ADAPTING FOR HETEROSCEDASTICITY IN REGRESSION MODELS

Raymond J. Carroll[1]
David Ruppert[2]
University of North Carolina

and

Leonard A. Stefanski
North Carolina State University

[1]Research supported by the Air Force Office of Scientific Research AFOSR-F-49620-85-C-0144.

[2]Research supported by The National Science Foundation MCS-8100748.

ABSTRACT

We investigate the limiting behavior of a class of one-step M-estimators in heteroscedastic regression models. The mean function is assumed to be known up to parameters, but the variance function is considered an unknown function of a p dimensional vector. The variance function is to be estimated nonparametrically by a function of the absolute residuals from the current fit to the mean. Under a variety of conditions, we discuss when the estimates adapt for scale, i.e., estimate the regresion parameter just as well as if the scale function was known. Connections with the theory of optimal semiparametric estimation are made.

Key Words and Phrases : Heteroscedasticity, adaptation, nonparametric regression, M-estimator, robustness, generalized least squares.

INTRODUCTION

We study aspects of the effect of estimating weights in a generalization of the heteroscedastic regression model considered by Carroll & Ruppert (1982a). The observations are (y_i, x_i) for i=1,...,N, where y_i is the response and x_i is the vector of predictors. Let $L(y|x)$ and $S(y|x)$ define location and scale for the distribution of y given x. For example, $L(y|x)$ could be the mean of y given x, while $S(y|x)$ could be the standard deviation. The model is specified conditionally on the x_i by

$$L(y|x) = f(x,\beta) \; ; \tag{1.1}$$

$$S^2(y|x) = v_0(x) = v(x,\theta) \; . \tag{1.2}$$

Throughout it will be understood that $S(y|x)$ depends on a p dimensional subvector of x or on a p dimensional predictor z.

If L is expectation and S is standard deviation, then (1.1)-(1.2) specify the usual heteroscedastic regression model. In (1.2), the vector parameter θ is typically unknown.

Carroll & Ruppert (1982a) consider a class of M-estimators for the regression parameter β. Let $\hat{\theta}$ be any $N^{1/2}$ consistent estimator of θ, and let Ψ be an odd function with derivative Ψ_1. Let f_β be the derivative of f with respect to β and v_θ the derivative of v with respect to θ. Define

$$\epsilon = y - f(x,\beta) \; : \; e(\beta,\theta) = \epsilon / v^{\frac{1}{2}}(x,\theta) \; : \; z(\beta,\theta) = -(\partial/\partial\beta)e(\beta,\theta). \tag{1.3}$$

We assume that f and v_0 are such that the following limit result holds for some positive definite covariance matrix W_1 :

$$(1.4) \qquad N^{-1/2} \sum_{i=1}^{N} z_i(\beta,\theta)\, \Psi\{ e_i(\beta,\theta) \} \Rightarrow \text{Normal}(\, 0,\, W_1 \,) \ .$$

Define $\hat{\beta} = \hat{\beta}(\gamma)$ to be a solution to the estimating equation

$$(1.5) \qquad N^{-1/2} \sum_{i=1}^{N} z_i(\beta,\gamma)\, \Psi\{ e_i(\beta,\gamma) \} = 0 \ .$$

If for some positive definite matrix W_2

$$(1.6) \qquad N^{-1} \sum_{i=1}^{N} z_i(\beta,\theta)\, z_i(\beta,\theta)^T\, \Psi_1\{ e_i(\beta,\theta) \} \overset{p}{\longrightarrow} W_2 \ ,$$

then we would have the asymptotic limit result

$$(1.7) \qquad N^{1/2}\, (\hat{\beta}(\theta) - \beta) \Rightarrow \text{Normal}(\, 0,\, W_2^{-1}\, W_1\, W_2^{-1} \,) \ .$$

In most applications θ is unknown and thus $\hat{\beta}(\theta)$ is also unknown. A natural estimator of β is then $\hat{\beta}(\hat{\theta})$, where $\hat{\theta}$ is a $N^{1/2}$-consistent estimator of θ. A major question is to find conditions under which $\hat{\beta}(\cdot)$ adapts for estimating θ, i.e., $\hat{\beta}(\theta)$ and $\hat{\beta}(\hat{\theta})$ have the same limiting distribution (1.7). This has important statistical consequences, because if adaptation for θ occurs, then at least for large samples we can pretend that θ is known and use standard methods of inference. For example, if $\Psi(w) = w$ then (1.1) and (1.2) generally imply (1.4) In this case, the solution to (1.5) is a generalized least squares estimator computed by a weighted least squares algorithm with weights the inverse of $v(x_i,\hat{\theta})$. When adaptation occurs, inference can be made as if the weights were known, at least in the limit. For discussions which use second order calculations to understand the effect of estimating weights, see Rothenberg (1984) and Carroll, Ruppert & Wu (1986).

An easy extension of Theorem 1 in Carroll & Ruppert (1982a) states that under regularity conditions, $\hat{\beta}$ has the same asymptotic normal limit distribution as if θ were known as long as

$$(1.8) \qquad N^{-1} \sum_{i=1}^{N} e_i(\beta,\theta)\, \Psi_1\{ e_i(\beta,\theta) \}\, z_i(\beta,\theta)\, v_\theta(x_i,\theta) / v(x_i,\theta) \xrightarrow{p} 0 .$$

For generalized least squares, $\Psi(w) = w$ and condition (1.8) is almost always satisfied. For general M-estimators, condition (1.8) essentially reduces to an assumption of symmetry of the distribution of e_i given x_i.

In practice, the form of the location function $f(x_i,\beta)$ may be easily specified but the scale function $v(x_i,\theta)$ less clear, especially when the dimension p of x is greater than one. There are at least three strategies for coping with this case. The first is to assume that the scale is a function of the mean response. This reduces the scale estimation problem to a single dimension. See Carroll & Ruppert (1982a) and Davidian & Carroll (1986).

A second strategy is to use an empirical model for the scale function $v(x,\theta)$, for example a response surface quadratic or its square root. This approach has not been tried too often in the literature, although Box & Meyer (1985) seem to suggest the idea.

A third approach which we investigate is to estimate the scale function nonparametrically. Consider the case that (1.2) involves the variance. Carroll (1982) proposed that the unknown variance function v_0 be estimated by nonparametric kernel regression on the squared residuals from an unweighted least squares fit. Though v_0 is unknown, he showed that one can estimate β by generalized least squares just as well as if v_0 were known, so that the limit result (1.8) holds with Ψ the identity function. Unfortunately, while the result is interesting the conditions of his proof are unnecessarily stringent.

Subsequent to Carroll's paper,Matloff, Rose & Tai (1984) performed a simulation study, while Muller & Stadtmuller (1986) considered a fixed design. Robinson (1986) obtained Carroll's result for more than one dimension under far weaker regularity conditions. This is a nice piece of work, and we will borrow techniques from it where appropriate.

In this paper, we study adaptation in a broader framework by considering a class of one-step weighted M-estimators with weights estimated nonparametrically. The model for the means is allowed to be nonlinear. We allow a fairly broad class of smoothers, including kernel and nearest neighbor regression estimates. It is not the particular smoother that matters, but rather that the smoother satisfy certain reasonable moment conditions. In particular, robust smoothers could be used.

In the next section we introduce the class of estimators and the basic set of conditions. In sections 3 & 4 we use our work to provide examples for which adaptation occurs. In section 5, we link our work to the theory of information bounds for semiparametric models. In section 6 we address a few brief remarks to the case that the scale functions are known to be dependent on the mean response. All proofs are in the Appendix.

Section 2 : Estimators and Main Results

The choice of one step weighted M-estimates avoids the necessity of specifying extraneous conditions for asymptotic normality in nonlinear regression. The estimators include those studied by Carroll (1982) and Robinson (1986) as special cases. Let $\hat{\beta}_*$ be any $N^{1/2}$ consistent estimate of β.

Write

$$\epsilon(\beta) = \epsilon(y,x,\beta) = y - f(x,\beta) \quad ; \quad d(\beta) = d(x,\beta) = f_\beta(x,\beta) \quad .$$

It is most compact to treat (y,x) as a random variable, the translation to the case of fixed x's being immediate. In what follows we use the notation of Pollard (1984), so that for any random variable G(y,x), $\mathbb{P}\{ G(x,y) \}$ is the espectation of G, and $\mathbb{P}_N\{ G(x,y) \}$ is the average of the values $G(x_i,y_i)$.

If $\hat{g}(x)$ is any estimate of the scaling function $v_0(x)$, then as in Bickel (1975) a weighted one step M-estimator of β is

(2.1)
$$\hat{\beta} = \hat{\beta}_* + \{ A_N(\hat{\beta}_*,\hat{g}) \}^{-1} B_N(\hat{\beta}_*,\hat{g}), \text{ where}$$
$$A_N(\beta,g) = \mathbb{P}_N\{ g(x)^{-1} d(\beta) d(\beta)^T \Psi_1(\epsilon(\beta) g(x)^{-1/2}) \} \quad ;$$
$$B_N(\beta,g) = \mathbb{P}_N\{ g(x)^{-1/2} d(\beta) \Psi(\epsilon(\beta) g(x)^{-1/2}) \} \quad .$$

Bickel (1975) calls this a one step estimate of Type 1. Our technical arguments are for the Type 1 estimates, but they are easily modified to apply to his Type 2 estimates. If $\Psi(w) = w$, then (2.1) is a generalized least squares estimate with weights the inverse of $\hat{g}(x)$. Ordinarily, one first computes $\hat{\beta}$ based on unweighted least squares residuals, updates the preliminary estimator and recomputes $\hat{\beta}$. This process will be repeated a few times.

Let $\eta_N = N^{-\gamma}$ for a sufficiently small positive γ. Let $\hat{g}(x)$ equal η_N plus an estimate of $v_0(x)$ based on the residuals $\epsilon_i(\hat{\beta}_*)$. The addition of the small amount η_N avoids problems with infinite weights. Let $\hat{v}_0(x)$ be an estimate of $v_0(x)$ based on the true unknown errors $\epsilon_i(\beta)$, and let $\hat{v}(x) = \hat{v}_0(x) + \eta_N$. In this and the next section, rather than adding η_N to the smoothers, without change we could have taken the maximum of η_N and the smoother. The key conditions about these estimates are as follows.

(A.1) $$\text{infimum } v_0(x) > 0 \quad .$$

(A.2) $$N^{\alpha(1)-\delta}\, \mathbb{P}_N\{\, \hat{v}_0(x) - v_0(x)\,\}^2 \xrightarrow{p} 0 \text{ for all } \delta > 0 \ ,$$
$$\text{where } \alpha(1) = 4 \,/\, (4 + p) \ .$$

(A.3) $$N^{\alpha(2)-\delta}\, \mathbb{P}_N\{\, \hat{g}(x) - \hat{v}_0(x)\,\}^2 \xrightarrow{p} 0 \text{ for all } \delta > 0 \ ,$$
$$\text{where } \alpha(2) \geq \alpha(1) \ .$$

Assumption (A.2) holds for kernel and nearest neighbor estimates under appropriate conditions, see section 4. The constant $\alpha(1)$ is the optimal rate of convergence for nonparametric regression estimates, although slower rates could have been used. We are not restricting to smoothing only squared residuals.

Condition (A.3) is often easy to check. For linear smoothers such as kernel and nearest neighbor estimates having "bandwidth" not depending on the responses, $\alpha(2) \geq 1 > \alpha(1)$, at least under minimal regularity conditions. In some instances, $\alpha(2) > 1$. It is easy to develop conditions for these to hold in the case of a linear smoother applied to squared residuals, but in the interest of space we forego the details.

If x has compact support, then under additional regularity conditions the convergence in (A.2) - (A.3) is uniform in x, e.g., in (A.2)

$$N^{\alpha(1)-\delta} \sup_x \left| \hat{v}_0(x) - v_0(x) \right|^2 \xrightarrow{p} 0 \ .$$

See section 4 for a further discussion. The regularity conditions on Ψ and f would then be constructed as in Bickel (1975), and we could take $\eta_N = 0$. In particular, Ψ would not need to be differentiable. Where possible, we wish to avoid the assumption of compactness. If the x's are not confined to a compact set, the convergence of $\hat{g}$ and $\hat{v}_0$ is not uniform and we must assume more

smoothness in Ψ. The following conditions are reasonable, most of them trivial for linear regression. We use the notation $\| \cdot \|$ to be the Euclidean norm of the argument.

(A.4) Ψ_1 exists and is continuous and for $M_\Psi^{(1)}$, $M_\Psi^{(2)} \le \infty$, $c_\Psi < \infty$

$$| \Psi(a) - \Psi(b) | \le c_\Psi \text{ minimum}(M_\Psi^{(1)} , | a - b |) ;$$

$$| \Psi_1(a) - \Psi_1(b) | \le c_\Psi \text{ minimum}(M_\Psi^{(2)} , | a - b |) .$$

(A.5) $$\mathbb{P}\{ \Psi_1(e(\beta)) \}^2 < \infty .$$

(A.6) $$\mathbb{P}\{ \| d(\beta) \| \Psi(e(\beta)) \}^2 < \infty .$$

(A.7) $$N^{-1/2} \mathbb{P}\left[\sup\{ \Psi_1^{\,2}(\epsilon(\beta+\Delta_1) / \Delta_2^{\,1/2}) \} \;\middle|\; \begin{array}{l} \| \Delta_1 \| \le MN^{-1/2} \\ | \Delta_2 | \ge \eta_N \end{array} \right] = O(1),$$

for all $M > 0$.

(A.8) $$N^{k/2} \mathbb{P}\left[\sup\{ \| d(\beta+\Delta_1) \|^2 \| d(\beta+\Delta_1) - d(\beta+\Delta_2) \|^k \} \;\middle|\; \begin{array}{l} \| \Delta_1 \| \le MN^{-1/2} \\ | \Delta_2 \| \le MN^{-1/2} \end{array} \right]$$

$= O(1)$ for all $M > 0$ and $k = 1,2$.

(A.9) $$\mathbb{P}[\| f_{\beta\beta}(x,\beta) \|^4 \{1 + |\Psi(e(\beta))| + |e(\beta) I(M_\Psi^{(1)} = \infty)|\}] < \infty .$$

(A.10) If $G(\Delta_1,\Delta_2) = d(\beta+\Delta_1)\, d(\beta+\Delta_1)^T \Psi_1\{\epsilon(\beta+\Delta_1)/\Delta_2^{\,1/2}\}$, then

$$\eta_N^{-1} \mathbb{P}\left[\sup\{ \| G(\Delta_1,\Delta_2) - G(0,\Delta_2) \| \} \;\middle|\; \begin{array}{l} \|\Delta_1\| \le MN^{-1/2} \\ \|\Delta_2\| \ge \eta_N^{1/2} \end{array} \right] \to\to\to 0 \text{ for all } M > 0.$$

(A.11) $$N^{-2\gamma} \mathbb{P}\left[\{\sup \| f_\beta(x,\beta) \|^{2k} s^{2j}\left(\frac{\epsilon(\beta)}{\Delta}\right) \} \;\middle|\; \Delta^2 \ge \eta_N \right] \to\to\to 0 ,$$

for $(k=1,2 : j=0,1 : s=\Psi_1)$ or $(k=1, j=1, s=\Psi)$.

(A.12) Write $H_{k,j,n}(\epsilon,x) = \epsilon(\beta)^j \| \dfrac{\partial^k}{\partial\beta^k} f(x,\beta) \|^{2n} / v_0(x)^n$. Then if $M_\Psi^{(2)} = \infty$ in (A.4), $\mathbb{P}\{ H_{k,j,n}(\epsilon,x) \} < \infty$ for $n=2, j=2, k=1$

if $M_{\Psi}^{(2)} < \infty$ in (A.4), $\mathbb{P}\{ H_{k,j,n}(\epsilon,x) \} < \infty$ for n=2,k=1,j=0.

(A.13) $$N^{-1/4}\, \mathbb{P}\left[\sup\left[\| f_{\beta\beta\beta}(x,\beta+\Delta_1) \|\, \Psi\{ \frac{\epsilon(\beta)}{\Delta_2} \} \right]^2 \,\middle|\, \begin{array}{l} \|\Delta_1\| \le MN^{-1/2} \\ |\Delta_2|^2 \ge \eta_N \end{array} \right] \longrightarrow 0$$

For linear regression and generalized least squares, (A.4) - (A.13) simplify to $\| x \|^4$, $\epsilon^4(\beta)$ and $\| x\, v_0(x) \|^2$ having finite expectations. The first step is a preliminary expansion for β.

LEMMA 1 : Define

$$C_N = N^{1/2}\, \mathbb{P}_N\left[\hat{g}(x)^{-1/2}\, d(\beta)\, \Psi(\hat{g}(x)^{-1/2}\, \epsilon(\beta)) \right] .$$

Suppose that $C_N = O_p(1)$. Then under the conditions (1.4), (1.6) and (A.1) - (A.13), we have that

(2.2) $$N^{1/2}(\hat{\beta} - \beta) - W_2^{-1}\, C_N \xrightarrow{p} 0.$$ □

In principle, we can obtain the asymptotic distribution of $\hat{\beta}$ by direct methods using (2.2). However, sometimes it is possible to replace $\hat{g}$ in (2.2) by $\hat{v}$, the estimator of v_0 based only on the errors $\epsilon_i(\beta)$.

THEOREM 1 : Suppose $\alpha(2) > 1.0$ in assumption (A.3). Define

$$D_N = N^{1/2}\, \mathbb{P}_N\left[\hat{v}(x)^{-1/2}\, d(\beta)\, \Psi(\hat{v}(x)^{-1/2}\, \epsilon(\beta)) \right] .$$

Suppose that $D_N = O_p(1)$, and that either of the following hold:

(B.1) $$\mathbb{P}\{ \epsilon^2(\beta)\, \| d(\beta) \|^2 \} < \infty .$$

(B.2) For some $q(1)$, $q(2)$,

$$N^{q(1)} \sup_x \left| \hat{v}_0(x) - v_0(x) \right| \xrightarrow{p} 0 \; ; \; \mathbb{P}\{ \|d(\beta)\|^2 / v_0(x) \} < \infty \; ;$$

$$N^{q(2)} \sup_x \left| \hat{g}(x) - \hat{v}(x) \right| \xrightarrow{p} 0 \; ; \quad \text{and}$$

$$N \, \mathbb{P}\left[\sup\left[\left(\frac{v_0(x)}{v_0(x)+\Delta_1} \right)^2 \left| H(\Delta_1,\Delta_2) - H(\Delta_1,0) \right|^2 \;\middle|\; \begin{matrix} \|\Delta_1\| \le MN^{-q(1)} \\ \|\Delta_2\| \le MN^{-q(2)} \end{matrix} \right] \right] \to 0 \; ,$$

where $H(\Delta_1,\Delta_2) = \Psi\left(\dfrac{\epsilon(\beta)}{(v_0(x) + \Delta_1 + \Delta_2)^{1/2}} \right)$.

Then, under the conditions of LEMMA 1,

(2.3) $$N^{1/2}(\hat{\beta} - \beta) - W_2^{-1} D_N \xrightarrow{p} 0.$$ □

Our device of adding a constant η_N to the nonparametric regression estimator of v_0 can be eliminated in certain circumstances.

COROLLARY 1 : In the definition of $\hat{g}$, do not add the positive constant η_N. Suppose that $\{\hat{g} - \hat{v}_0\}$ and $\{\hat{v}_0 - v_0\}$ converge in probability to zero uniformly at the rate $N^{-\gamma}$. Then the conclusion to THEOREM 1 still holds. □

SECTION 3 : M-estimators in the Symmetric Case

If, given x_i, $\epsilon(x_i,\beta) = y_i - f(x_i,\beta)$ is symmetrically distributed about zero, then M-estimators adapt for heteroscedasticity, i.e., they have the same distribution as if the scaling function v_0 were known. In the previous section

we assumed only that the estimator $\hat{v}_0$ was a function of the errors $\epsilon(x_i,\beta)$. In this section we make the further assumption that $\hat{v}_0$ is an even function of these errors. In effect, when estimating the scaling function v_0, we are restricting to nonparametric regression estimators which are functions of the absolute residuals from the current fit $\hat{\beta}_*$, as did Carroll (1982) and Robinson (1986). It makes perfectly good sense to use the residuals to gain understanding of the variance structure.

Assumption (C.2) below is typically even weaker than (A.2), because in proving the latter one typically shows that the expectation of the left hand side of (A.2) converges to zero.

THEOREM 2 : Assume

(C.1) $$\mathbb{P}\{ \|d(\beta)\|^2 [1 + \|d(\beta)\|^2 \Psi^4(\epsilon(\beta)/v_0(x)^{\frac{1}{2}}) / v_0^2(x)] \} < \infty .$$

(C.2) $$\eta_N^{-6} E[\mathbb{P}_N\{ | \hat{v}_0(x) - v_0(x) |^2 \}] \rightarrow\rightarrow 0 .$$

(C.3) $$\text{If } M_\Psi^{(1)} = \infty \text{ in (A.4), then } \mathbb{P}\{ \epsilon^4(\beta) / v_0^2(x) \} < \infty .$$

Further assume (1.4) and the conclusion to THEOREM 1. Then the estimators (2.1) adapt for heteroscedasticity and have the same limit distribution (1.7) as if the scaling function v_0 were known.

□

For asymmetric errors when Ψ is not the identity function, THEOREM 2 typically fails. As we indicate below, it is possible to compute the limit distribution in this case, although we will not pursue the matter in any detail. To see what happens, consider a problem of dimension p = 1. By a Taylor series and (A.2), under sufficient regularity conditions, setting $\eta_N = 0$ for convenience, we have

$$(3.1)\qquad D_N = N^{1/2}\,\mathbb{P}_N\{\, d(\beta)\, v_0(x)^{-\frac{1}{2}}\, \Psi(\, \epsilon(\beta) \,/\, v_0(x)^{\frac{1}{2}}\,)\,\}$$
$$+\, N^{1/2}\,\mathbb{P}_N\{\, G(\epsilon,\beta,x)\,[\,\hat{v}_0(x) - v_0(x)\,]\,\} + o_p(1)\ ,$$
$$G(\epsilon,\beta,x) = (\partial/\partial w)\,\{\, w^{-1/2}\, d(\beta)\, \Psi(\epsilon(\beta)\,/\,w^{1/2})\,\}\,\Big|_{w=v_0(x)}\ .$$

As indicated in the next section, we can replace $G(\epsilon,\beta,x)$ by its conditional expectation given x, which is

$$G_*(x) = -(1/2)\, v_0(x)^{-2}\, d(\beta)\, E\,[\,\epsilon(\beta)\,\Psi_1\{\epsilon(\beta)/v_0(x)^{1/2}\}\,|\,x\,]\ .$$

The new second component of (3.1) has a nontrivial limit distribution. Thus, the limit distribution of $\hat{\beta}$ typically will not be the same as if v_0 were known.

SECTION 4 : Adaptation for Generalized Least Squares

If $\Psi(w) = w$, the estimator (1.7) is a generalized least squares estimator. Under weak conditions, Robinson (1986) proves adaptation in linear regression using a variant of the nearest neighbor device, the smoother being applied to squared residuals. His $\hat{g}(x_i)$ does not use the ith residual $y_i - f(x_i,\hat{\beta}_*)$. His proof is easily extended to nonlinear regression, as we now indicate.

First consider linear smoothers based on squared residuals:

$$(4.1)\qquad \hat{v}_0(x) = N^{-1}\sum_{i=1}^{N} c_N(x_i,x)\,\epsilon_i^2(\beta)\ ;\ \hat{g}(x) = N^{-1}\sum_{i=1}^{N} c_N(x_i,x)\,\epsilon_i^2(\hat{\beta}_*) + \eta_N,$$

$$(4.2)\qquad N^{-1}\sum_{i=1}^{N} c_N(x_i,x) = 1 \quad \text{for all } x\ .$$

Define the expectation of $\hat{v}_0$ given the design as

$$(4.3) \qquad v_*(x) = N^{-1} \sum_{i=1}^{N} c_N(x_i,x)\ E\{\epsilon_i^2(\beta) \mid x_i\}\ .$$

LEMMA 1 and **THEOREM 1** still hold if we replace $v_0(x)$ in (A.2) by $v_*(x)$. The modified (A.2) is easy to verify under weak conditions. Following Robinson (1986), choose the weight function $c_N(x,z)$ to be of nearest neighbor type with $c_N(x,x) = 0$. **THEOREM 1** eliminates the nonlinear regression function, and we can apply Robinson's proof virtually without change, the only complication being the addition of η_N.

Here is a second application of Theorem 1. Suppose the support of x is compact. If the variance is to be modeled nonparametrically as a function of p $\leq$ 3 predictors, then we can obtain an adaptation result based on the usual nearest neighbor regression estimators which use the ith residual in computing an estimate of $v_0(x)$, i.e., $c_N(x,x) \neq 0$. For reasons of space and interest we forego the details.

Some background is useful. Uniform convergence of nonparametric regression estimators when x has compact support has been proved for kernel estimators by Collomb & Haerdle (1986) and for nearest neighbor estimators by Devoyre (1978). Applying these results to our problem does not yield the weakest conditions, as the results assume that the marginal distribution of x is very smooth.

Average mean squared error convergence as in (A.2) have been discussed by Marron & Haerdle (1986) for kernel estimators with x being compactly supported and having a smooth distribution. Results of Mack (1981) can be used to prove (A.2) for nearest neighbor estimators under similar conditions. For nearest neighbor estimators it is far easier to prove (A.2) with v_0 replaced by v_* and

then note that (D.5) is the same as Robinson's Lemma 8, which by the way is a powerful result.

Here is a third application of THEOREM 1. One difficulty with using squared residuals to estimate the scaling function is the tendency for a few wildly aberrent values to distort the picture, see for example Figure 2 in Hinkley (1985). We have found it more useful to smooth absolute rather than squared residuals, a robust smoother being even better. Adaptation holds when absolute rather than squared residuals are smoothed. We give one set of strong conditions. Set $c_N(x,x) = 0$ and define

$$s_i(\beta) = s(x_i,\beta) = \mathbb{P}\{\ |\epsilon(\beta)|\ \mid x_i\}\ ;\quad v_0(x) = s^2(x,\beta)\ .$$

Consider the smoothers

$$\hat{g}(x)^{\frac{1}{2}} = N^{-1}\sum_{i=1}^{N} c_N(x_i,x)|\epsilon_i(\hat{\beta}_*)| + \eta_N\ ;\ \hat{v}(x)^{\frac{1}{2}} = N^{-1}\sum_{i=1}^{N} c_N(x_i,x)|\epsilon_i(\beta)| + \eta_N;$$

$$v_*(x)^{\frac{1}{2}} = N^{-1}\sum_{i=1}^{N} c_N(x_i,x)\ s(x_i,\beta) + \eta_N\ .$$

THEOREM 3 Consider linear regression in which the support of x is bounded. Since this is linear regression β has dimension p. Assume that the fourth moment of $\epsilon(\beta)$ given x is bounded and that $s(x,\beta)$ is strictly bounded away from zero and ∞. Suppose further that (A.4) through (A.13) hold and that

(D.1) $$\mathbb{P}\ [\mathbb{P}_N\{\ |v_*(x)^{1/2} - v_0(x)^{1/2}|^2\ \}\ ] \to\to 0\ .$$

(D.2) $\hat{\beta}_*$ is restricted to the set of values $(kN^{1/2})^{-1}[i_1,\ldots i_p]$, where $i_1,\ldots i_p$ are integers .

(D.3) $\hat{v}(x)^{1/2} - v_*(x)^{1/2}$ satisfies (A.2)

(D.4) $$\eta_N^{-4}\ N^{\frac{1}{2}}\ \mathbb{P}_N\{\ |\epsilon(\beta)|\ |\hat{v}(x)^{1/2} - v_*(x)^{1/2}|^2\ \} \xrightarrow{p} 0\ .$$

(D.5) $$N^{\frac{1}{2}}\ \mathbb{P}_N\{\ d(\beta)\epsilon(\beta)v_*(x)^{-\frac{1}{2}}\ [\hat{v}(x)^{-\frac{1}{2}} - v_*(x)^{-\frac{1}{2}}]\ \} \xrightarrow{p} 0\ .$$

Then then limit distribution of generalized least squares is the same as if $\hat{g}$ were replaced by v_0. □

(D.1) is the same as Robinson's key and powerful Lemma 8. (D.4) holds if the variance is a function of three or fewer predictors. Assumption (D.5) is essentially Robinson's key Proposition 2, substituting our $|\epsilon(\beta)|$ for his $\epsilon^2(\beta)$ and our $d(\beta)/v_*(x)^{\frac{1}{2}}$ for his $d(\beta)$.

SECTION 5 : Semiparametric Models and Information Bounds

An alternative approach is optimal estimation in semiparametric models, see Begun, et. al. (1983) and Bickel (1982). The easiest context for applying this theory is the location-scale model

$$e_i(\beta) = \{ y_i - f(x_i,\beta) \} / v_0^{1/2}(x_i) ,$$

Here, given x_i, the $\{e_i\}$ are independent and identically distributed random variables with a known density function $\omega(\cdot)$, but the scaling function $v_0(\cdot)$ is unknown. Our previous results do not require that the $e_i(\beta)$ be independent and identically distributed, conditionally on the x_i. The usual approach in the semiparametric literature is to find out how well one can do estimating $\beta$when v_0 is unknown, i.e., find a matrix $I(\omega)$ such that

$$N^{1/2} (\hat{\beta} - \beta) \Rightarrow \text{Normal}(0,S) \text{ implies } S \geq I(\omega)^{-1} .$$

The matrix $I(\omega)$ is called the semiparametric information bound. Any estimator achieving this bound is said to be asymptotically efficient.

We can use the results of section 3 to construct asymptotically efficient estimators for β. Suppose $\omega(\cdot)$ is a symmetric density. Let

$$\Psi(w) = - (\partial/\partial w) \log\{ h(w) \} .$$

Suppose further that Ψ, the regression function f and the nonparametric regression estimators $(\hat{g},\hat{v}_0)$ satisfy (A.1) - (A.13), one of (B.1) - (B.2) with $\alpha(2) > 1.0$, and (C.1) - (C.3). Then our estimators (2.1) are asymptotically efficient in the semiparametric sense, because they have the same limit distribution as maximum likelihood with v_0 known. As an added benefit of THEOREM 2, even if the errors are incorrectly specified but still symmetric, our estimators will behave as if v_0 were known.

That the information bound is the same as if v_0 were known is an easy informal calculation using the theory of Begun, et. al. (1983). To apply their theory, set $v_0(x) = \sigma^2 \; v_*(x)$, where v_* is a density with respect to a dominating measure. The calculations are immediate.

It is also clear that if one is willing to assume symmetric, independent and identically distributed errors, then the information bound does not change even if the density $\omega(\cdot)$ is unknown. It should be possible to construct asymptotically efficient estimators in this case. This program has already been carried out by Bickel (1982) for homoscedastic regression. One can also comtemplate $\omega(\cdot)$ being known and asymmetric or unknown and possibly asymmetric, but we leave this problem to others.

SECTION 6 : Variance a Function of the Mean

Often, the variance can be modeled as a function of the mean response. Assume that the data are normally distributed with mean $f(x,\beta)$ and variance modeled parametrically as $\sigma^2 v_{mp}(f(x,\beta),\theta)$. It is well known that generalized least squares estimates of β have the same limit distribution (1.7) as weighted least squares with known weights. Jobson & Fuller (1980) showed that the normal theory maximum likelihood estimate of β is, at the normal model, asymptotically more efficient than generalized least squares. However, the maximum likel estimate has some disadvantages. Generalized least squares has the robustness property that its asymptotic distribution is not dependent on assumptions about higher moments. McCullagh (1983) shows that this is not the case for the maximum likelihood estimate. It does not take a particularly nasty distibution before the maximum likelihood estimate becomes less efficient than generalized least squares. Carroll & Ruppert (1982b) note that the influence function of the maximum likelihood estimate is quadratic in the errors, and that unlike generalized least squares, the maximum likelihood estimate it not robust to a small misspecification in the variance function. We have seen no real data examples where even the potential gain in efficiency for using maximum likelihood is over 30%, and it our belief that "asymptopia" occurs for the maximum likelihood estimate only for much larger sample sizes than necessary for generalized least squares.

With this background in mind, let us turn to the semiparametric location-scale model with scale depending on the regression parameter. Referring for notation to (1.3), assume that the errors satisfy

$$(6.1) \qquad e_i = \frac{\epsilon_i(\beta)}{v_0^{1/2}(x_i)} = \frac{\epsilon_i(\beta)}{v_m^{1/2}\{f(x_i,\beta)\}} \quad ,$$

where, given x_i, the $\{e_i\}$ are independent and identically distributed. Each of the ways of writing the errors (6.1) has an interesting interpretation. The first indicates that the variances are not constant, so that we could apply the techniques of sections 2-4. This is almost certain to be inefficient if the dimension of x_i is of any size. The second way of writing the model suggests a variant of generalized least squares. Take the squared residuals from the current fit and regress them nonparametrically on the predicted values from the current fit. For symmetric errors, linear regression and with stringent regularity conditions, Carroll (1982) showed that this form of generalized least squares has the limit distribution (1.7). It would be interesting to improve upon Carroll's result, but we leave this to another time. The second form of (6.1) also suggests a semiparametric model. We consider only the case that the errors (6.1) have known symmetric density $\omega(\cdot)$.

As in the parametric case, it does not follow that, having found an asymptotically efficient semiparametric estimator, one should use it. Our calculations outlined below indicate that for normally distributed data the efficient semiparametric estimator will suffer the drawbacks that the maximum likelihood estimate does in the parametric case. The efficient influence function is quadratic and hence the estimators will be sensitive to nonnormality. We conjecture that there is an analogue to the Carroll & Ruppert (1982b) result, i.e., the estimators will also be affected by small misspecifications of the variance model. It is not clear that it is often the case that the increase in efficiency at the model can be achieved in reasonable size samples and for realistic problems.

The calculations we give here are informal, meant to indicate the form of the efficient influence function. We assume that the mean function is that of the linear model. Define the auxilliary functions

$$\mu = x^T\beta \ ; \ q(\mu) = (1/2)\,\frac{\partial}{\partial\,\mu}\,\ln\{\ v_m(\mu)\ \} \ ; \ t(\mu) = E\{\ x \mid x^T\beta=\mu\ \} \ ;$$
$$SC(e(\beta)) = \frac{\partial}{\partial\,t}\,\ln\{\ \omega(t=e(\beta))\ \} \ .$$

$SC(\cdot)$ is the usual score function in the univariate case. The derivatives of the loglikelihood with respect to β and σ are

$$(6.2) \qquad \ell_\beta = -q(m)\ x\ \{1 + e(\beta)\ SC(e(\beta))\} - x\ SC(e(\beta))\ /\ \{\sigma^2 v_{m*}(\mu)\}^{1/2} \ .$$

$$(6.3) \qquad \ell_\sigma = -\sigma^{-1}\ \{1 + e(\beta)\ SC(e(\beta))\ \} \ .$$

Using the notation of Begun, et. al. (1983), if the density function of the data is f_d, the local derivative of the function v_{m*} is

$$(6.4) \qquad 2(A_1 b_1)/f_d^{1/2} = -b_1(\mu)\ \{1 + e(\beta)\ SC(e(\beta))\}\ /\ v_{m*}(\mu)^{1/2} \ ;$$

This local derivative depends on the data only through μ and $|e(\beta)|$. As in Begun, et. al. (1983) or Bickel & Ritov (1986), the efficient score function is defined by

$$(6.5) \qquad [\ \ell_\beta - E\{\ell_\beta|\mu,|e(\beta)|\}\ ] - d\ [\ \ell_\sigma - E\{\ell_\sigma|\mu,|e(\beta)|\}\ ] \ .$$

Here d is to be chosen so that the expectation of the product of (6.5) and (6.4) is zero. Since ℓ_σ is itself a function only of μ and $|e(\beta)|$, it suffices to set d = 0. Thus, the efficient score function is

$$(6.6) \qquad \overline{\ell}_{\beta\sigma} = -q(m)\ \{x - E[x|\mu]\}\ \{1 + e(\beta)\ SC(e(\beta))\} - x\ SC(e(\beta))\ /\ \{\sigma^2 v_{m*}(\mu)\}^{1/2} .$$

The information bound $I(\omega)$ is the covariance matrix of $\overline{\ell}_{\beta\sigma}$, which is easy to compute because the two components of (6.6) are uncorrelated by symmetry. The covariance matrix of the second term in (6.6) is the information bound of section 5, where one did not know that the variance was a function of the mean. This means that there is extra information available when one knows that the variance is a function of the mean response. The only exception is when $x = E[x|\mu]$. Since the efficient semiparametric score $\overline{\ell}_{\beta\sigma}$ differs from the usual score (6.2), this is a case where there is a loss of asymptotic efficiency for not knowing the form of the variance function.

For normally distributed errors, (6.6) shows that the efficient semiparametric score function is quadratic in the errors, not linear as is the case for generalized least squares. The same thing happens in the parametric case, and the same concerns we have raised previously apply. The possible gains in efficiency are less than in the parametric case. It is not clear to us that when one is assuming $h(\cdot)$ is known, one should try to extract the extra information in the data.

We can carry out the same informal calculations when the density of the errors $\omega(\cdot)$ is symmetric and the density of the design $s(\cdot)$ is unknown. The local derivatives for h and s are (respectively)

$$2(A_2 b_2)/f_d^{1/2} = 2\ b_2(e(\beta))\ /\ h^{1/2}(e(\beta))\ ;$$

$$2(A_3 b_3)/f_d^{1/2} = 2\ b_3(x)\ /\ s^{1/2}(x)\ .$$

These are orthogonal to (6.6), so the efficient score function and the information bound are the same as if h and s were known. While one can

construct uniformly efficient estimates of β when the errors are symmetric, it would be more interesting to know if such estimates are actually efficient in samples of small to moderate size.

REFERENCES

Begun, J. M., Hall, W. J., Huang, W. M. & Wellner, J. A. (1983). Information and asymptotic efficiency in parametric - nonparametric models. Annals of Statistics 11, 432-452.

Bickel, P. J. (1975). One-step Huber estimates in the linear model. Journal of the American Statistical Association 70, 428-436.

Bickel, P. J. (1982). On adaptive estimation. Annals of Statistics 10, 647-671.

Bickel, P. J. & Ritov, Y. (1986). Efficient estimation in the errors in variables model. Annals of Statistics 14, 000-000.

Box, G. E. P. & Myer, R. D. (1985b). Dispersion effects from fractional designs. Technometrics, 28, 19-28.

Carroll, R. J. (1982a). Adapting for heteroscedasticity in linear models, Annal of Statistics 10, 1224-1233.

Carroll, R. J., and Ruppert, D. (1982a). A comparison between maximum likelihood and generalized least squares in a heteroscedastic linear model. Journal of the American Statistical Association 77, 878-882.

Carroll, R. J., and Ruppert, D. (1982b). Robust estimation in heteroscedastic linear models, Annals of Statistics 10, 429-441.

Carroll, R. J., Ruppert, D., & Wu, C. F. J. (1986). Variance expansions and the bootstrap in generalized least squares. Preprint.

Collomb, G. & Haerdle, W. (1986). Strong uniform convergence rates in robust nonparametric time series analysis and prediction : kernel regression estimations from dependent observations. Stochastic Processes and Their Applications, to appear.

Davidian, M. & Carroll, R. J. (1986). An asymptotic theory of variance function estimators. Preprint.

Devoyre, L. (1978). Uniform convergence of nearest neighbor regression function estimators and their applications in optimiation. IEEE

Transactions on Information Theory, IT42, 142-151.

Hinkley, D. V. (1985). Transformation diagnostics for linear models. Biometrika 72, 487-496.

Jobson, J. D. & Fuller, W. A. (1980). Least squares estimation when the covariance matrix and parameter vector are functionally related. Journal of the American Statistical Association 75, 176-181.

Mack, Y. P. (1981). Local properties of K-NN regression estimators. SIAM Journal on Algebraic and Discrete Methods 2, 311-323.

Marron, J. S. & Haerdle, W. (1986). Random approximations to some measures of accuracy in nonparametric curve estimation. Journal of Multivariate Analysis, to appear.

Matloff, N., Rose, R. & Tai, R. (1984). A comparison of two methods for estimating optimal weights in regression analysis. Journal of Statistical Computation and Simulation 19,265-274.

McCullagh, P. (1983). Quasi-likelihood functions. Annals of Statistics 11, 59-67.

Muller, H. G. & Stadtmuller, U. (1986). Estimation of heteroscedasticity in regression analysis. Preprint.

Pollard, D. (1984). Convergence of Stochastic Processes. Springer-Verlag, New York.

Robinson, P. M. (1986). Asymptotically efficient estimation in the presence of heteroscedasticity of unknown form. Econometrica 56, 000-000.

Rothenberg, T. J. (1984). Approximate normality of generalized least squares estimates. Econometrica 52, 811-825.

Appendix : The proof of **LEMMA 1** is accomplished through a series of propositions. Let $h(\beta,v) = d(x,\beta)/v(x)^{\frac{1}{2}}$ and $r(\beta,v) = \epsilon(\beta)/v(x)^{\frac{1}{2}}$. Throughout, T_N will be used as a current random variable under discussion.

PROPOSITION 1 : As $N \to \infty$,

$$N^{\frac{1}{2}} \mathbb{P}_N\{ h(\hat{\beta}_*,\hat{g})\, T_N \} \xrightarrow{p} 0 \text{ , where}$$

$$T_N = \Psi\{r(\hat{\beta}_*,\hat{g})\} - \Psi\{r(\beta,\hat{g})\} - h(\hat{\beta}_*,\hat{g})^T \Psi_1(r(\hat{\beta}_*,\hat{g}))\ \{\hat{\beta}_* - \beta\}\ .$$

Proof : By a Taylor series in β, with $\hat{\beta}_p$ on the line connecting $\hat{\beta}_*$ and β, it suffices to show that if $N^{\frac{1}{2}}(\hat{\beta}_p-\beta) = O_p(1)$ then

$$\mathbb{P}_N\{h(\hat{\beta}_*,\hat{g})\ [\ h(\hat{\beta}_p,\hat{g})\ \Psi_1(r(\hat{\beta}_p,\hat{g}) - h(\hat{\beta}_*,\hat{g})\ \Psi_1(r(\hat{\beta}_*,\hat{g})\ ]\} \overset{p}{\longrightarrow} 0\ .$$

By assumption (A.4), it suffices to show that

$$(7.1)\qquad \mathbb{P}_N\{\hat{g}(x)^{-\frac{1}{2}}\ h(\hat{\beta}_*,\hat{g})\ \Psi_1\{r(\hat{\beta}_p,\hat{g})\}\ [\ d(\hat{\beta}_*) - d(\hat{\beta}_p)\ ]\} \overset{p}{\longrightarrow} 0\ ;$$

$$(7.2)\qquad \mathbb{P}_N(\hat{g}(x)^{-\frac{1}{2}}\ h(\hat{\beta}_*,\hat{g})\ h(\hat{\beta}_*,\hat{g})^T\ [\ d(\hat{\beta}_*) - d(\hat{\beta}_p)\ ]\} \overset{p}{\longrightarrow} 0\ .$$

Since $\hat{g}(x) \geq \eta_N$ and γ is small, (7.2) follows from (A.8). Applying Cauchy-Schwarz, (7.1) follows from (A.7), (A.8) and since γ is small. □

PROPOSITION 2 : Let $s(x,\epsilon)$ be any positive function. Then $T_N = \mathbb{P}_N\{s(x,\epsilon)\hat{g}(x)^{-2}\} = O_p(1)$ if $\mathbb{P}\{s^2(x,\epsilon)\} < \infty$.

Proof : Since $\mathbb{P}_N\{s(x,\epsilon)/v_0^2(x)\} = O_p(1)$, it suffices that

$$\mathbb{P}_N\{s(x,\epsilon)\ |\ \hat{g}(x)^{-2} - v_0(x)^{-2}\ |\ \} = O_p(1)\ .$$

Now if v_0^{-1} and v_0^{-2} are bounded by $c > 1$, then

$$|\hat{g}^{-2}(x) - v_0^{-2}(x)| \leq 2c^2|\hat{g}(x) - v_0(x)|\ /\ \eta_N^2\ .$$

It thus suffices to show that

$$\eta_N^{-2}\ \mathbb{P}_N\{s(x,\epsilon)\ |\hat{g}(x) - v_0(x)|\ \} = O_p(1).$$

Using Cauchy-Schwarz and assumptions (A.2) and (A.3) for γ sufficiently small, the result follows. □

PROPOSITION 3 : As $N \to \infty$,

$$N^{\frac{1}{2}} \mathbb{P}_N\{\hat{g}(x)^{-1} [d(\hat{\beta}_*) - d(\beta)] \Psi(r(\beta,\hat{g})\} \xrightarrow{p} 0 .$$

Proof : By a two term Taylor series of f_β done componentwise, it suffices that

$$\mathbb{P}_N\{ \hat{g}(x)^{-1} f_{\beta\beta}(x,\beta) \Psi(r(\beta,\hat{g})\} \xrightarrow{p} 0 \tag{7.3}$$

$$N^{-\frac{1}{2}} \mathbb{P}_N\{ \hat{g}(x)^{-1} f_{\beta\beta\beta}(x,\hat{\beta}_p) \Psi(r(\beta,\hat{g}) \} \xrightarrow{p} 0 . \tag{7.4}$$

Assumption (A.13) is sufficient to prove (7.4). For (7.3), it suffices that

$$\mathbb{P}_N\{ \hat{g}(x)^{-1} f_{\beta\beta}(x,\beta) [\Psi\{r(\beta,\hat{g})\} - \Psi\{r(\beta,v_0)\}]\} \xrightarrow{p} 0 \tag{7.5}$$

$$\mathbb{P}_N\{f_{\beta\beta}(x,\beta) \Psi\{r(\beta,v_0)\} [\hat{g}(x)-v_0(x)] [\hat{g}(x)v_0(x)]^{-1}\} \xrightarrow{p} 0 . \tag{7.6}$$

Because of (A.1), (A.9) and Cauchy-Schwarz. (7.6) follows if

$$\mathbb{P}_N\{ \hat{g}(x)^{-2} |\hat{g}(x) - v_0(x)|^2 \} \xrightarrow{p} 0 . \tag{7.7}$$

By adding and subtracting $\hat{v}(x) = \hat{v}_0(x) + \eta_N$ in the numerator of (7.7) and then applying (A.2) and (A.3), (7.7) holds as long as

$$\eta_N^2 \mathbb{P}_N\{\hat{g}(x)^{-2}\} \xrightarrow{p} 0 ,$$

which follows from **PROPOSITION 2** with $s(x,\epsilon) \equiv 1$. It thus suffices to prove (7.5). First consider the case that $M_\Psi^{(1)} = \infty$ in (A.4). Writing

$$\Delta_N(x) = |\hat{g}(x) - v_0(x)|^2 / \eta_N ,$$

the square of (7.5) is bounded by

$$\mathbb{P}_N\{\hat{g}(x)^{-2} \| f_{\beta\beta}(x,\beta) e(\beta) \|^2\} \mathbb{P}_N\{ \Delta_N(x) \} = C_N^{(1)} \times C_N^{(2)} .$$

By (A.9) and **PROPOSITION 2**, $C_N^{(1)} = O_p(1)$. By (A.2) - (A.3),

$$C_N^{(2)} \xrightarrow{p} 0 .$$

Thus, (7.5) holds if $M_\Psi^{(1)} = \infty$, and it suffices to consider $M_\Psi^{(1)} < \infty$. Let $\Psi_{Hc}(v) = \max(-c,\min(v,c))$ be the Huber Ψ function. **PROPOSITION 2**, (A.9) and Cauchy-Schwarz combine to show that it suffices that

$$\mathbb{P}_N\{ [\Psi\{r(\beta,\hat{g})\} - \Psi\{r(\beta,v_0)\}]^2 \} \xrightarrow{p} 0 .$$

Using (A.4), it suffices that for every $c < \infty$,

$$\mathbb{P}_N\{ \Psi_{Hc}\left[\frac{\epsilon^2(\beta) \mid \hat{g}(x) - v_0(x) \mid^2}{v_0(x)\, \hat{g}(x)\, [v_0(x)^{\frac{1}{2}} + \hat{g}(x)^{\frac{1}{2}}]^2} \right] \} \xrightarrow{p} 0 .$$

By (A.1), monotonicity of Ψ_{Hc} and the fact that $\hat{g}(x) \geq \eta_N$, it suffices that

$$\mathbb{P}_N[\Psi_{Hc}\{ \epsilon^2(\beta)\, \Delta_N(x) \}] \xrightarrow{p} 0 . \tag{7.8}$$

Fix $\kappa > 0$. Decompose the left hand side of (7.8) as

$$A_{1N} + A_{2N} = \mathbb{P}_N\{\Psi_{Hc}\{\epsilon^2(\beta)\, \Delta_N(x)\}\, [\, I(\epsilon^2(\beta) \geq \kappa) + I(\epsilon^2(\beta) < \kappa)\,]\} ;$$

$$A_{1N} \leq c\, \mathbb{P}_N\{ I(\epsilon^2(\beta) \geq \kappa) \} \xrightarrow{} c\, \mathbb{P}\{ I(\epsilon^2(\beta) \geq \kappa)\} ;$$

$$A_{2N} \leq c^2\, \mathbb{P}_N\{ \Delta_N(x) \} \xrightarrow{p} 0 .$$

If we let $N \to \infty$ and then $\kappa \to \infty$, (7.8) follows. □

PROPOSITION 4 : As $N \to \infty$,

$$T_N = \mathbb{P}_N\{ \left|\hat{g}(x)^{-1} - v_0(x)^{-1}\right|^2 \} = O_p(N^{-2\gamma}) .$$

Proof : By (A.1), for some constants c_1, c_2,

$$T_N \le c_1 \mathbb{P}_N\{ \hat{g}(x)^{-2} \, |\hat{g}(x) - v_0(x)|^2 \}$$
$$\le c_2 \mathbb{P}_N\{ \hat{g}(x)^{-2} \, [|\hat{g}(x) - \hat{v}(x)|^2 + |\hat{v}_0(x) - v_0(x)|^2 + \eta_N^2] \} .$$

The proof is completed by (A.2) - (A.3) and **PROPOSITION 2**. □

PROPOSITION 5 : As $N \to \infty$,

$$\mathbb{P}_N\{ h(\hat{\beta}_*, \hat{g}) \, h(\hat{\beta}_*, \hat{g})^T \Psi_1\{r(\hat{\beta}_*, \hat{g})\} \} \xrightarrow{p} W_2 .$$

Proof : By (A.10) and (1.6), it suffices that

$$\mathbb{P}_N\{ h(\beta, \hat{g}) h(\beta, \hat{g})^T \Psi_1(r(\beta, \hat{g})) - h(\beta, v_0) h(\beta, v_0)^T \Psi_1(r(\beta, v_0)) \} \xrightarrow{p} 0 .$$

By **PROPOSITION 4** and (A.11),

$$\mathbb{P}_N\{ d(\beta) \, d(\beta)^T \Psi_1(r(\beta, \hat{g})) \, [\hat{g}(x)^{-1} - v_0(x)^{-1}] \} \xrightarrow{p} 0,$$

so it suffices that

$$\mathbb{P}_N\{ h(\beta, v_0) h(\beta, v_0)^T \, [\Psi_1(r(\beta, \hat{g})) - \Psi_1(r(\beta, v_0)) \} \xrightarrow{p} 0 . \tag{7.9}$$

If $M_\Psi^{(2)} = \infty$ in (A.4), then by (A.12) and Cauchy-Schwarz it suffices that

$$\mathbb{P}_N\{ \, |\hat{g}(x)^{-\frac{1}{2}} - v_0(x)^{-\frac{1}{2}}|^2 \} \xrightarrow{p} 0 .$$

For a constant c, by (A.1) this last term can be bounded by

$$\mathbb{P}_N\{ \hat{g}(x)^{-1} \, |\hat{g}(x) - v_0(x)|^2 \} \le \eta_N^{-1} \mathbb{P}_N\{ \, |\hat{g}(x) - v_0(x)|^2 \} \xrightarrow{p} 0$$

by **PROPOSITION 4**. If $M_\Psi^{(2)} < \infty$ in (A.4), by (A.12) to prove (7.9) it suffices that for every $c < \infty$,

$$\mathbb{P}_N\{\ \Psi_{Hc}(\ \epsilon^2(\beta)\ \Delta_N(x)\ )\ \} \xrightarrow{p} 0\ .$$

This is (7.8), which was proved in PROPOSITION 3. □

Proof of Lemma 1 : Combine PROPOSITIONS 1-5. □

PROPOSITION 6 : As $N \to \infty$,

$$T_N = N^{\frac{1}{2}}\ \mathbb{P}_N\{\ \left[\ \frac{d(\beta)\Psi(r(\beta,\hat{g}))\ \{\hat{g}(x)^{\frac{1}{2}} - \hat{v}(x)^{\frac{1}{2}}\}}{\hat{g}(x)^{\frac{1}{2}}\quad \hat{v}(x)^{\frac{1}{2}}}\ \right]\ \} \xrightarrow{p} 0\ .$$

Proof : Under (B.1) or (B.2) with $\alpha(2) > 1.0$, applying Cauchy-Schwarz, (A.3) and choosing γ,δ small shows that it suffices that

$$\eta_N^2\ \mathbb{P}_N\{\ \|d(\beta)\|^2\ \Psi^2(r(\beta,\hat{g}))\ \} = O_p(1)\ . \tag{7.10}$$

This follows from (A.11). □

Proof of THEOREM 1 : By (1.4) and PROPOSITION 6, we must show that

$$T_N = N^{\frac{1}{2}}\ \mathbb{P}_N\{\ h(\beta,\hat{v})\ [\ \Psi(r(\beta,\hat{g})) - \Psi(r(\beta,\hat{v}))\ ]\ \} \xrightarrow{p} 0\ .$$

This is routine for (B.1) and (B.2). □

Proof of THEOREM 2: It suffices that $A_N^{(1)} \xrightarrow{p} 0$, $A_N^{(2)} \xrightarrow{p} 0$. where

$$A_N^{(1)} = N^{\frac{1}{2}}\ \mathbb{P}_N\{\ d(\beta)\Psi(\epsilon(\beta)/v_0(x)^{\frac{1}{2}})[\hat{v}(x)^{-\frac{1}{2}} - v_0(x)^{-\frac{1}{2}}]\ \}\ ,$$

$$A_N^{(2)} = N^{\frac{1}{2}} \mathbb{P}_N\{ \hat{v}(x)^{-\frac{1}{2}} d(\beta) [\Psi(\epsilon(\beta)/\hat{v}(x)^{\frac{1}{2}}) - \Psi(\epsilon(\beta)/v_0(x)^{\frac{1}{2}})] \} .$$

Recall that $\hat{v}(x) = \hat{v}_0(x) + \eta_N$, where $\hat{v}_0(x)$ is an even function of $\{\epsilon_i(\beta)\}$. Because Ψ is odd, $\{\epsilon_i(\beta)\}$ has a symmetric distribution and $\hat{v}_0$ is even, it follows that for $j=1,2$, $A_N^{(j)}$ is $N^{1/2}$ times an average of N mean zero uncorrelated (not independent) random variables. Thus,

$$\|\mathbb{P}A_N^{(1)}A_N^{(1)^T}\| \leq \mathbb{P}[\, \mathbb{P}_N\{\|d(\beta)\|^2 \Psi^2(\epsilon(\beta)/v_0(x)^{\frac{1}{2}}) \, |\hat{v}(x)^{-\frac{1}{2}} - v_0(x)^{-\frac{1}{2}}|^2\} \,] .$$

By (C.1) and Cauchy-Schwarz, $A_N^{(1)} \xrightarrow{p} 0$ as long as

$$(7.11) \qquad \mathbb{P}[\, \mathbb{P}_N\{ \, |\hat{v}(x)^{\frac{1}{2}} - v_0(x)^{\frac{1}{2}}|^2 \, / \, \hat{v}(x) \} \,] \to 0.$$

This follows routinely from (A.1) and (A.2). Because $A_N^{(2)}$ is $N^{\frac{1}{2}}$ times a sum of uncorrelated random variables, we have by (A.4) and (C.1) that $\mathbb{P}||A_N^{(2)}||^2 \to 0$ as long as

$$(7.12) \qquad \mathbb{P}[\, \mathbb{P}_N\{T_N\} \,] \xrightarrow{p} 0, \text{ where}$$

$$T_N = \hat{v}(x)^{-1} \min^2[\, M_\Psi^{(1)}, \, |\epsilon(\beta)| \, |\hat{v}(x)^{-\frac{1}{2}} - v_0(x)^{-\frac{1}{2}}| \,]$$

Fix $c > 0$ and write $T_N = T_N I(|\epsilon(\beta)|>c) + T_N I(|\epsilon(\beta)|\leq c) = T_N^{(1)} + T_N^{(2)}$. Because of (A.1) and (C.2)

$$\mathbb{P}[\, \mathbb{P}_N\{T_N^{(2)}\} \,] \leq c^2 \, \mathbb{P}[\, \mathbb{P}_N\{\hat{v}(x)^{-1}|\hat{v}(x)^{-\frac{1}{2}} - v_0(x)^{-\frac{1}{2}}|^2\} \,]$$
$$\leq c_1 \, \mathbb{P}[\, \mathbb{P}_N\{ \, \hat{v}(x)^{-2}|\hat{v}(x) - v_0(x)|^2\} \,] \leq \eta_N^2 \, c_1 \, \mathbb{P}[\, \mathbb{P}_N\{\hat{v}(x)^{-2}\} \,] + o(1).$$

But by (C.2) and mimicking the proof of Proposition 2,

$$(7.13) \qquad \mathbb{P}[\ \mathbb{P}_N\ \{\hat{v}(x)^{-2}\}\ ] < \infty\ .$$

We thus need only show that

$$(7.14) \qquad \lim_{c\to\infty}\lim_{N\to\infty} \mathbb{P}[\ \mathbb{P}_N\ \{T_N^{(1)}\}\ ] = 0.$$

If $M_\psi^{(1)} < \infty$, then by Cauchy-Schwarz,

$$\mathbb{P}[\ \mathbb{P}_N\{T_N^{(1)}\}\ ] \le [\mathbb{P}\{I(|\epsilon(\beta)|>c)\}]^{\frac{1}{2}}\ N^{-1}\sum_{i=1}^{N}[\mathbb{P}\{\hat{v}(x)^{-2}\}]^{\frac{1}{2}}$$

$$\le [\mathbb{P}\ \{I(|\epsilon(\beta)|>c)\}]^{\frac{1}{2}}\ \{\ \mathbb{P}[\ \mathbb{P}_N\{\hat{v}(x)^{-2}\ ]\ \}^{\frac{1}{2}}\ .$$

Invoking (7.13), equation (7.14) now follows. We thus need to verify (7.12) when $M_\psi^{(1)} = \infty$. By (C.2) - (C.3) and a proof similar to that of Proposition 2, we have that

$$\mathbb{P}[\ \mathbb{P}_N\{\epsilon^2(\beta)\ v_0(x)^{-1}\ \hat{v}(x)^{-2}\}\ ] = O(1),$$

so that (7.12) follows by applying Cauchy-Schwarz and (7.11). □

<u>PROOF OF THEOREM 3</u> : (Sketch) As in Bickel (1982, page 653), from (D.2) it suffices to prove all steps with $\hat{\beta}_*$ replaced by $\beta_* = \beta + \Delta_N/N^{\frac{1}{2}}$, where $\Delta_N \to \Delta_0$ is a deterministic sequence. Using compactness and the boundedness of v_*,

$$|\hat{g}(x) - \hat{v}(x)| \le c_1 N^{-\frac{1}{2}}\ \{1 + |\hat{v}(x)^{\frac{1}{2}} - v_*(x)^{\frac{1}{2}}|\ ,$$

so that (A.3) holds with $\alpha(2) \ge 1.0$ and v_0 replaced by v_*. This assures that Lemma 1 holds. To prove <u>THEOREM 1</u>, it suffices to prove <u>PROPOSITION 6</u>, as the other step is similar. Write T_N as in <u>PROPOSITION 6</u> and

$$S_N = N^{\frac{1}{2}}\ \mathbb{P}_N\{\ v_*(x)^{-3/2}\ d(\beta)\ \epsilon(\beta)\ [\hat{g}(x)^{\frac{1}{2}} - \hat{v}(x)^{\frac{1}{2}}]\ \}\ .$$

Then, since $d(\beta) = d(x,\beta) = x$ is bounded, for some $c > 0$,

$$|T_N - S_N| \le c\,\eta_N^{-4}\,N^{\frac{1}{2}}\,\mathbb{P}_N\{\,|\epsilon(\beta)|\;|\hat{g}(x)^{\frac{1}{2}} - \hat{v}(x)^{\frac{1}{2}}|^2\,\}$$

$$+ c\,\eta_N^{-4}\,N^{\frac{1}{2}}\,\mathbb{P}_N\{\,|\epsilon(\beta)|\;|\hat{g}(x)^{\frac{1}{2}} - \hat{v}(x)^{\frac{1}{2}}|\;|\hat{v}(x)^{\frac{1}{2}} - v_*(x)^{\frac{1}{2}}|\,\}\ .$$

Since we have replaced $\hat{\beta}_*$ by β_*.

$$|\hat{g}(x)^{\frac{1}{2}} - \hat{v}(x)^{\frac{1}{2}}| \le c_4\,N^{-\frac{1}{2}}\ ,$$

so that $|T_N - S_N| \xrightarrow{p} 0$ by (D.4) and (A.6). Thus, to prove an analogue to <u>THEOREM 1</u>, it thus suffices that $S_N \xrightarrow{p} 0$. Using Bickel's trick and the boundedness of $d(\beta) = d(x,\beta) = x$ and the conditional fourth moment of $\epsilon(\beta)$, one shows directly that the covariance of S_N converges to zero as desired. The next step in the proof of Theorem 3 is to show that

$$Q_N = N^{\frac{1}{2}}\,\mathbb{P}_N\{\,d(\beta)\,\epsilon(\beta)\,[\hat{v}(x)^{-1} - v_*(x)^{-1}]\,\} \xrightarrow{p} 0\ .$$

By algebra,

$$Q_N = C_N^{(1)} + C_N^{(2)},\ \text{where}$$

$$C_N^{(1)} = 2\,N^{\frac{1}{2}}\,\mathbb{P}_N\{\,d(\beta)\,\epsilon(\beta)\,v_*(x)^{-\frac{1}{2}}\,[\hat{v}(x)^{-\frac{1}{2}} - v_*(x)^{-\frac{1}{2}}]\,\}\ ;$$

$$C_N^{(2)} = N^{\frac{1}{2}}\,\mathbb{P}_N\{\,d(\beta)\,\epsilon(\beta)\,[\hat{v}(x)^{-\frac{1}{2}} - v_*(x)^{-\frac{1}{2}}]^2\,\}\ .$$

These two terms converge to zero by (D.5) and (D.4) respectively. We are finished once we show that

$$N^{\frac{1}{2}}\,\mathbb{P}_N\{\,d(\beta)\,\epsilon(\beta)\,[v_*(x)^{-1} - v_0(x)^{-1}]\,\} \xrightarrow{p} 0.$$

Note that since $s(x,\beta)$ is bounded away from zero, so too is $v_*(x)$. The proof is completed by noting that this last random variable is a mean zero random variable whose covariance converges to zero by (D.1).