_0^2 = \left(\sum_{j=j_1}^{j_2} \omega_j Y_{j+\nu_0} \right)^2, \tag{2.1}$$

where $m \geq 2$ is a fixed integer and $j_1 = -[m/2]$, $j_2 = [m/2 - \frac{1}{4}]$. Here $[a]$ denotes the largest integer $\leq a$. In order to obtain an asymptotically unbiased estimate, we have to require that

$$\sum_{j=j_1}^{j_2} \omega_j = 0 \quad \text{and} \quad \sum_{j=j_1}^{j_2} \omega_j^2 = 1. \tag{2.2}$$

Furthermore, we assume that there exists a function $\mu_4(t) \in \mathrm{Lip}_\beta([0,1])$ for some $\beta \in (0,1]$ s.t.

$$E(\varepsilon_i^4) = \mu_4(t_i) < \infty, \quad \text{for } 1 \leq i \leq n,\ \mu_{4,0} := \mu_4(t_{\nu_0}). \tag{2.3}$$

We require the design (t_i) in the regression model (1.1) to be a regular sequence in the sense of Sacks and Ylvisaker (1970) generated by a design density f, i.e.,

$$\int_0^{t_i} f(x)\,dx = \frac{i-1}{n-1}, \qquad i = 1, \ldots, n, \tag{2.4}$$

where $f \in \mathrm{Lip}_1([0,1])$ is positive on $[0,1]$. Obviously, this implies that $\max_{1 \leq i \leq n}(t_i - t_{i-1}) = O(n^{-1})$.

LEMMA 2.1. *Assume that Assumption* A *and conditions* (2.2)–(2.4) *hold and that* $g \in \mathrm{Lip}_\alpha([t_{\nu_0} - z, t_{\nu_0} + z])$ *for some* $z > 0$. *Then*

$$\begin{aligned} &\text{(i)} \quad E(\tilde{\sigma}_0^2) = \sigma_0^2 + O(n^{-\min(2\alpha,\gamma)}), \\ &\text{(ii)} \quad \mathrm{Var}(\tilde{\sigma}_0^2) = (\mu_{4,0} - 3\sigma_0^4) \sum_{j=j_1}^{j_2} \omega_j^4 + 2\sigma_0^4 + O(n^{-\min(\alpha,\beta,\gamma)}). \end{aligned} \tag{2.5}$$

PROOF.

$$\begin{aligned} \text{(i)} \quad & E\left(\left(\sum_{j=j_1}^{j_2} \omega_j\left(\varepsilon_{j+\nu_0} + g\left(t_{j+\nu_0}\right)\right)\right)^2\right) \\ &= E\left(\left(\sum_{j=j_1}^{j_2} \omega_j \varepsilon_{j+\nu_0} + g_0 \sum_{j=j_1}^{j_2} \omega_j + O(n^{-\alpha})\right)^2\right) \\ &= \sigma_0^2 \sum_{j=j_1}^{j_2} \omega_j^2 + O(n^{-\gamma}) + O(n^{-2\alpha}), \quad \text{by Assumption A and (2.2).} \end{aligned}$$

$$\begin{aligned} \text{(ii)} \quad \mathrm{Var}(\tilde{\sigma}_0^2) &= E\left(\left(\sum_{j=j_1}^{j_2} \omega_j \varepsilon_{j+\nu_0} + O(n^{-\alpha})\right)^4\right) - \sigma_0^4 + O(n^{-\min(2\alpha,\gamma)}) \\ &= \mu_{4,0} \sum_{j=j_1}^{j_2} \omega_j^4 + 3 \sum_{i \neq j} \omega_i^2 \omega_j^2 \sigma_0^4 - \sigma_0^4 + O(n^{-\alpha}) \\ &\quad + O(n^{-\beta}) + O(n^{-\min(2\alpha,\gamma)}). \end{aligned}$$

By (2.2), we obtain (2.5)(ii). □

Observe that $\mu_{4,0} - 3\sigma_0^4 = \sigma_0^4\gamma_2$, where $\gamma_2 = (\mu_{4,0}/\sigma_0^4) - 3$ is the kurtosis [see Kendall and Stuart (1969)]. If the error variables (ε_i) are normally distributed, then $\gamma_2 = 0$, and hence in this case the variance of $\hat{\sigma}_0^2$ does not depend on the choice of weights, provided (2.2) holds.

In other cases it is possible to minimize the variance if one knows whether γ_2 is positive or negative. In case $m = 2$ the weights (ω_j) are completely determined by (2.2) to be $\omega_{j_1} = 1/\sqrt{2}$, $\omega_{j_2} = -1/\sqrt{2}$. Considering cases $m \geq 3$, we write $\omega = (\omega_{j_1}, \ldots, \omega_{j_2})$. Minimizing the variance of $\tilde{\sigma}_0^2$ corresponds to minimizing or maximizing (according to the sign of γ_2) $S(\omega) = \sum_{j=j_1}^{j_2} \omega_j^4$ under the side condition (2.2).

The following results on this variational problem can be obtained:

(i) If $m = 3$, we see that $S(\omega) = \frac{1}{2}$ for all weights satisfying (2.2).

(ii) If $m = 2\mu$, $\mu \geq 2$, $S(\omega)$ is minimized by the weights $\omega_j^* = (-1)^{e_j}/\sqrt{m}$, $j_1 \leq j \leq j_2$, where $e_j \in \{\pm 1\}$ and $\sum_{j=j_1}^{j_2} e_j = 0$; these weights yield $S(\omega^*) = 1/m$.

(iii) If $m = 2\mu + 1$, $\mu \geq 2$, $S(\omega)$ is minimized asymptotically (for large m) by the weights

$$\omega_j^* = \begin{cases} \pm\mu/(\mu(\mu+1)(2\mu+1))^{1/2}, & j_1 \leq j \leq j_1 + \mu, \\ \mp(\mu+1)/(\mu(\mu+1)(2\mu+1))^{1/2}, & j_1 + \mu < j \leq j_2, \end{cases}$$

or any permutation, yielding $S(\omega^*) = (\mu^3 + (\mu+1)^3)/m^2\mu(\mu+1)$.

(iv) If $m \geq 4$, $S(\omega)$ is maximized asymptotically by the weights

$$\omega_j^{**} = \begin{cases} \pm(m-1)/(m(m-1))^{1/2}, & j = j_1, \\ \mp 1/(m(m-1))^{1/2}, & j_1 < j \leq j_2, \end{cases}$$

or any permutation, yielding $S(\omega^{**}) = ((m-1)^3 + 1)/(m^2(m-1))$.

(v) Examples are for

$m=3$: $\omega = \frac{1}{\sqrt{6}}(1, -2, 1)$, $S(\omega) = \frac{1}{2}$,

$m=4$: $\omega^* = \frac{1}{2}(1, -1, 1, -1)$, $S(\omega^*) = \frac{1}{4}$,

$\omega^{**} = \frac{1}{2\sqrt{3}}(-1, -1, 3, -1)$, $S(\omega^{**}) = \frac{7}{12}$,

$m=5$: $\omega^* = \frac{1}{\sqrt{30}}(2, -3, 2, -3, 2)$, $S(\omega^*) = \frac{7}{30}$,

$\omega^{**} = \frac{1}{2\sqrt{5}}(-1, -1, 4, -1, -1)$, $S(\omega^{**}) = \frac{13}{20}$.

In order to estimate a constant global variance in a nonparametric regression model, Rice (1984) proposed the special cases for $m = 2$ and $m = 3$ of the above estimators. A related idea was already discussed by Breiman and Meisel (1976). Another approach that was, e.g., considered by Silverman (1985), is to estimate a global or local variance by taking a moving average of squared ordinary (or

deleted) residuals. An analysis of the latter proposal shows that (2.2) is approximately satisfied, but the remainder terms in (2.5) get worse. This reflects the empirical finding that this method entails a relatively large bias as compared to our approach.

3. Kernel smoothing of initial local variance estimates. Within the framework of Assumption A and condition (2.4), which we assume in the following, we proposed initial variance estimates in Section 2 that, however, are not consistent according to Lemma 2.1. In order to overcome this difficulty and to obtain consistent estimators, it is a natural approach to smooth neighboring values of $\tilde{\sigma}_i^2$ exploiting the smoothness properties of $\sigma^2(\cdot)$.

For this purpose, we view the initial estimates $\tilde{\sigma}_i^2$ as measurements coming from the following regression model:

$$\tilde{\sigma}_{i,n}^2 = \sigma^2(t_{i,n}) + \tilde{\varepsilon}_{i,n}, \qquad 1 \le i \le n. \tag{3.1}$$

In order to analyze this model, we assume that the original error variables in Assumption A [see (1.1)] satisfy, in addition,

$$E\left(|\varepsilon_i|^{2s}\right) \le M < \infty, \quad \text{for } 1 \le i \le n \text{ and some } s > 1. \tag{3.2}$$

Now it is easy to see (using Lemma 2.1) that the error variables $\tilde{\varepsilon}_{i,n}$ satisfy the following Assumption B with $\rho = \min(2\alpha, \gamma)$ (for α see Lemma 2.1, for γ see Assumption A).

ASSUMPTION B. The error variables $\varepsilon_i = (\varepsilon_{i,n})_{i=1,n}$ form a triangular array of rowwise $(m-1)$-dependent r.v.'s [$m \ge 1$, see, e.g., Billingsley (1979): ε_i, ε_j are independent for $|i-j| > m-1$] and satisfy $E(|\varepsilon_{i,n}|^s) \le M < \infty$ for some $s > 2$ and $\max_{1 \le i \le n} |E(\varepsilon_{i,n})| = O(n^{-\rho})$ for some $\rho > 0$.

For carrying out the smoothing procedure, we apply the kernel estimators [see Gasser and Müller (1984)]

$$\hat{\sigma}^2(t) = \frac{1}{b} \sum_{j=1}^{n} \int_{s_{j-1}}^{s_j} K\left(\frac{t-u}{b}\right) du\, \tilde{\sigma}_j^2, \tag{3.3}$$

$$\text{with } s_0 = 0,\ s_n = 1,\ s_j = \frac{t_j + t_{j+1}}{2},\ 1 \le j \le n-1.$$

The sequence of bandwidths $b = b(n)$ has to satisfy

$$b \to 0, \qquad nb \to \infty, \quad \text{as } n \to \infty, \tag{3.4}$$

and K denotes the kernel function. Given an integer $k \ge 0$ and a $\zeta \in [0,1]$, we require that $K \in \mathcal{M}_{k+\zeta}$, where

$$\mathcal{M}_{k+\zeta} = \left\{ h \in \mathrm{Lip}_1([-1,1]) : \mathrm{support}(h) = [-1,1], \right.$$
$$\left. \int_{-1}^{1} h(t) t^j\, dt = \begin{cases} 0, & 0 < j <]k+\zeta[, \\ 1, & j = 0, \\ B_{k+\zeta} \ne 0, & j =]k+\zeta[, \end{cases} \right\} \tag{3.5}$$

where $]k+\zeta[$ denotes the smallest integer $\ge k+\zeta$.

For the following main result we introduce the smoothness classes

$$S_{k,\zeta}([0,1]) := \left\{ g \in \mathscr{C}^k([0,1]) \quad \text{and} \quad g^{(k)} \in \mathrm{Lip}_\zeta([0,1]) \right\},$$

where $\zeta \in [0,1]$, $k \geq 0$, is an integer and $\mathrm{Lip}_0([0,1]) := \mathscr{C}([0,1])$. A straightforward application of Theorem 5.1B and Lemma 2.1 now yields the following convergence properties of the estimators (3.3), observing that the assumptions made imply Assumption B with $\rho \geq \frac{1}{2}$ for model (3.1).

THEOREM 3.1. *Assume that* $\sigma^2(\cdot) \in S_{k,\zeta}([0,1])$ *for some integer* $k \geq 0$ *and some* $0 \leq \zeta \leq 1$, *such that* $k + \zeta \geq \frac{1}{2}$, *and that* $g \in \mathrm{Lip}_\alpha([0,1])$ *for some* $\alpha \geq \frac{1}{4}$. *Furthermore, suppose that* $K \in \mathscr{M}_{k+\zeta} \cap \mathrm{Lip}_1(\mathbb{R})$ *and that for some* $s > 4 + 2(k+\zeta)^{-1}$, $E(|\varepsilon_i|^{2s}) \leq M < \infty$, $1 \leq i \leq n$, *in the model* (1.1). *If the bandwidth in* (3.3) *is chosen according to* $b \sim [\log n/n]^{1/(2(k+\zeta)+1)}$, *we obtain*

$$\sup_{t \in I} |\hat{\sigma}^2(t) - \sigma^2(t)| = O\left(\left(\frac{\log n}{n} \right)^{(k+\zeta)/(2(k+\zeta)+1)} \right) \quad a.s., \tag{3.6}$$

for any compact subinterval $I \subset (0,1)$.

REMARKS. (i) Using modified boundary kernels [compare Gasser and Müller (1984)] or assuming that data are available on $[-z, 1+z]$, for some $z > 0$, the uniform consistency result extends to $[0,1]$.

(ii) In case we replace in (3.6) the a.s. convergence by convergence in probability, the moment requirements can be weakened, assuming only $s > 2 + (k+\zeta)^{-1}$; according to Theorem 5.1A they depend on the rate to be attained.

(iii) Carroll (1982) assumes for his related Theorem 5.1 that $k = 1$ and $\zeta = 0$ and obtains in the linear regression model $n^{-1/4}$ as the rate of convergence. Under the same smoothness assumptions, we obtain in the nonparametric model a rate of $(\log n/n)^{1/3}$. If $\sigma^2(\cdot)$ is sufficiently smooth, the rate can be improved further.

Our rates correspond to the optimal ones given for a different model by Stone (1982).

(iv) The bandwidth for $\hat{\sigma}(\cdot)$ can be chosen by cross validation suitably modified for $(m-1)$-dependent data. Preliminary simulation results are encouraging.

(v) The result remains unchanged if we have multiple measurements at each point t_i, that is,

$$Y_{ij} = g(t_i) + \varepsilon_{ij}, \qquad 1 \leq i \leq n, 1 \leq j \leq m_i > 1,$$

and

$$\tilde{\sigma}_i^2 = \frac{1}{m_i - 1} \sum_{j=1}^{m_i} \left(Y_{ij} - \bar{Y}_i \right)^2.$$

(vi) Our approach generalizes easily to the multivariate regression problem g: $\mathbb{R}^d \to \mathbb{R}$, for some $d > 1$. Here we base the initial estimates (2.1) on the m nearest neighbors of a given point $t \in \mathbb{R}^d$. Then we use multivariate kernel

estimators, as described in Müller (1983), and obtain again a uniform convergence result if $2k + d \geq kd$ and $K \in \mathcal{M}_{k+\zeta}$ (obvious generalization): the rate of convergence is $(\log n/n)^{(k+\zeta)/(2(k+\zeta)+d)}$.

4. Applications of local variance estimation.

4.1. *Possible applications of local variances.* We suggest here several applications of local variance estimation and explore two proposals further.

(i) *Controlling the error of measurements over time.* Considering, e.g., longitudinal regression studies, there is a genuine interest to observe the behavior of the error variance over time. The main objective is to draw conclusions on the underlying mechanism generating the data, but another aim is to control the accuracy of the measurements that may decrease during the course of time because the measurement devices lose precision or the measurement team works less carefully.

(ii) *Choice of optimal design in nonparametric regression.* If, in a nonparametric setup, a constant bandwidth kernel estimator is used to estimate g, we can find for a follow-up study the optimal design density f^*, given by $f^*(t) = \sigma(t)/\int_0^t \sigma(x)\,dx$ [see Müller (1984)] and a natural estimate for f^* would be $\hat{f}^*(t) = \hat{\sigma}(t)/\int_0^t \hat{\sigma}(x)\,dx$, employing (3.3).

(iii) *Improvement of the estimation of the regression function itself.*

(a) A first possibility, applicable to parametric and nonparametric regression, is the transformation of the data towards homoscedasticity by $Y_i \to Y_i/\hat{\sigma}(t_i)$, fitting then a parametric or nonparametric regression curve $\tilde{g}$ (e.g., by ordinary least squares in the parametric case). Afterwards we retransform and obtain the estimate $\hat{g}(t_i) = \tilde{g}(t_i)\hat{\sigma}(t_i)$.

(b) In the case of parametric regression under Assumption A, we know by the Gauss–Markov theorem [see, e.g., Rao (1973)] that an optimal unbiased linear estimator for the parameters is given by the weighted least-squares method. The optimal weights are $1/\sigma^2(t_i)$ and can be estimated by $1/\hat{\sigma}^2(t_i)$ [$\hat{\sigma}(t_i)$ defined in (3.3)]. We will show in Section 4.2 that weighted least-squares estimators with weights $1/\hat{\sigma}(t_i)$ are asymptotically efficient. A related procedure for linear regression models was investigated by Carroll (1982), Matloff, Rose and Tai (1984) and Rose (1978).

(c) In the case of nonparametric regression one could use $\hat{\sigma}(\cdot)$ to do weighted kernel estimation. This is of interest if the local variance changes considerably on intervals of length b, since the smoothing window of the kernel estimate is $[t - b, t + b]$ if the curve is to be estimated at t. But asymptotically (for a large number of observations) this method is equivalent to the ordinary kernel estimate since then $b \to 0$ and $\sigma^2(\cdot)$ is approximately constant on $[t - b, t + b]$. Another improvement of kernel estimation is to choose estimated local optimal bandwidths instead of global (constant) bandwidths by applying $\hat{\sigma}^2(\cdot)$. Details will be given in Section 4.3, where we shall show that the integrated mean squared error (IMSE) can indeed be reduced by this local bandwidth choice.

4.2. *Data adaptive efficient parametric regression.* In this section we deal with the following parametric model (4.1), which is a special case of model (1.1):

$$Y_i = a(t_i)^T\theta + \varepsilon_i, \qquad 1 \le i \le n, 0 \le t_1 \le \cdots \le t_n \le 1, \tag{4.1}$$

where $a(t)^T = (a_1(t), \ldots, a_p(t))$ is a vector of $p < n$ linearly independent given functions on $[0,1]$; then the matrix $A = (a_i(t_k))_{i=1,k=1}^{p,n}$ has full rank p for sufficiently large n, where the points of measurements (t_k) follow (2.4). $\theta \in \mathbb{R}^p$ is a vector of unknown parameters that have to be estimated. Furthermore, we assume that (4.1) is heteroscedastic, that is, Assumption A is in force and that we can neglect boundary effects. It is well known that the best linear unbiased estimator in the sense of the Gauss–Markov theory is the weighted least-squares estimator $\tilde{\theta}$, the minimizer of

$$\sum_{j=1}^{n} \sigma^{-2}(t_j)\left(Y_j - a(t_j)^T\theta\right)^2 = \min! \quad \text{w.r.t. } \theta \in \mathbb{R}^p. \tag{4.2}$$

Usually, $\sigma^2(\cdot)$ is unknown and our suggestion is to estimate θ by $\hat{\theta}$, which is the minimizer of (4.2), with $\sigma^{-2}(\cdot)$ replaced by $\hat{\sigma}^{-2}(\cdot)$, the variance estimate (3.3). The following result shows that this adaptive procedure is asymptotically efficient and that we lose nothing when replacing $\sigma^2(\cdot)$ by $\hat{\sigma}^2(\cdot)$ in (4.2) in terms of the asymptotic distribution.

THEOREM 4.1. *Assume that for the heteroscedastic regression model* (4.1) *under Assumption* A *and* (2.4) *we have* $E(|\varepsilon_i|^{2s}) \le M < \infty$ *for some* $s > 2$, $1 \le i \le n$, $\sigma^2(\cdot) \in \mathrm{Lip}_1([0,1])$ *and* $\inf_{[0,1]}\sigma(t) > 0$. *Furthermore, assume that* $a_i(\cdot) \in \mathrm{Lip}_\alpha[0,1]$ *for* $1 \le i \le n$, *for some* $\alpha \ge \frac{1}{2}$. *If* $\hat{\theta}$ *and* $\tilde{\theta}$ *are defined as above and if we use for the kernel smoother* (3.3) *defining* $\hat{\sigma}^2(\cdot)$ *bandwidths* b *satisfying* $b \to 0$, $nb^2 \to \infty$ *and* $\liminf_{n\to\infty}(1/n^2)(nb/\log n)^{s-\eta} > 0$ *as* $n \to \infty$ *for some* $\eta \in (0, s-2)$, *then we have*

$$n^{1/2}(\hat{\theta} - \tilde{\theta}) \to_p 0, \quad \text{as } n \to \infty.$$

Therefore, $\hat{\theta}$ *is asymptotically efficient since* $n^{1/2}(\tilde{\theta} - \theta)$ *is asymptotically normally distributed.*

REMARKS. (i) Under the assumptions of Theorem 4.1, a weaker version of Theorem 5.1A gives $\sup_{t\in[0,1]}|\hat{\sigma}^2(t) - \sigma^2(t)| = o_p(1)$ as $n \to \infty$ according to Remark (ii) after Theorem 5.1.

(ii) In the case of simple linear regression, a similar result was obtained by Carroll (1982), Theorem 1, where the (t_i) are assumed to be random.

(iii) A multivariate version is possible [compare Remark (vi) after Theorem 3.1].

PROOF. Subtracting the normal equations for the weighted least-squares estimators $\hat{\theta}$ and $\tilde{\theta}$ one from another, we obtain [writing $a_{j\lambda} = a_\lambda(t_j)$, $\sigma_j = \sigma(t_j)$

and $\hat{\sigma}_j = \hat{\sigma}(t_j)$]

$$\sum_{\mu=1}^{p} n^{1/2}\left(\hat{\theta}_\mu - \tilde{\theta}_\mu\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{a_{i\lambda}a_{i\mu}}{\sigma_i^2}\left(1 + \left(\frac{\sigma_i^2}{\hat{\sigma}_i^2} - 1\right)\right)\right)$$

$$= \sum_{\mu=1}^{p} n^{1/2}\left(\theta_\mu - \tilde{\theta}_\mu\right)\left(\frac{1}{n}\sum_{i=1}^{n}\frac{a_{i\lambda}a_{i\mu}}{\hat{\sigma}_i^2}\left(\frac{\hat{\sigma}_i^2}{\sigma_i^2} - 1\right)\right)$$

$$+ n^{1/2}\sum_{i=1}^{n}\frac{a_{i\lambda}}{\sigma_i^2\hat{\sigma}_i^2}\left(\hat{\sigma}_i^2 - \sigma_i^2\right)\varepsilon_i, \qquad 1 \le \lambda \le p.$$

Writing

$$Q = \frac{1}{\sqrt{n}}\left(\frac{a(t_1)}{\sigma_1} \cdots \frac{a(t_n)}{\sigma_n}\right),$$

we find, using Riemann sums for integrals, that

$$Q^TQ \to \tilde{Q} = \left(\int_0^1 a_i(t)a_k(t)\frac{f(t)}{\sigma^2(t)}\,dt\right), \qquad 1 \le i, k \le p, \text{ as } n \to \infty.$$

$\tilde{Q}$ is positive definite, hence regular and thus $\tilde{Q}^{-1}$ has bounded elements. Observing now that

$$n^{1/2}\left(\theta_\mu - \tilde{\theta}_\mu\right) = O_p(1), \qquad \max_{1\le i\le n}|\hat{\sigma}_i^2 - \sigma_i^2| = o_p(1) \quad \text{[see Remark (i) above]}$$

and that the equation above can be written

$$Q^TQn^{1/2}(\hat{\theta} - \tilde{\theta})\left(1 + o_p(1)\right) = n^{-1/2}\left(\sum_{i=1}^{n}\frac{a_{i\lambda}}{\sigma_i^2\hat{\sigma}_i^2}\left(\sigma_i^2 - \hat{\sigma}_i^2\right)\varepsilon_i\right)_\lambda + o_p(1),$$

it suffices to show that the right-hand term in the last equation tends to zero in probability in order to prove Theorem 4.1. This can be shown by tedious calculations, demonstrating that the first two moments tend to zero, where we use

$$\hat{\sigma}_i^2 - \sigma_i^2 = \sum_j \frac{1}{b}\int_{s_{j-1}}^{s_j} K\left(\frac{t_i - u}{b}\right)du\left(\tilde{\sigma}^2(t_j) - \sigma_j^2\right) + o(1), \quad \text{uniformly in } i,$$

the compact support of K, the dependence structure of $(\tilde{\sigma}^2(t_j))$ and (ε_i): $((\tilde{\sigma}(t_j))$ are $(m-1)$-dependent and $\tilde{\sigma}(t_j)$ and ε_i are independent if $|j - i| > m/2$, the uniform boundedness of the $a_{i\lambda}$ and the fact that $E(\tilde{\sigma}^2(t_j) - \sigma_j^2) = O(1/n)$ by Lemma 2.1. □

4.3. *Improvement on nonparametric regression.* The idea here is to vary the bandwidth in nonparametric kernel regression locally and to smooth more at points where high residual variances occur. We show that such a procedure, which is a generalization of local bandwidth variation as proposed by Müller and Stadtmüller (1987), reduces the integrated mean squared error (IMSE) of kernel estimates.

Given a function b_t: $[0,1] \to \mathbb{R}_+$, we define variable bandwidth kernel estimators in the model (1.1) by

$$\hat{g}(t, b_t) = \frac{1}{b_t} \sum_{j=1}^{n} \int_{s_{j-1}}^{s_j} K\left(\frac{t-u}{b_t}\right) du\, Y_j. \tag{4.3}$$

If $b_t \equiv b \equiv$ const., i.e., if $\hat{g}$ is a global bandwidth estimator, we refer to it as $\hat{G}(\cdot, b)$.

Minimizing the asymptotic mean squared error (MSE) of (4.3) w.r.t. the bandwidth b yields under suitable assumptions ($g \in S_{k,0}$ and using a kernel $K \in \mathcal{M}_k$) the locally optimal bandwidth

$$b_t^* = \left(\frac{1}{n}\frac{1}{2k}\frac{V}{\tilde{B}_k^2}\frac{\sigma^2(t)}{f(t)g^{(k)}(t)^2}\right)^{1/(2k+1)}, \quad \text{provided } g^{(k)}(t) \neq 0, \tag{4.4}$$

where $V = \int K(x)^2\,dx$ and $\tilde{B}_k = (-1)^k \int K(x)x^k\,dx/k!$. For the globally optimal bandwidth b^* of G w.r.t. IMSE see Gasser and Müller (1984). From (4.4), local bandwidth variation would be desirable w.r.t.

(a) local variance of the error (heteroscedasticity) $\sigma^2(\cdot)$;
(b) local nonequidistancy of the design $f(\cdot)$;
(c) local curvature of the true function $g^{(k)}(\cdot)$.

A complete procedure for (c) was worked out in Müller and Stadtmüller (1987), using pilot estimators of $g^{(k)}(t)$. The simulation study provided there demonstrates for various examples the superiority of local bandwidth choice over global bandwidth choice already for small samples. Silverman (1984) showed that smoothing splines adapt locally like (b), but not with the correct exponent. In the general situation one would like to adapt w.r.t. (a)–(c) jointly. This can be done by replacing in (4.4) σ^2, f and $g^{(k)}$ by consistent estimates, obtaining a consistent estimate of b_t. From a theoretical point of view such a procedure might be quite noisy [compare Hall and Marron (1987)] but our simulation results are encouraging.

If we use a density estimator $\hat{f}$ for f (e.g., kernel estimator) if f is unknown, $\hat{\sigma}$ for σ (3.3) and $\hat{G}^{(k)}$ for $g^{(k)}$, we obtain under mild conditions that these estimators are weakly consistent for any $t \in [0,1]$ (neglecting or modifying for boundary effects). Such conditions are given by Parzen (1962) or Rosenblatt (1956) for $\hat{f}$, by Gasser and Müller (1984) for $\hat{G}^{(k)}$ and by Theorem 5.1A, observing Remark (ii) after Theorem 3.1, for $\hat{\sigma}$. Writing $\tilde{b}_t$ for the resulting estimate of b_t^* and setting

$$b_t^0 = \inf\{b_t^*, vn^{-1/(2k+1)}\}$$

and

$$\hat{b}_t = \inf\{\tilde{b}_t, vn^{-1/(2k+1)}\}, \quad \text{for some large } v > 0,$$

we arrive at

$$\hat{b}_t / b_t^0 \to_p 1, \quad \text{as } n \to \infty. \tag{4.5}$$

Applying the results of Müller and Stadtmüller (1987) and the uniform integrability lemma of Stadtmüller (1986b), we can show

THEOREM 4.2. *Under the assumptions above and* $\int_0^1 g^{(k)}(t)^2\,dt > 0$ *we have*

$$\limsup_{n\to\infty}\left(\mathrm{IMSE}\left(\hat{g}\left(t,\hat{b}_t\right)\right)/\mathrm{IMSE}\left(\hat{G}\left(t,b^*\right)\right)\right) \le 1, \tag{4.6}$$

and if $g^{(k)}(t) \neq 0$ *on* $[0,1]$, *the limit of the l.h.s. of* (4.6) *exists and is equal to*

$$\int_0^1 \left(\sigma^2(t)/f(t)\right)^{2k/(2k+1)}\left(g^{(k)}(t)\right)^{2/(2k+1)}dt \Bigg/ \left(\int_0^1 \sigma^2(t)/f(t)\,dt\right)^{2k/(2k+1)} \times \left(\int_0^1 g^{(k)}(t)^2\,dt\right)^{1/(2k+1)}, \tag{4.7}$$

which is at most 1 *by Hölder's inequality.*

Theorem 4.2 shows that a local bandwidth estimator with consistently estimated local bandwidths behaves never worse than a constant bandwidth estimator using the optimal bandwidth b w.r.t. IMSE. Moreover, the result shows that in case of equidistant data, i.e., $f \equiv 1$, local bandwidth variation is especially rewarding if $\sigma(\cdot)^{2k/(2k+1)}$ and $g^{(k)}(\cdot)^{2/(2k+1)}$ are close to be orthogonal in L^2, that is, e.g., if $\sigma(\cdot)$ is approximately antisymmetrical and $g^{(k)}(\cdot)$ is approximately symmetrical around 0.5 or vice versa.

5. Rates of uniform convergence. In this section we prove a rather general result on strong uniform convergence of weighted averages in Lemma 5.2, which is applied to kernel estimates in Theorem 5.1. We assume here that the error variables $(\varepsilon_i) = (\varepsilon_{i,n})$ of model (1.1) follow Assumption B (see Section 3).

First we give an exponential inequality for bounded *m-dependent* r.v.'s.

LEMMA 5.1. *Assume that the* $(\varepsilon_{i,n})$ *satisfy Assumption* B *but* $E(\varepsilon_{i,n}) \equiv 0$. *Furthermore, suppose that* $|\varepsilon_{i,n}| \le M < \infty$ *and* $E(\varepsilon_{i,n}^2) \le R_{i,n}^2$, $1 \le i \le n$. *Then we have for* $S_n = \sum_{i=1}^n \varepsilon_{i,n}$ *and all* $x \in [0, 2/M]$

$$E\left(\exp(xS_n)\right) \le \exp\left(\tfrac{3}{2}(m+1)x^2 \sum_{j=1}^n R_{j,n}^2\right). \tag{5.1}$$

PROOF. If $m = 0$, this inequality is well known [see, e.g., Lamperti (1966), Chapter II, Section 11, Lemma 1]. If $m > 0$ we separate S_n into $(m+1)$ partial sums $S_n^{(1)}, S_n^{(2)}, \ldots, S_n^{(m+1)}$ with independent random variables in each sum and use Hölder's inequality to obtain

$$E\left(\prod_{j=1}^{m+1}\exp\left(xS_n^{(j)}\right)\right) \le \prod_{j=1}^{m+1}\left(E\left(\exp\left(x(m+1)S_n^{(j)}\right)\right)^{1/(m+1)}\right)$$

and then apply (5.1) for $m = 0$ to the r.h.s. of the last inequality. □

Now we consider weighted averages

$$g_n(t) = \sum_{i=1}^{n} W_{i,n}(t) Y_i \tag{5.2}$$

as estimators of the regression function g in the model (1.1). The kernel estimate (3.3) and other curve estimates like smoothing splines are a special case. We derive rates of strong uniform convergence for the estimator (5.2) under Assumption B Cheng and Ling (1981) have derived related results under Assumption A for kernel estimators. A bound for the stochastic deviation is given in the following lemma, where indices n have been omitted whenever appropriate.

LEMMA 5.2. *Assume that the regression model* (1.1) *follows Assumption* B. *Furthermore, assume that the weight functions* $W_i(t)$ *satisfy for some* $0 < \delta \leq 1$ *and* $L_\delta > 0$

$$\sup_{1 \leq i \leq n} |W_i(t_1) - W_i(t_2)| \leq L_\delta |t_1 - t_2|^\delta, \quad \text{for all } t_1, t_2 \in [0,1] \tag{5.3}$$

and for some $c > 0$

$$\max_{1 \leq i \leq n} |W_i(t)| \geq cn^{-1}, \quad \text{uniformly for } t \in [0,1]. \tag{5.4}$$

Finally, suppose that there is a sequence $\alpha_n \downarrow 0$, *and constants* $\eta \in (0, s-2)$ *and* $K > \frac{1}{2}$ *s.t. for all* $t \in [0,1]$

$$n^{2/(s-\eta)} \max_{1 \leq i \leq n} |W_i(t)| \log n \leq \alpha_n / K \tag{5.5}$$

and

$$\left(\sum_{i=1}^{n} W_i^2(t) \log n \right)^{1/2} \leq \alpha_n / K. \tag{5.6}$$

Then $\sup_{t \in [0,1]} |g_n(t) - E(g_n(t))| = O(\alpha_n)$ *a.s.*

REMARK. Conditions (5.5) and (5.6) relate the variance of a weighted sum of r.v.'s with the maximum weight; conditions of this type are common for limit theorems of weighted averages.

PROOF. Defining $\mu = 3/\delta$, $r = s - \eta$ and $I = [0,1]$, we consider a sequence of $n^{-\mu}$-neighborhoods U_n covering I. Choosing proper middle points τ_n for U_n we need $O(n^\mu)$ sets U_n. Let $U_n(\tau_n(t))$ be s.t. $t \in U_n(\tau_n(t))$. Using (5.3)–(5.5), we find (with some constant K')

$$\begin{aligned} & n^{2+1/r} \sup_{t \in I} \max_{1 \leq i \leq n} \sup_{u \in U_n(\tau_n(t))} |W_i(t) - W_i(u)| \\ & \leq L_\delta n^{-1+1/r} \leq K' \max_{1 \leq i \leq n} |W_i(t)| n^{1/r} \leq \alpha_n / K, \end{aligned} \tag{5.7}$$

for n sufficiently large.

Defining $\bar{\varepsilon}_i = \varepsilon_i \chi(|\varepsilon_i| \le (in)^{1/r})$, $h(t) = \Sigma W_i(t)\varepsilon_i$, $\bar{h}(t) = \Sigma W_i(t)\bar{\varepsilon}_i$, we proceed as follows (where $\|\cdot\|_\infty := \sup_I |\cdot|$):

$$\begin{aligned}\|g_n(\cdot) - E(g_n(\cdot))\|_\infty \le{}& \|h(\cdot) - \bar{h}(\cdot)\|_\infty + \|\bar{h}(\cdot) - \bar{h}(\tau_n(\cdot))\|_\infty \\ &+ \|\bar{h}(\tau_n(\cdot)) - E(\bar{h}(\tau_n(\cdot)))\|_\infty \\ &+ \|E(\bar{h}(\tau_n(\cdot))) - E(\bar{h}(\cdot))\|_\infty \\ &+ \|E(\bar{h}(\cdot)) - E(h(\cdot))\|_\infty.\end{aligned}$$

Using $P(|\varepsilon_i| > (in)^{1/r}) \le E(|\varepsilon_i|^s)/(in)^{s/r}$ and the Borel–Cantelli lemma we show that for almost all ω of the underlying probability space Ω, there exists N_ω s.t. for $n > N_\omega$ we have $|\varepsilon_{i,n}(\omega)| \le (in)^{1/r} \le n^{2/r}$, $1 \le i \le n$. It follows that the first and the last term in the inequality above can be bounded by

$$O\left(\sup_{t\in I} \max_{1\le i\le n} |W_i(t)| \sum_{p=1}^{n} \sum_{i=1}^{p} |\bar{\varepsilon}_{i,p} - \varepsilon_{i,p}|\right),$$

resp., by

$$O\left(n^{2/r} \sup_{t\in I} \max_{1\le i\le n} |W_i(t)|\right),$$

and both terms by $O(\alpha_n)$ almost surely resp. strictly according to (5.5), bearing in mind that the $\varepsilon_i = \varepsilon_{i,n}$ form a triangular array.

For the second and fourth terms we apply (5.7). For the third term we define

$$\beta_n(t) := \alpha_n^{-2} \max_{1\le i\le n} |W_i(t)| n^{2/r} (\log n)^2,\ \eta_n(t) = \alpha_n \beta_n(t)$$

and apply Lemma 5.1 to the random variables $\beta_n(t)W_i(t)(\bar{\varepsilon}_i - E(\bar{\varepsilon}_i))$, choosing $x = (\beta_n(t)n^{2/r}\max_{1\le i\le n}|W_i(t)|)^{-1/2}$.

Observing $P(S_n > a) \le e^{-ax}E(e^{xS_n})$, we obtain for any constant $T > 0$

$$\begin{aligned}&P\left(\beta_n(t)\left(\sum W_i(t)(\bar{\varepsilon}_i - E(\bar{\varepsilon}_i))\right) > T\eta_n(t)\right) \\ &\quad\le \exp\left\{c_1 \frac{\beta_n(t)\Sigma W_i^2(t)}{n^{2/r}\max_{1\le i\le n}|W_i(t)|} - \frac{c_1 T\eta_n(t)}{\{\beta_n(t)n^{2/r}\max_{1\le i\le n}|W_i(t)|\}^{1/2}}\right\} \\ &\quad\le n^{c_1K^{-2}-c_2T},\end{aligned}$$

with suitable constants $c_1, c_2 > 0$, where (5.6) is needed to establish the last inequality. By symmetry, we conclude that

$$\sum_{n=1}^{\infty} P\left(\|\bar{h}(\tau_n(\cdot)) - E\bar{h}(\tau_n(\cdot))\|_\infty > T\alpha_n\right) \le C_T \sum_{n=1}^{\infty} n^{\mu} n^{c_1K^{-2}-c_2T} < \infty,$$

for T sufficiently large. The assertion follows from the Borel–Cantelli lemma. □

REMARK. In case the (ε_i) form a linear scheme, the term $n^{2/(s-\eta)}$ in (5.5) can be replaced by $n^{1/(s-\eta)}$, which implies that it suffices for Lemma 5.2 that the moment requirement in Assumption B is satisfied for $s/2$. The same holds if we consider bounds in probability.

Lemma 5.2 gives a result for the stochastic part of the deviation of weighted averages (5.2). For the bias, the deterministic approximation properties of (5.2) have to be investigated. This is done for the special case of kernel estimates for $g(t)$ in the model (1.1), i.e.,

$$\hat{g}_n(t) = \frac{1}{b}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K\left(\frac{t-u}{b}\right) du\, Y_i, \tag{5.8}$$

where we assume that $b \to 0$, $nb \to \infty$, as $n \to \infty$. Then we obtain for the bias part of the deviation, applying Taylor expansions with Lagrange and integral remainder terms [see Müller (1984)]:

LEMMA 5.3. *Assume that in the model* (1.1), $g \in S_{k,\zeta}$ $(k \geq 0,\ \zeta \in [0,1])$, *that the kernel used in* (5.8) *satisfies* $K \in \mathcal{M}_{k+\xi}$ [*see* (3.5)] *and that the error variables* (ε_i) *satisfy Assumption* A *or* B. *Defining* $\xi = \min(1, k+\zeta, \rho)$ (*for* ρ *see Assumption* B), *we have for any compact subinterval* $I \subset (0,1)$

(i) $$E(\hat{g}_n(t)) - g(t) = g^{(k)}(t)\cdot b^k \tilde{B}_k + o(b^k) + O(n^{-\xi}),$$

uniformly on I [*where* $\tilde{B}_k$ *is defined after* (4.4)];

(ii) $$\sup_{t\in I}\left|E(\hat{g}_n(t)) - g(t)\right| \to 0, \quad \text{as } n \to \infty,$$

and if $k + \zeta > 0$, *then*

$$\sup_{t\in I}\left|E(\hat{g}_n(t)) - g(t)\right| = O(b^{k+\zeta} + n^{-\xi}).$$

(iii) *If the error variables satisfy Assumption* A, *then*

$$\operatorname{Var}(\hat{g}_n(t)) = \frac{\sigma^2(t)}{f(t)nb} V(1 + o(1)), \quad \text{uniformly on } I,$$

where

$$V = \int K(x)^2\, dx.$$

Summing up, we obtain

THEOREM 5.1. *Assume that* $g \in S_{k,\zeta}$ *and* $K \in \mathcal{M}_{k+\zeta}$ *and that the error variables satisfy Assumption* B *with some given* $\rho > 0$. *Assume that* K *is Lipschitz continuous on* $\mathbb{R}$.

(A) *Let* $\xi = \min(1, k+\zeta, \rho)$. *If* b *satisfies for some* $\delta > 0$ *and some* $\eta \in (0, s-2)$

$$\liminf_{n\to\infty} nb^{1+\delta} > 0, \tag{5.9}$$

$$\liminf_{n\to\infty} n^{\xi} b^{k+\zeta} > 0, \tag{5.10}$$

$$\liminf_{n\to\infty} (nb/\log n)^{1/2} n^{-2/(s-\eta)} > 0, \tag{5.11}$$

then we have on any compact interval $I \subset (0,1)$ *for the estimator* (5.8)

$$\sup_{t\in I}|\hat{g}_n(t) - g(t)| = \begin{cases} O\left(\left(\frac{\log n}{nb}\right)^{1/2} + b^{k+\zeta}\right), & \text{if } k+\zeta > 0, \\ o(1), & \text{if } k+\zeta = 0, \end{cases} \quad a.s.$$

(B) *If* $k + \zeta \geq \frac{1}{2}$, *in Assumption* B $\rho \geq \frac{1}{2}$ *and* $s > 4 + 2/(k+\zeta)$ *and if we choose* $b \sim (\log n/n)^{1/(2(k+\zeta)+1)}$, *then we have*

$$\sup_{t\in I}|\hat{g}_n(t) - g(t)| = O\left(\left(\frac{\log n}{n}\right)^{(k+\zeta)/(2(k+\zeta)+1)}\right) \quad a.s.$$

PROOF. Observing that $K \in \mathrm{Lip}_\delta(\mathbb{R})$ for all $0 < \delta \leq 1$, we conclude from (5.9) that (5.3) is valid.

The fact that $\Sigma W_i(t) \equiv 1$ for kernel estimates (5.8) implies (5.4). Setting $\alpha_n = (\log n/nb)^{1/2}$, (5.5) follows from the fact that $\sup_{t\in I}\max_{1\leq i\leq n}|W_i(t)| = O((nb)^{-1})$ and from (5.11). (5.6) follows using Lemma 5.3(iii). (5.10) and Lemma 5.3(ii) yield the bias term $O(b^{k+\zeta})$. (B) is an immediate consequence of (A). □

REMARKS. (i) If we apply smooth boundary kernels near the ends of the interval $[0,1]$, the result extends to $I = [0,1]$.

(ii) In case we deal with a linear scheme of (ε_i) or we want to obtain results on convergence in probability, condition (5.11) in Theorem 5.1A can be relaxed to $\liminf_{n\to\infty}(nb/\log n)^{1/2}n^{-1/(s-\eta)} > 0$ (see the remark after Lemma 5.2), and accordingly we need then in Theorem 5.1B the weaker condition $E|\varepsilon_{i,n}|^s < M < \infty$ with $s > 2 + 1/(k+\zeta)$.

(iii) Of related interest are results by Bierens (1983) on the uniform consistency of the kernel estimator which he obtained for correlated data in the random design regression model under stationarity assumptions.

(iv) The results can be extended to the estimation of derivatives as well as to higher dimensions. If we estimate a mixed partial derivative of the orders $(\nu_1, \nu_2, \ldots, \nu_d)$ of a regression function g: $\mathbb{R}^d \to \mathbb{R}$ and write $\nu = \nu_1 + \cdots + \nu_d$, the choice $b \sim (\log n/n)^{1/(2(k+\zeta)+d)}$ under appropriate conditions (which are obvious generalization of the conditions of Theorem 5.1) yields the uniform rate $O((\log n/n)^{(k+\zeta-\nu)/(2(k+\zeta)+d)})$.

REFERENCES

ANSCOMB, F. J. and TUKEY, J. W. (1963). The examination and analysis of residuals. *Technometrics* **5** 141–160.

BIERENS, J. (1983). Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Amer. Statist. Assoc.* **78** 699–707.

BILLINGSLEY, P. (1979). *Probability and Measure*. Wiley, New York.

BREIMAN, L. and MEISEL, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.* **71** 301–307.

CARROLL, R. J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10** 1224–1233.

CHENG, K. F. and LIN, P. E. (1981). Nonparametric estimates of a regression function. *Z. Wahrsch. verw. Gebiete* **57** 223–233.

FULLER, W. A. and RAO, J. N. K. (1978). Estimation for a linear regression model with unknown diagonal covariance matrix. *Ann. Statist.* **6** 1149–1158.

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

HALL, P. and MARRON, J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab. Theory Rel. Fields* **74** 567–582.

KENDALL, M. G. and STUART, A. (1969). *The Advanced Theory of Statistics* **1**. Griffin, London.

LAMPERTI, J. (1966). *Probability*. Benjamin, New York.

MATLOFF, N., ROSE, R. and TAI, R. (1984). A comparison of two methods for estimating optimal weights in regression analysis. *J. Statist. Comput. Simulation* **19** 265–274.

MÜLLER, H.-G. (1983). Beiträge zur nichtparametrischen Kurvenschätzung. Dissertation, Universität Ulm.

MÜLLER, H.-G. (1984). Optimal designs for nonparametric kernel regression. *Statist. Probab. Lett.* **2** 285–290.

MÜLLER, H.-G. and STADTMÜLLER, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15** 182–201.

PARZEN, E. (1962). On estimation of a probability density and mode. *Ann. Math. Statist.* **33** 1065–1076.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Nonparametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385–392.

RAO, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.

REINSCH, C. (1967). Smoothing by spline functions. *Numer. Math.* **10** 177–183.

RICE, J. (1984). Bandwidth choice for nonparametric kernel regression. *Ann. Statist.* **12** 1215–1230.

ROSE, R. L. (1978). Nonparametric estimation of weights in least-squares regression analysis. Ph.D. dissertation, Univ. of California, Davis.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 642–649.

SILVERMAN, B. W. (1984). Smoothing splines—the equivalent variable kernel method. *Ann. Statist.* **12** 898–916.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–50.

SACKS, J. and YLVISAKER, D. (1970). Designs for regression problems with correlated errors. III. *Ann. Math. Statist.* **41** 2057–2074.

STADTMÜLLER, U. (1986a). Asymptotic properties of nonparametric curve estimates. *Period. Math. Hungar.* **17** 83–108.

STADTMÜLLER, U. (1986b). An inequality between kernel estimators with global and local bandwidths. *Statistics and Decisions* **4** 353–361.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.

UNIVERSITÄT MARBURG
INSTITUT FÜR MEDIZINISCH-BIOLOGISCHE STATISTIK
ERNST-GILLER-STRASSE 20
D-3550 MARBURG
FEDERAL REPUBLIC OF GERMANY

UNIVERSITÄT ULM
ABTEILUNG FÜR MATHEMATIK
OBERER ESELSBERG
D-7900 ULM
FEDERAL REPUBLIC OF GERMANY