The Annals of Statistics
1982, Vol. 10, No. 4, 1224–1233

ADAPTING FOR HETEROSCEDASTICITY IN LINEAR MODELS[1]

BY RAYMOND J. CARROLL

University of North Carolina at Chapel Hill

In a heteroscedastic linear model, it is known that if the variances are a parametric function of the design, then one can construct an estimate of the regression parameter which is asymptotically equivalent to the weighted least squares estimate with known variances. We show that the same is true when the only thing known about the variances is that they are determined by an unknown but smooth function of the design or the mean response.

1. Introduction. We are interested in efficient regression parameter estimation in a heteroscedastic linear model given by

$$Y_{ij} = x_i'\beta + \sigma_i\varepsilon_{ij},\ i = 1, \cdots, n, j = 1, \cdots, m_i, \Sigma m_i = N. \tag{1.1}$$

Here Y_{ij} is the response of the jth replicate at the design point x_i (a p-vector), β is the unknown regression parameter of interest, $\{\sigma_i\}$ express the heteroscedasticity in the model and $\{\varepsilon_{ij}\}$ are i.i.d. with variance one and distribution function F assumed symmetric about zero but otherwise unknown. Theoretical analysis of the model (1.1) has traditionally fallen into one of the two areas we describe below.

The parametric approach generally assumes

$$\sigma_i^2 = H(x_i, \theta) \quad \text{or} \quad H(x_i'\beta, \theta), \quad H \text{ known}. \tag{1.2}$$

See Hildreth and Houck (1968), Froehlich (1973), Dent and Hildreth (1977), Box and Hill (1974), Jobson and Fuller (1980) and Carroll and Ruppert (1982). Once a parametric assumption such as (1.2) is made, one computes estimates of the $r \times 1$ unknown parameter θ, next estimates

$$\hat\sigma_i^2 = H(x_i, \hat\theta) \quad \text{or} \quad H(x_i'\hat\beta, \hat\theta),$$

as appropriate, and then constructs a weighted estimate of β. If we denote the weighted least squares estimate based on the true weights by $\hat\beta_T$ and the weighted estimate based on the estimates $\{\hat\sigma_i\}$ by $\hat\beta_E$, we get a well-known result:

RESULT 1. For the parametric approach, in large samples there is no cost due to estimating $\{\sigma_i\}$, i.e., $\hat\beta_T$ and $\hat\beta_E$ have the same limiting normal distribution.

This result is proved rigorously and extended to robust estimation by Carroll and Ruppert (1982); Carroll (1982) shows it holds even if the dimension p of β increases with N, e.g., $p^2/N \to 0$ generally suffices. See also Williams (1975).

The nonparametric approach differs quite radically. Here,

$$\sigma_i^2 = H(x_i) \quad \text{or} \quad H(x_i'\beta), \quad H \text{ unknown}. \tag{1.3}$$

Since $H(\cdot)$ is assumed completely unknown, the standard method is to get information about $H(\cdot)$ by replication ($m_i > 1$). Fuller and Rao (1978) consider the situation often seen

Received May 1981; revised April 1982.

[1] University of North Carolina at Chapel Hill, Department of Statistics. This work was supported by the Air Force Office of Scientific Research under Contract AFOSR-80-0080.

AMS 1970 *subject classification.* Primary 62J05; secondary 62G35.

Key words and phrases. Linear models, regression, heteroscedasticity, nonparametric regression.

in practice that the number of design points $n \to \infty$, but each m_i stays bounded. Their method is to fit least squares estimates $(\hat{\beta}_L)$ to the data, compute predicted values $(t_i = x_i'\hat{\beta}_L)$ and residuals $(r_{ij} = Y_{ij} - t_i)$ and estimate

$$\hat{\sigma}_i^2 = \frac{1}{m_i} \sum_j r_{ij}^2. \tag{1.4}$$

With these estimates one then performs weighted least squares, obtaining what we shall denote by $\hat{\beta}_{FR}$. By delicate and very interesting calculations, they obtain an important result which had not been previously known or appreciated:

RESULT 2. In the nonparametric approach, there is a cost due to not knowing $\{\sigma_i\}$, i.e., $\hat{\beta}_T$ and $\hat{\beta}_{FR}$ have different limiting distributions.

We explore here the possibility of closing the wide gap between Results 1 and 2, at least in an asymptotic sense. Specifically, we explore methods for which the nonparametric approach (1.3) is used but for which Result 1 obtains. In other words, we will show that situations exist in which nothing specific is known about the variance function, but estimation of β can be done asymptotically as well as if the variance function were completely known.

A key feature of many—but as Fuller and Rao (1978) note, not all—heteroscedastic regression problems is that the variances appear to be smooth functions of the design or mean response; we use the term "smooth" loosely here, but generally will mean that the variance function $H(\cdot)$ has a continuous first derivative. This smoothness suggests that if x_1 and x_2 are very close, so too should be $H(x_1)$ and $H(x_2)$. This suggests that information about $H(x_1)$ can be obtained from data at x_2. Hence, we will study the nonparametric models

$$\sigma_i^2 = H(x_i) \quad \text{or} \quad H(x_i'\beta), \quad H \text{ unknown but smooth.} \tag{1.5}$$

This approach of sharing information contrasts with that of the nonparametric method (1.3)—(1.4), which only uses data at x_1 to estimate $H(x_1)$. By sharing information we should now get good consistent estimates of $H(\cdot)$, which enables us in certain circumstances to get better estimates of β for which Result 1 holds.

We specifically consider only two cases. In Section 2 we discuss simple linear regression, while in Section 3 we assume that the variance is a smooth function of the mean. The technical details are not trivial and the notation is rather messy, but the basic idea is simple and can be described as follows. Under the second part of (1.5) for example, we have

$$E(Y_{ij} - x_i'\beta)^2 = H(x_i'\beta).$$

Thus, for the residuals r_{ij} we have

$$Er_{ij}^2 = E(Y_{ij} - x_i'\hat{\beta}_L)^2 \approx H(x_i'\beta). \tag{1.6}$$

Equation (1.6) puts us in the realm of nonparametric regression of squared residuals on a function $H(\cdot)$. Even if one goes no further, there is already a huge literature which can be exploited to define nonparametric regression estimates (Watson, 1964; Rosenblatt, 1969; Stone, 1977; Mack and Silverman, 1980; Johnston, 1982); this we do. If one goes further and makes the often reasonable assumption that $H(\cdot)$ is monotone, isotonic regression could be used (Wright, 1978). Such isotonic estimates should work well in practice but we have been unable to develop a theory for them.

Throughout this paper, x and β refer to p-vectors, while α and c are scalars. For example, the heteroscedastic simple linear regression model is, with $x_i' = (1, c_i)$ and $\beta' = (\alpha_0, \alpha_1)$,

$$Y_{ij} = x_i'\beta + \sigma_i\varepsilon_{ij} = \alpha_0 + \alpha_1 c_i + \sigma_i\varepsilon_{ij}. \tag{1.7}$$

NOTE ADDED IN PROOF. After this paper was accepted for publication, I was informed by Professor N. Matloff (Department of Statistics and Electrical and Computer Engineering, University of California at Davis) that in 1978 his student Dr. Robin Lawrence Rose, in an unpublished dissertation, proposed methods of estimation similar to those investigated here, and performed Monte-Carlo experiments for these methods.

2. Simple linear regression. We first consider simple linear regression. This is the only case for which we have been able to obtain results in which the variance is a function of the design alone, as would be the case in the random coefficient model of Hildreth and Houck (1978). In the next section, we discuss the situation in which the variance is a function of the mean response.

Thus, in this section, the model is given by (1.7), where

$$\sigma_i^2 = H(c_i), \quad H(\cdot) \text{ unknown.}$$

Much of the literature for the nonparametric regression problem assumes that the independent or predictor variables are themselves random. In order to make the most efficient presentation, we will follow this lead, making the assumption for model (1.7) that $\{\varepsilon_{ij}\}$ and $\{c_i\}$ are sets of i.i.d. random variables independent of one another. After the statement of Theorem 1, we will discuss the case that $\{c_i\}$ is a set of fixed constants. We will first present and discuss the assumptions, and then state the first result.

First, from Watson (1964) and Johnston (1982), a plausible kernel-type estimate of H is

$$(2.1) \qquad \hat{H}_N(c) = \sum_{i=1}^n \sum_{j=1}^{m_i} r_{ij}^2 K\left(\frac{c_i - c}{b(N)}\right) \left\{\sum_{i=1}^n \sum_{j=1}^{m_i} K\left(\frac{c_i - c}{b(N)}\right)\right\}^{-1}.$$

The weighted estimate $\hat{\beta}_w$ is formed by setting

$$\hat{\sigma}_i^2 = \hat{H}_N(c_i)$$

and then performing weighted least squares.

In order that information about the scalar function $H(\cdot)$ can be shared and in order to avoid being subject to the Fuller and Rao Result 2, we need the design to be eventually dense in a set such as an interval. This will enable us to estimate $H(\cdot)$ uniformly well. One can do this under the following assumption.

ASSUMPTION 1. $\{x_i\}$ have density function f positive on its compact support $\mathscr{I}$. Further, on $\mathscr{I}$, f has two continuous derivatives.

Note that Assumption 1 is really designed for regression problems and not for factorial designs. Naturally, we also require that H be smooth:

ASSUMPTION 2. H and its first derivative are continuous on $\mathscr{I}$.

In order to make sure that no infinite weights occur in our weighted regression, we need

ASSUMPTION 3. H has a positive infimum on $\mathscr{I}$.

We also need some assumptions on the kernel $K(\cdot)$ and bandwidth $b(N)$ in (2.1).

ASSUMPTION 4. $K(\cdot)$ is a symmetric density function. It has compact support, three continuous derivatives, and its support includes an open set containing $\mathscr{I}$.

ASSUMPTION 5. The bandwidth $b(N)$ satisties $Nb(N)^4 \to 0$.

ASSUMPTION 6. The bandwidth $b(N)$ satisfies $N^{5/4}b(N)^4 \to \infty$.

Finally, we make assumptions relating to the uniqueness of the design; these are reasonably standard assumptions even in the parametric approach. Recall $x' = (1, c)$.

ASSUMPTION 7. $E(xx')$ and $E\{xx'H^{-1}(c)\}$ are positive definite.

THEOREM 1. *Under Assumptions* 1–7 *and the condition that* $E\varepsilon_1^6 < \infty$, $\hat{\beta}_W$ *and* $\hat{\beta}_T$ *have the same normal limit distribution, with mean* β *and covariance* $N^{-1}Exx'H^{-1}(c)$.

The proof is in Section 5. Assumptions 5 and 6 probably can be weakened.

REMARKS. The problem discussed in this section is a special case of one in which the variance is a function of the design and $p \geq 2$; when $p \geq 3$, H is a function of a vector argument. For larger values of p, the rate of convergence to H of the estimator (2.1) will be slower and the proof given in the appendix will break down, since Theorem 5.1 will not be true. We believe brute force (Taylor series) can be used to extend Theorem 1 to the case $p \geq 3$, but an alternative approach would be preferable.

Theorem 1 can be extended to the case where the predictor variables $\{c_i\}$ are fixed constants by assuming that they "act i.i.d." in all essential aspects; this is rather untidy. Alternatively, one could replace (2.1) by the Priestley-Chao estimator studied by Benedetti (1977). For this estimator, certain technical difficulties can be avoided because there is no random denominator term as in (2.1); the assumptions, however, remain basically unchanged with the exception of Assumption 1.

3. Variance a function of the mean. Here we consider the model (1.1) with

$$\sigma_i^2 = H(x_i'\beta) = H(\tau_i), \quad H \text{ unknown.} \tag{3.1}$$

The variance is often considered a function of the mean as in (3.1) because residual plots fall in a fan-shaped pattern; see Box and Hill (1974), Bickel (1978), Jobson and Fuller (1980) and Carroll and Ruppert (1982).

Note that

$$\tau_i = \text{true mean response} = x_i'\beta$$

and define the predicted values as

$$t_i = x_i'\hat{\beta}_L,$$

where $\hat{\beta}_L$ is least squares estimator. Following the same reasoning as in the previous section, the estimator of H becomes

$$\hat{H}_N(s) = \{Nb(N)\}^{-1}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m_i} r_{ij}^2 K\left(\dfrac{t_i - s}{b(N)}\right)\left[\{Nb(N)\}^{-1}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{m_i} K\left(\dfrac{t_i - s}{b(N)}\right)\right]^{-1},$$

and the estimated variances are $\{\hat{H}_N(t_i)\}$.

THEOREM 2. *Under the assumptions of Theorem* 1, *but in Assumption* 7 *replacing* $H(c)$ *by* $H(x'\beta)$, $\hat{\beta}_W$ *and* $\hat{\beta}_T$ *have the same normal limit distribution with mean* β *and covariance* $N^{-1}E\{xx'H^{-1}(x'\beta)\}$.

The proof is in Section 5.

4. A Monte-Carlo study. We performed a small Monte-Carlo experiment to see if the previous results make any sense even in an ideal situation. We took the model to be simple linear regression.

$$Y_i = \alpha_0 + \alpha_1 c_i + \sigma_i\varepsilon_i, \qquad i = 1, \cdots, N = 60. \tag{4.1}$$

Here $\{\varepsilon_i\}$ are standard normal random variables, $(\alpha_0, \alpha_1) = (50, 60)$, and $\{c_i\}$ are i.i.d.

uniform on the interval (−½, ½). The normal random numbers were generated by the IMSL routine GGNPM, while the uniform numbers used GGUBS. The number of Monte-Carlo simulations for each situation was 500.

We estimated the function H by $\hat{H}_N$ of (2.1), with

$$K(v) = \begin{cases} 3(1 - |v|)^2/2 & |v| \leq 1, \\ 0 & |v| \geq 1, \end{cases} \tag{4.2}$$

$$b(N) = 0.13. \tag{4.3}$$

The particular choice for $b(N)$ was arbitrary, although on average approximately 8 observations are used in constructing $\hat{H}_N$ at each design point. While $K(\cdot)$ does not strictly satisfy Assumption 5, it does have a continuous first derivative which should suffice. In Table 1, the weighted least squares estimate with weights generated by $\hat{H}_N$ is denoted NONPAR. The least squares estimate is LSE.

Three models for the variances were considered. The first, given by Jobson and Fuller (1980), is

$$\sigma_i^2 = a_1 + a_2\tau_i^2, \qquad \tau_i = \alpha_0 + \alpha_1 c_i. \tag{4.4}$$

For our simulations we chose $a_1 = 100$, $a_2 = 0.25$. Our second model is one of more severe heteroscedasticity,

$$\sigma_i = a_1 \exp(a_2 | \tau_i |), \tag{4.5}$$

where $a_1 = 0.25$ and $a_2 = 0.04$. This type of model is mentioned by Bickel (1978). The third model is one of severe heteroscedasticity

$$\sigma_i = a_1 \exp(a_2\tau_i^2), \tag{4.6}$$

where

$$(a_1, a_2) = (¼, 1/3200).$$

We also constructed a third estimator PARM based on the parametric model (4.4). Our intentions in doing this were (a) to see if the nonparametric estimate is at all reasonable when compared to an estimate based on a correct parametric model (4.4) for the variances, and (b) to see if the nonparametric estimate is more robust than the parametric estimate if the variance model is badly misspecified, i.e., (4.6) holds but estimation is done as if (4.4) holds.

TABLE 1

Results of the Monte-Carlo Study of Section 4 for the model $\tau_i = EY_i = 50 + 60\, c_i$, c_i *Uniform* (−½, ½). *The models for* $\text{Var}(Y_i) = \sigma_i^2$ *are: Model 1* $\sigma_i^2 = a_1 + a_2\tau_i^2$, $(a_1, a_2) = (100, 0.25)$; *Model 2* $\sigma_i = a_1\exp(a_2 | \tau_i |)$, $(a_1, a_2) = (0.25, 0.04)$; *Model 3* $\sigma_i = a_1\exp(a_2\tau_i^2)$, $(a_1, a_2) = (0.25, 1/3200)$.

		$\alpha_0 = 50$		$\alpha_1 = 60$	
Estimator	Variance Model	Bias	MSE*	Bias	MSE*
LSE	1	.052	13.27	.925	172.07
PARM	1	.040	13.27	.711	138.58
NONPAR	1	.066	14.80	.727	144.46
LSE	2	.016	12.87	.106	231.20
PARM	2	.011	11.45	.049	100.00
NONPAR	2	.007	10.18	.037	80.34
LSE	3	.004	10.98	.032	200.41
PARM	3	.003	9.85	.016	96.09
NONPAR	3	.002	9.19	.014	88.88

* *The actual* MSE *for Model* 2 *is the figure given divided by* 10^2, *while the figure for Model* 3 *should be divided by* 10^3.

The estimate PARM is constructed as follows:

(i) Define P as in Jobson and Fuller (1980).

(ii) Let $\hat\beta_L$ = LSE, with $\hat\beta'_L = (\hat\alpha_0, \hat\alpha_1)$.

(iii) Let r^2 be the vector of squared residuals, i.e. squares of $Y_i - x_i'\hat\beta_L$.

(iv) Minimize $(r^2 - P\mathbf{\hat a})'(r^2 - P\mathbf{\hat a})$ for $\mathbf{\hat a} \geq 0$, where $\mathbf{\hat a}' = (\hat a_1, \hat a_2)$

(v) Define $\hat\sigma_i^2 = \hat a_1 + \hat a_2(\hat\alpha_0 + \hat\alpha_1 c_i)^2$

(vi) Compute a weighted estimate $\hat\beta_p$ and residuals $r_{iw} = Y_i - x_i'\hat\beta_p = Y_i - \hat\alpha_{0p} - \hat\alpha_{1p}c_i$.

(vii) Repeat steps (iv) and (v), replacing $(\hat\alpha_0, \hat\alpha_1)$ by $(\hat\alpha_{0p}, \hat\alpha_{1p})$ in (v).

(viii) Recompute a weighted estimate, call it PARM.

The outcomes of the simulations are given in Table 1. The results are quite encouraging and suggest that there are instances where our nonparametric estimation of the variances can work well, particularly for larger sample sizes.

5. Proof. Because the details are lengthy, we sketch the proofs only for the case $m_i \equiv 1$. As a shorthand notation, identify Assumptions 1–7 as A1, A2, $\cdots$, A7. Consider first simple linear regression in Section 2. We have the following.

THEOREM 5.1. *If the supremum is taken over the support of the design $\mathscr{I}$ (assumed compact), then*

$$N^{1/4}\sup|\hat H_N(c) - H(c)| \to_p 0.$$

PROOF OF THEOREM 5.1. Rewrite (2.1) as $\hat H_N = \hat G_N/\hat f_N$ and

$$\hat G_N(c) = \hat G_{N1}(c) - 2\hat G_{N2}(c) + \hat G_{n3}(c)$$
$$= \{Nb(N)\}^{-1}\textstyle\sum_{i=1}^N [\sigma_i^2\varepsilon_i^2 - 2\sigma_i\varepsilon_i x_i'(\hat\beta_L - \beta) + \{x_i'(\hat\beta_L - \beta)\}^2]K\left(\dfrac{c_i - c}{b(N)}\right).$$

Because both $K(\cdot)$ and the support of the design are bounded and $\hat\beta_L = \beta + O_p(N^{-1/2})$, we have

$$N^{1/4}\sup|\hat G_{N3}(c)| \to_p 0.$$

Routine but detailed weak convergence arguments using Theorem 12.3 of Billingsley (1968), A4–A7 and $E\varepsilon^6 < \infty$ can be used to show that

$$\begin{aligned}&N^{1/4}\sup|\hat G_{N2}(c)| \to_p 0, && N^{1/4}\sup|\hat f_N(c) - E\hat f_N(c)| \to_p 0,\\ &N^{1/4}\sup|\hat G_{N1}(c) - E\hat G_{N1}(c)| \to_p 0, && N^{1/4}\sup|E\hat G_{N1}(c)/E\hat f_N(c) - H(c)| \to 0.\end{aligned} \tag{5.1}$$

The first part of (5.1) is simple enough. It follows from direct weak convergence arguments after one shows that, from the central limit theorem, one can replace $\hat\beta_L - \beta$ by

$$\{E(xx')\}^{-1}N^{-1}\textstyle\sum_{i=1}^N x_i\sigma_i\varepsilon_i.$$

The second and third parts of (5.1) can be shown directly by weak convergence arguments, but they are also essentially known from the nonparametric regression literature. The fourth part of (5.1) is merely tedious algebra; one has to be quite careful with end points.

Now recall once again that the model for Theorem 1 is

$$Y_i = x_i'\beta + \sigma_i\varepsilon_i = \alpha_0 + \alpha_1 c_i + \sigma_i\varepsilon_i, \qquad \mathrm{Var}(\varepsilon_i) = 1.$$

PROOF OF THEOREM 1. First note because $E\varepsilon^2 < \infty$ and A7 holds, $N^{1/2}(\hat\beta_T - \beta)$ has the normal limit distribution claimed in Theorem 1. It thus suffices to show that

$$N^{1/2}(\hat\beta_w - \hat\beta_T) \to_p 0.$$

Recall, $\hat\beta_w$ is the weighted estimator based on the adaptive weights (2.1). Because $\hat\beta_T$ is asymptotically normal, the design is bounded, $H(c) > 0$ and $E\varepsilon^6 < \infty$, one can use Theorem 5.1 to see that it suffices to show that

$$N^{-1/2}\Sigma\, x_i\varepsilon_i\{\hat{H}_N(c_i) - H(c_i)\}/\sigma_i^3 \to_p 0, \tag{5.2}$$

where $x_i' = (1, c_i)$ as before. By the proof of Theorem 5.1 it suffices to show

$$N^{-1/2}\sum x_i\varepsilon_i[\hat{G}_N(c_i) - E\hat{G}_{N1}(c_i)]/\sigma_i^3 E\hat{f}_N(c_i) \to_p 0, \tag{5.3}$$

$$N^{-1/2}\sum x_i\varepsilon_i E\hat{G}_{N1}(c_i)[\hat{f}_N(c_i) - E\hat{f}_N(c_i)]/\sigma_i^3(E\hat{f}_N(c_i))^2 \to_p 0, \tag{5.4}$$

$$N^{-1/2}\sum x_i\varepsilon_i\{E\hat{G}_{N1}(c_i)/E\hat{f}_N(c_i) - H(c_i)\}/\sigma_i^3 \to_p 0. \tag{5.5}$$

In the expressions above, $E\hat{G}_{N1}(c_i)$ refers to $E\hat{G}_{N1}(c)$ evaluated at $c = c_i$, and similarly for $E\hat{f}_N(c_i)$. We will only sketch (5.3) as (5.4) and (5.5) are much easier. Rewrite (5.3) as

$$\begin{aligned}&\{N^{3/2}b(N)\}^{-1}\textstyle\sum_i\sum_j \dfrac{x_i\varepsilon_i}{\sigma_i^3E\hat{f}_N(c_i)}\left\{\sigma_j^2K\left(\dfrac{c_j - c_i}{b(N)}\right) - E\hat{G}_{N1}(c_i)\right\}\\ &\qquad + \{N^{3/2}b(N)\}^{-1}\textstyle\sum_i\sum_j \dfrac{x_i\varepsilon_i}{\sigma_i^3E\hat{f}_N(c_i)}\\ &\qquad\qquad \cdot\,[\sigma_j^2(\varepsilon_j^2 - 1) - 2\sigma_j\varepsilon_jx_j'(\hat{\beta}_L - \beta) + \{x_j'(\hat{\beta}_L - \beta)\}^2]K\left(\dfrac{c_j - c_i}{b(N)}\right).\end{aligned} \tag{5.6}$$

Each term in (5.6) converges in probability to zero. The first term and the first part of the second term only require computing second moments, remembering that $\{x_i\}$ and $\{\sigma_i\}$ are uniformly bounded and noting that $\{E\hat{f}_N(c_i)\}$ are bounded away from zero. The third part of the second term is easy. For the second part of the second term, it suffices to prove the result when we replace $\hat{\beta}_L - \beta$ by

$$\{E(xx')\}^{-1}N^{-1}\textstyle\sum_{i=1}^N x_i\varepsilon_i\sigma_i. \tag{5.7}$$

Having done this, one then computes second moments. In these steps the full strength of the assumption $E\varepsilon^6 < \infty$ is used.

We next sketch the proof for Theorem 2. The first step is a version of Theorem 5.1. Recall that $\tau_i = x_i'\beta = EY_i$, $t_i = x_i'\hat{\beta}_L$. The definitions of $\hat{H}_N$ and $\hat{\beta}_w$ are given in Section 3, while $\hat{f}_N$ is the inverted term in the definition of $\hat{H}_N$.

THEOREM 5.2.

$$N^{1/4}\sup|\hat{H}_N(s) - H(s)| \to_p 0.$$

PROOF OF THEOREM 5.2. It is first of all possible to show by weak convergence techniques that

$$N^{1/4}\sup|\hat{f}_N(s) - E_*\hat{f}_N(s)| \to_p 0, \tag{5.8}$$

where $f(\cdot)$ is the density of $\{x_i'\beta\}$,

$$\hat{f}_N(s) = \{Nb(N)\}^{-1}\textstyle\sum_{i=1}^N K\left(\dfrac{t_i - s}{b(N)}\right),$$

$$E_*\hat{f}_N(s) = E\{Nb(N)\}^{-1}\textstyle\sum_{i=1}^N K\left(\dfrac{\tau_i - s}{b(N)}\right);$$

i.e., E_* means we replace t_i by $\tau_i = x_i'\beta$ and then take expectations. To show (5.8), first recall that the support of the design is bounded, so that $|t_i - \tau_i| = O_p(N^{-1/2})$ uniformly in i. This means that, uniformly in i,

$$\{(t_i - \tau_i)/b(N)\}^3 \to_p 0.$$

Using this and compactness, one expands to get

$$\begin{aligned}&\hat{f}_N(s) = \{Nb(N)\}^{-1}\textstyle\sum_{i=1}^N\\ &\quad\left\{K\left(\dfrac{\tau_i - s}{b(N)}\right) + \left(\dfrac{t_i - \tau_i}{b(N)}\right)K'\left(\dfrac{\tau_i - s}{b(N)}\right) + \dfrac{1}{2}\left(\dfrac{t_i - \tau_i}{b(N)}\right)^2K''\left(\dfrac{\tau_i - s}{b(N)}\right)\right\} + o_p(1).\end{aligned} \tag{5.9}$$

That the third term on the r.h.s. of (5.9) convergences in probability to zero at rate $N^{1/4}$ follows directly from A6. Denote the first term by $V_{N1}(s)$ and the second by $V_{N2}(s)$. The same weak convergence argument used in Theorem 5.1 shows $N^{1/4}\{V_{N1}(s) - E_* \hat{f}_N(s)\}$ converges in probability to zero uniformly on compacts. Dealing with $V_{N2}(s)$ is quite tricky. One first shows that it suffices to make the substitution (5.7) for $\hat{\beta}_L - \beta$. Then, a simple second moment computation shows that the finite dimensional distributions of the resulting modified process

$$V_{N2}^*(s) = \{N^2 b^2(N)\}^{-1} \textstyle\sum_\iota \sum_J x_\iota' \{E(xx')\}^{-1} x_J \sigma_J \varepsilon_J K'\left(\dfrac{\tau_i - s}{b(N)}\right)$$

converge in probability to zero; here, as in the tightness argument to follow, we use the fact that the support of K strictly includes the support of $\{x_\iota'\beta\}$ and, since K is a symmetric density, $\int K'(y)\, dy = 0$. Finally, tightness can be proven by using Theorem 12.3 of Billingsley (1968) (use his equation (12.51) with $\gamma = 2$ and $\alpha = 1 + a$, a very small); in doing this calculation, one must separate the cases $|t_2 - t_1| \geq db(N)$ and $< db(N)$ for a large constant d (t_1 and t_2 refer to Billingsley's notation). Because of (5.8) and Theorem 5.1, we now only need to prove Theorem 5.2 for

$$H_N^*(s) = \{Nb(N)\}^{-1} \textstyle\sum_{\iota=1}^N r_\iota^2 K\left(\dfrac{t_\iota - s}{b(N)}\right) - \{Nb(N)\}^{-1} E \sum_{\iota=1}^N \varepsilon_\iota^2 \sigma_\iota^2 K\left(\dfrac{x_\iota'\beta - s}{b(n)}\right). \tag{5.10}$$

One first makes the expansion of (5.10), as in (5.9), about $K((\tau_\iota - s)/b(N))$, and then argues as above and in the proof of Theorem 5.1; the assumption $E\varepsilon^6 < \infty$ is again vital here.

PROOF OF THEOREM 2. As in the proof of Theorem 1 we must show

$$N^{-1/2} \textstyle\sum x_\iota \varepsilon_\iota \{\hat{H}_N(t_\iota) - H(\tau_i)\}/\sigma_\iota^3 \to_p 0. \tag{5.11}$$

The proof parallels that of Theorem 1. Here, the difficult case is to show

$$N^{-1/2} \textstyle\sum_{\iota=1}^N x_\iota \varepsilon_\iota \{\hat{G}_N(t_\iota) - Q_N(\tau_\iota)\}/\{\sigma_\iota^3 E\hat{f}_N(\tau_\iota)\} \to_p 0, \tag{5.12}$$

where

$$\hat{H}_N = \hat{G}_N/\hat{f}_N, \qquad Q_n(x) = \{Nb(N)\}^{-1} E \textstyle\sum \varepsilon_\iota^2 \sigma_\iota^2 K\left(\dfrac{\tau_\iota - x}{b(N)}\right).$$

Rewrite (5.12) as

$$\begin{aligned} &\{N^{3/2}b(N)\}^{-1} \textstyle\sum_\iota \sum_j \dfrac{x_\iota \varepsilon_i}{\sigma_\iota^3 E\hat{f}_N(\tau_\iota)} r_J^2 \left\{K\left(\dfrac{t_J - t_\iota}{b(N)}\right) - K\left(\dfrac{\tau_J - \tau_\iota}{b(N)}\right)\right\} \\ &\quad + \{N^{3/2}b(N)\}^{-1} \textstyle\sum_\iota \sum_J \dfrac{x_i \varepsilon_\iota}{\sigma_\iota^3 E\hat{f}_N(\tau_\iota)} \left\{r_J^2 K\left(\dfrac{\tau_J - \tau_\iota}{b(N)}\right) - b(N) Q_N(\tau_n)\right\}. \end{aligned} \tag{5.13}$$

By a messy argument similar to that of (5.10), the second term in (5.13) can be shown to converge in probability to zero. For the first term, it suffices to show that for every $M > 0$,

$$\sup_{|\Delta| \leq M} |V_N(\Delta)| \to_p 0, \tag{5.14}$$

where

$$V_N(\Delta) = \{N^{3/2}b(N)\}^{-1} \textstyle\sum_\iota \sum_J \dfrac{x_\iota \varepsilon_\iota r_J^2}{\sigma_\iota^3 f(\tau_\iota)} \left\{K\left(\dfrac{\tau_J - \tau_\iota}{b(N)} + \dfrac{(x_\iota' - x_J')\Delta}{N^{1/2}b(N)}\right) - K\left(\dfrac{\tau_J - \tau_\iota}{b(N)}\right)\right\}.$$

Because, uniformly in i,

$$\{x_\iota'(\hat{\beta}_L - \beta)\}^2 = O_p(N^{-1}),$$

by A6 we must merely show (5.14) for the process $V_{N^*}(\Delta)$ which, in $V_N(\Delta)$, replaces r_j^2 by

$$\sigma_j^2\varepsilon_j^2 - 2\sigma_j\varepsilon_j x_j'(\hat{\beta}_L - \beta).$$

Divide V_{N^*} into the two processes

$$V_{N^*}(\Delta) = V_{N^*}^{(1)}(\Delta) + V_{N^*}^{(2)}(\Delta).$$

We now invoke the results of Bickel and Wichura (1971) on multiparameter stochastic processes, changing their equation (3) to

$$E\,|X(B)|^2 \le \mu(B)^{1+\gamma},$$

for some $\gamma > 0$. This shows (in order) that it suffices to show the results when we replace

$$N^{-1}\sum x_i x_i'$$

in the definition of $\hat{\beta}_L - \beta$ by $E(xx')$, and then that $V_{N^*}^{(j)}$ is tight with finite dimensional distributions converging in probability to zero. This proves (5.14) and completes the proof of Theorem 2.

NOTE. Handwritten detailed proofs are available from the author.

Acknowledgment. I would like to express my thanks to David Ruppert and Kent Bailey for stimulating discussions which led to this work. I would also like to thank the referees and associate editor for reading a difficult first draft, and to the editor for insisting on a readable presentation.

REFERENCES

BENEDETTI, J. K (1977). On the nonparametric estimation of regression functions. *J. Roy. Statist. Soc. B* **39** 248–253.

BICKEL, P. J. (1978). Using residuals robustly I: Tests for heteroscedasticity, nonlinearity. *Ann. Statist.* **6** 266–291.

BICKEL, P. J. and WICHURA, M. J. (1971). Convergence criteria for multiparameter stochastic processes and some applications. *Ann. Math. Statist.* **42** 1656–1670.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures.* Wiley, New York.

BOX, G. E. P. and HILL, W. J. (1974). Correcting inhomogeneity of variance with power transformation weighting. *Technometrics* **16** 385–389.

CARROLL, R. J. (1982). Estimation in heteroscedastic models when there are many parameters. *J. Statist. Plann. Infer.* To appear.

CARROLL, R. J. and RUPPERT, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10** 429–441.

DENT, W. T. and HILDRETH, C. (1977). Maximum likelihood estimation in random coefficient models. *J. Amer. Statist. Assoc.* **72** 69–72.

FROEHLICH, B. R . (1973). Some estimators for a random coefficient regression model. *J. Amer. Statist. Assoc.* **68** 329–335.

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.

HILDRETH, C. and HOUCK, J. P. (1968). Some estimators for a linear model with random coefficients. *J. Amer. Statist. Assoc.* **63** 584–595.

JOBSON, J. D. and FULLER, W. A. (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. *J. Amer. Statist. Assoc.* **75** 176–181.

JOHNSTON, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12** 404–414.

LENTH, R. V. (1977). Robust splines. *Commun. Statist. A* **6** 847–854.

MACK, Y. P. and SILVERMEN, B. W. (1980). Weak and strong uniform consistency of kernel regression estimates. Preprint.

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

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

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

WILLIAMS, J. S. (1975). Lower bounds on convergence rates of weighted least squares to best linear unbiased estimators. In *A Survey of Statistical Design and Linear Models* (J. N. Srivastava, ed.) 555–569. Academic, New York.

WRIGHT, I. W. and WEGMAN, E. J. (1980). Isotonic, convex and related splines. *Ann. Statist.* **8** 1023–1035.

WRIGHT, F. T. (1978). Estimating strictly increasing regression functions. *J. Amer. Statist. Assoc.* **73** 636–639.

9810 PARKWOOD DRIVE
BETHESDA, MARYLAND 20814