A NOTE ON ADAPTING FOR HETEROSCEDASTICITY WHEN THE VARIANCES DEPEND ON THE MEAN

Raymond J. Carroll* and David Ruppert**

University of North Carolina

Leonard A. Stefanski

Cornell University

*Supported by Air Force Office of Scientific Research contract AFOSR F49620-82-C-0009. Work partially completed while visiting the Mathematics Research Center of the University of Wisconsin - Madison with the support of Army Research Office Contract DAAG29-80-C-0041.

**Supported by the National Science Foundation Grant MCS-87-3129.

ABSTRACT

We consider the normal-theory regression model when the error disturbances are heteroscedastic, i.e., have non-constant variances. We distinguish two cases: (i) predictor heteroscedasticity, where the variances depend on a function g of known quantities and (ii) mean heteroscedasticity, where the variances depend on a function g of the means. For the case where g is unknown, Carroll (1982) showed by construction that, in certain cases, it is possible to estimate the regression parameter asymptotically as well as if g were known and weighted least squares applied. We reconsider this problem from the information bound theory of Begun, Hall, Huang & Wellner (1983). For mean heteroscedasticity, we obtain a rather surprising result. If g were known in this case, Jobson & Fuller (1980) showed that the maximum likelihood estimate is asymptotically more efficient than weighted least squares with known weights. When g is unknown the full Jobson & Fuller improvements are not possible; however, we show that one can, in theory, attain asymptotically better performance than weighted least squares with known weights.

AMS (MOS) Subject Classification: Primary 62J05; Secondary 62F10

Key Words: Adaptation, linear models, nonconstant variances, weighted least squares, maximum likelihood

1. INTRODUCTION

We consider the following heteroscedastic linear regression model. The observed data are (Y_i, x_i) for $i = 1, \ldots, N$. Here Y is a scalar and x is a p-vector. For technical reasons, we will assume that (Y_i, x_i) are independent and identically distributed according to the model.

(1.1) Given x_i, Y_i is distributed with mean $x_i^T\theta$ variance $\sigma^2 Q(x_i,\theta)$ for some function Q. The density of $(Y_i - x_i^T\theta)(\sigma^2 Q(x_i,\theta))^{-1/2}$ is $h(\cdot)$, which is symmetric about zero and continuous.

(1.2) The $\{x_i\}$ are bounded, independent and identically distributed random variables possessing, except for a possible intercept term, an absolutely continuous density $s(\cdot)$ with respect to some sigma-finite measure μ.

The model (1.1) includes a wide variety of special cases, of which two are the most important. The first we shall call mean-heterogeneity, where the variance is a continuously differentiable function of the mean, i.e.,

(1.3) Mean-Heterogeneity: $\mathrm{Var}(Y_i | x_i) = \sigma^2 g(x_i^T\theta)$.

The second special case is predictor-heterogeneity, where the variance depends on known quantities through a continuously differentiable function g, i.e.,

(1.4) Predictor Heterogeneity: $\mathrm{Var}(Y_i | x_i) = \sigma^2 g(x_i)$.

In (1.3) and (1.4), g is a density function with respect to some σ-finite measure on the support of $x^T\theta$ and x, respectively.

It may appear odd that g is assumed to be a density. This was done so that the general theory of Begun et al. (1983) is immediately applicable. It should be relatively easy to extend their theory to include classes of nuisance parameters g which are not necessarily densities, and this extension would be natural in our setting. However, the extra work would lengthen this article without providing any substantially new insights. Since σ can be adjusted depending on g, in practice it would not be difficult to standardize all the functions g so that they are densities with respect to some measure.

The function $g(\cdot)$ in (1.3)-(1.4) is rarely known exactly, while the density $h(\cdot)$ in (1.1) is usually assumed to be that of a standard normal random variable. When $g(\cdot)$ is unknown except for a finite number of parameters, a huge literature can be employed to estimate θ. For mean-heterogeneity, see Pritchard, Downie & Bacon (1977), Jobson & Fuller (1980) and Carroll & Ruppert (1982a), among others. For predictor-heterogeneity, see Hildreth and Houck (1968), Carroll & Ruppert (1983), and Johansen (1983).

There is less literature on estimating θ when the function $g(\cdot)$ in (1.3) or (1.4) is unknown and must be estimated, see Carroll (1982) for a theoretical study and Matloff, Rose and Tai (1984) as well as an unpublished report by Cohen, Dalal and Tukey for empirical studies. Let $\hat{\theta}_W$ be the weighted least squares estimate with known weights in model (1.1), i.e.,

$$\hat{\theta}_W = \left(\sum_{i=1}^{N} x_i x_i^T / Q(x_i, z_i, \theta)\right)^{-1} \sum_{i=1}^{N} x_i Y_i / Q(x_i, z_i, \theta) \ .$$

Then, under regularity conditions,

(1.5) $$N^{1/2}(\hat\theta_W - \theta) \Rightarrow \text{Normal }(0,S_W), \text{ where}$$

$$S_W^{-1} = \underset{N\to\infty}{\text{plim}}\ \sigma^{-2}N^{-1}\sum_{i=1}^{N} x_i x_i^T/Q(x_i,z_i,\theta)\ .$$

We shall call $\hat\theta_W$ the optimal normal theory weighted least squares estimate to indicate that it is based on knowing the weights and assuming the $\{Y_i\}$ are normally distributed. Neither situation is likely to arise too often in practice.

For mean-heterogeneity (1.3), Carroll (1982) showed by construction that by using nonparametric kernel regression estimation of squared least squares residuals on least squares predicted values $x_i^T\hat\theta_L$, an estimate $\hat g_N(\cdot)$ of $g(\cdot)$ can be constructed with the following property. Let $\hat\theta_K$ be the weighted least squares estimate based on the estimated weights $1/\hat g_N(x_i^T\hat\theta_L)$. Then

(1.6) $$N^{1/2}(\hat\theta_k - \theta) \Rightarrow \text{Normal }(0,S_W)\ .$$

Comparing (1.5) with (1.6) we see that asymptotically one can do as well as the optimal normal theory weighted least squares estimate even if the variance function is completely unknown. Carroll (1982) also showed a similar result for a special class of predictor-heterogeneity models.

If $g(\cdot)$ were known either exactly or up to a finite number of parameters, and if the error density $h(\cdot)$ in (1.1) is the normal density, then one could consider the normal theory maximum likelihood estimate $\hat\theta_M$. For mean-heterogeneity, there is information about θ in the variances as well as the mean, and Jobson & Fuller (1980) were able to show that the maximum likelihood estimate of θ is asymptotically preferable to optimal weighted least squares. More precisely,

$$(1.7) \qquad N^{1/2}(\hat{\theta}_M - \theta) \Rightarrow \text{Normal } (0, S_M), \text{ where } S_M \leq S_W \, .$$

For reasons of robustness, we have some doubts as to whether maximum likelihood should be the method of choice for small samples, see Carroll & Ruppert (1982b).

In the predictor-heteroscedasticity model (1.4), with normally distributed observations, we thus know that it is possible in some circumstances to achieve the asymptotic performance of optimal weighted least squares even when the variance function $g(x)$ is completely unknown. One purpose of this note is to explore the generality of this phenomenon. In particular, if the symmetric error density $h(\cdot)$ in (1.1) is unknown as well as the variance function $g(\cdot)$ in (1.4), we show that the information available for estimating θ is the same as when $h(\cdot)$ and $g(\cdot)$ are completely known. We obtain our results by applying the theory of Begun, et al. (1983) to models (1.3) and (1.4).

For normally distributed observations in a mean-heteroscedasticity model, we have two asymptotic facts. First, it is possible to reproduce optimal weighted least squares even when the variance function $g(x^T\theta)$ in (1.3) is completely unknown. Second, if the form of $g(\cdot)$ is known up to parameters, at the normal model it is possible to improve upon optimal weighted least squares by using maximum likelihood. This leaves unanswered two interesting questions. First, if $g(\cdot)$ is unknown in model (1.3), is it possible to achieve the performance of maximum likelihood with known $g(\cdot)$? Using the theory of Begun, et al, we show that the answer to this question is no, which in retrospect is perhaps not surprising. Second, if $g(\cdot)$ is unknown in model (1.3), is it still possible to improve upon optimal weighted least squares? We obtain only a partial but perhaps surprising positive answer to this question. More precisely, we show that if $g(\cdot)$ in (1.3) and $s(\cdot)$ in (1.2) are smooth, then the information available for estimating θ is more than that provided by weighted least squares.

The paper is organized as follows. In Section 2 we outline the theory developed by Begun, et al. In Sections 3-5 we apply this theory to our problems, discussing in Section 6 the possibility of constructing estimators which achieve the relevant information bounds.

2. THE BEGUN, HALL, HUANG & WELLNER THEORY

Lower bounds for estimation in semiparametric models is an area undergoing considerable development. Suppose that $z_1, z_2, \ldots, z_N$ are independent and identically distributed random vectors possessing a density function $f(\cdot,\theta,\sigma,g)$ with respect to a sigma-finite measure μ. Here θ is a vector of parameters of interest, σ is a vector of nuisance parameters and $g = (g_1, g_2, g_3)$ are densities with respect to sigma finite measures ν_1, ν_2, ν_3 respectively. Begun, et al. (1983) provide upper bounds on the information available for estimating θ when (σ,g) is unknown. Informally, their major result can be summarized as follows. Let $\ell(\cdot,\theta,\sigma,g)$ be the logarithm of $f(\cdot,\theta,\sigma,g)$ and let ℓ_θ, ℓ_σ be the derivatives of the log-likelihood ℓ with respect to θ and σ, respectively. Define

$$B = E\ell_\theta \ell_\sigma^T, \quad D = E\ell_\sigma \ell_\sigma^T \quad \text{and} \quad \ell_{\theta,\sigma} = \ell_\theta - BD^{-1}\ell_\sigma .$$

Let $|\cdot|$ denote the Euclidean normal and $\|\cdot\|_{\nu_i}$ the $L^2_{\nu_i}$ norm.

Suppose that there are a bounded linear operators $A_i = L^2_{\nu_i} \to L^2_\mu$ for which

$$n^{1/2}|\theta_n - \theta_0| - h_1 \to 0$$

(2.1) $$n^{1/2}|\sigma_n - \sigma_0| - h_2 \to 0$$

$$\|n^{1/2}(g_{i,n}^{1/2} - g_i^{1/2}) - \beta_i\|_{\nu_i} \to 0 \ (i=1,2,3)$$

implies for $g_n = (g_{n1}, g_{n2}, g_{n3})$

$$E\{2\, n^{1/2}[f^{1/2}(\cdot,\theta_n,\sigma_n,g_n) - f^{1/2}(\cdot,\theta_o,\sigma_o,g)]/f^{1/2}(\cdot,\theta_o,\sigma_o,g)$$

$$- h_1 \ell_\theta - h_2 \ell_\sigma - 2 \sum_{k=1}^{3} (A_k\beta_k)/f^{1/2}(\cdot,\theta_o,\sigma_o,g)\}^2 \to 0.$$

When this holds, $f^{1/2}(\cdot,\theta,\sigma,g)$ is said to be Hellinger differentiable.

As discussed by Begun et al. (1983), ℓ_θ and ℓ_σ are the score functions for θ and σ and $\ell_{\theta,\sigma}$ is the effective score for θ when σ is a nuisance parameter but the g_i are known. Begun et al. have a small technical error in their remark 3.2 where they compute the "effective score for θ" in the presence a $\mathbf{R}^s$ valued nuisance parameter η, which corresponds here to σ, and a density nuisance parameter g. Briefly, the effective score for θ is the part of score for θ orthogonal to the subspace spanned by the score for η and the score for g. Begun et al. compute this score by finding $\rho_{\theta\cdot\eta}$, the score for θ in the presence of η, and then taking the part of $\rho_{\theta\cdot\eta}$ orthogonal to the space of scores for g. Although convenient, their computational method is correct only under the condition that the score for η is orthogonal to the score for g. In our notation this condition is equivalent to having

$$(2.2) \qquad E[(\ell_\sigma)(\sum_{k=1}^{3} A_k\beta_k/f^{1/2})] = 0$$

for all choices of β_k in B_k. For ease of computation we will choose the measure ν_k so that (2.2) holds.

To continue the Begun et al. method of computing the effective score for θ, if θ is p-dimensional, find p-vectors β_{*1}, β_{*2}, β_{*3} for which

$$(2.3) \qquad E\{(\ell_{\theta,\sigma} - 2\sum_{k=1}^{3} (A_k\beta_{*k}/f^{1/2})] \, [2 \sum_{k=1}^{3} A_k\beta_k/f^{1/2}]\} = 0$$

for all β_1, β_2, β_3 in appropriate sets of functions B_1, B_2, B_3. Technically, the sets B_k must be closed subspaces of $L^2(\nu_k)$ such that

$$(2.4) \qquad \int g_k^{1/2}\, \beta_k \, d\nu_k = 0 \quad \text{if} \quad \beta_k \in B_k .$$

Equation (2.4) would not be necessary if we dropped the requirment that g_k be a density with respect to ν_k. However, by a judicious choice of ν_k (2.4) implies (2.2), and this fact is at least a minor convenience. If β_{*k} can be computed and is an element of B_k, then the information bound is

$$(2.5) \qquad I_* = E\, \tilde{\ell}_\theta \, \tilde{\ell}_\theta^T ,$$

where

$$(2.6) \qquad \tilde{\ell}_\theta = \ell_{\theta,\sigma} - 2\sum_{k=1}^{3} A_k\beta_{*k}/f^{1/2} .$$

The function (2.6) is the efficient score.

In a sense made precise by Begun, et al., for any regular estimator of θ for which

$$(2.7) \qquad N^{1/2}(\hat{\theta}_N - \theta) \Rightarrow N(0, \ddagger) ,$$

the best one can hope to do is to have $\ddagger = I_*^{-1}$, i.e., we must have that in (2.7),

$$\ddagger \geq I_*^{-1}. \tag{2.8}$$

3. INFORMATION BOUNDS FOR PREDICTOR-HETEROSCEDASTICITY

Consider the model defined by (1.1), (1.2) and (1.4). We wish to estimate θ, with σ unknown. Also, the density $s(\cdot)$ of x is unknown, as is the symmetric density $h(\cdot)$ of $(Y-x^T\theta)/(\sigma g^{1/2}(x))$ and the variance function $g(\cdot)$. We will show that the information bound I_* in (2.4) is exactly the usual parametric information I_o computed with (g,h,s) all known. In the language of Begun, et al., this is a situation for which adaptation is possible.

The density function for the predictor-heterogeneity model is

$$f = [\sigma g^{1/2}(x)]^{-1}\, s(x)\, h\{(Y-x^T\theta)/(\sigma g^{1/2}(x))\}. \tag{3.1}$$

Writing $r = (Y-x^T\theta)/(\sigma g^{1/2}(x))$, it follows directly that

$$\ell_\theta = -\frac{x}{\sigma g^{1/2}(x)}\,\frac{\dot{h}(r)}{h(r)} \quad \text{and} \quad \ell_\sigma = -\frac{1}{\sigma}\left(1 + \frac{\dot{h}(r)r}{h(r)}\right), \tag{3.2}$$

where $\dot{h}(r)/h(r)$ is an odd function of r. Since $B = E\ell_\theta\ell_\sigma = 0$, we have $\ell_{\theta,\sigma} = \ell_\theta$. If we let

$$\|n^{1/2}(g_n^{1/2} - g^{1/2}) - \beta_1\|_{\nu_1} \to 0 \tag{3.3}$$

$$\|n^{1/2}(h_n^{1/2} - h^{1/2}) - \beta_2\|_{\nu_2} \to 0 \tag{3.4}$$

(3.5) $\| n^{1/2}(s_n^{1/2} - s^{1/2}) - \beta_3 \|_{\nu_3} \to 0$,

then one finds that

$$2(A_1\beta_1)/f^{1/2} = -\beta_1(x)\{r\, \dot{h}(r)/h(r) + 1\}/g^{1/2}(x) ,$$

$$2(A_2\beta_2)/f^{1/2} = 2\ \beta_2(r)/h^{1/2}(r)$$

$$2(A_3\beta_3)/f^{1/2} = 2\ \beta_3(x)/s^{1/2}(x) .$$

Since $r\ \dot{h}(x)/h(r)$ is an even function, as is $\beta_2(r)/h^{1/2}(r)$, we have that for k=1,2,3

(3.6) $E\{\ell_{\theta,\sigma}(2\ A_k\beta_k/f^{1/2})\} = 0$.

Note that (2.2) holds if $E[\beta_1(X)/g^{1/2}(x)] = 0$, i.e., if $\int \beta_1(\mu)/g^{1/2}(\mu)\ s(\mu)d\mu = 0$. Therefore (2.2) is implied by (2.4) if ν_1 is chosen so that $g_0(\mu)d\nu_1(\mu) = s_0(\mu)d\mu$ where s_0 and g_0 are the true values of the density parameters s ang g. It follows from (2.2) and (2.5) that $\beta_{*k} = 0$ for k=1,2,3, so that the efficient score is $\tilde{\ell}_\theta = \ell_\theta$ and

(3.7) $$I_* = E\ell_\theta\ell_\theta^T$$

$$= E\{\dot{h}(r)/h(r)\}^2\ E\left\{\frac{xx^T}{\sigma^2 g(x)}\right\} .$$

Note that this is the same information bound as for the purely parametric case that g,h,s are all known. Thus, in principle at least, asymptotically we should be able to adapt to (g,h,s), i.e., estimate θ as well as if (g,h,s) were known.

4. INFORMATION BOUNDS FOR MEAN-HETEROSCEDASTICITY

The model is (1.1)-(1.3). Again $s(x)$ is the density of x and $h(\cdot)$ is the density of $(Y-x^T\theta)/(\sigma g^{1/2}(x^T\theta)$, but now the variance of Y is $\sigma^2 g(x^T\theta)$. In this section we compute the information bound I_* for estimating θ when (g,h,s) is unknown, and find that it is between the parametric information when (g,h,s) is known and the asymptotic variance of a "weighted" likelihood estimate.

The density function is given by

$$(4.1) \qquad f = (\sigma g^{1/2}(x^T\theta))^{-1}\, s(x)\, h\{(Y-x^T\theta)/(\sigma g^{1/2}(x^T\theta))\} .$$

Again letting $r = (Y-x^T\theta)/(\sigma g^{1/2}(x^T\theta))$, we find that

$$(4.2) \qquad \ell_{\theta,\sigma} = -(r\,\dot{h}(r)/h(r) + 1)\left\{\frac{x\,\dot{g}(x^T\theta)}{2\,g(x^T\theta)} - E\left(\frac{x}{2}\frac{\dot{g}}{g}\right)\right\} - \frac{x\,\dot{h}(r)}{\sigma g^{1/2}(x^T\theta)h(r)}$$

Considering (3.2)-(3.5) with the difference that now g is a function of $x^T\theta$, we obtain

$$2(A_1\beta_1)/f^{1/2} = -\beta_1(x^T\theta)\,\{r\,\dot{h}(r)/h(r) + 1\}\,/g^{1/2}(x^T\theta)$$

$$2(A_2\beta_2)/f^{1/2} = 2\,\beta_2(r)/h^{1/2}(r)$$

$$2(A_3\beta_3)/f^{1/2} = 2\,\beta_3(x)/s^{1/2}(x) .$$

The orthogonality condition (3.6) holds for $A_2\beta_2$ and $A_3\beta_3$ but not for $A_1\beta_1$. Suppose that $x^T\theta$ is not constant on the support of x, where θ is the true parameter. Noting that $(A_1\beta_1)/f^{1/2}$ is a function of the data only through $|r|$ and $x^T\theta$, it follows from the least squares projection theorem that

$$(4.3)\qquad 2(A_1\beta_{*1})/f^{1/2} = E\{\ell_{\theta,\sigma} \mid x^T\theta,\ |r|\}$$

$$= -(r\,\dot{h}(r)/h(r) + 1)\left\{\frac{\dot{g}\ (x^T\theta)}{2\ g(x^T\theta)}\,E(x|x^T\theta) - E(\frac{x}{2}\frac{\dot{g}}{g})\right\}.$$

To see this, we must check two points. The first is (2.2), which follows immediately upon noting that $A_2\beta_2$ and $A_1\beta_1$ depend on r only through $|r|$, while it holds for $A_3\beta_3$ since $E\{r\,\dot{h}(r)/h(r)\} = -1$. As before, (2.2) follows from the judicious choice of ν_1. To see this let

$$t(v) = E(x|x^T\theta=v),$$

we see that

$$2\ g^{1/2}(v)\beta_*(v) = \dot{g}(v)t(v) - g(v)\ E(x\ \dot{g}/g)\ .$$

Suppose that we are interested in the information bound at $(\theta_o,\sigma_o,g_o,h_o,s_o)$. Assume that g_o is continuous, that h_o,s_o are densities with respect to measures ν_2 and ν_3 respectively, and that the random variable $x^T\theta_o$ has continuous density ξ_o with respect to Lebesgue measure. Define $\nu_1(da) = (\xi_o(a)/g_o(a))da$. Then condition (2.3) will hold.

From (4.3) it follows that the efficient score function is

$$(4.4)\qquad \tilde{\ell}_{\theta,\sigma} = \frac{-x\ \dot{h}(r)}{\sigma g^{1/2}(x^T\theta)h(r)} - \tfrac{1}{2}(r\ \dot{h}(r)/h(r) + 1)(\dot{g}(x^T\theta)/g(x^T\theta))$$

$$\cdot\ \{x - E(x|x^T\theta)\}\ .$$

The information bound is then

$$(4.5) \qquad I_* = S_W^{-1} + (1/4)\, E\{r\, \dot{h}(r)/h(r) + 1\}^2$$

$$\cdot\, E\{(\dot{g}(x^T\theta)/g(x^T\theta))^2 (x - E(x|x^T\theta))(x - E(x|x^T\theta))^T\}\ .$$

When $x^T\theta$ is constant over the support of x, $A_1\beta_1/f^{1/2}$ is a function of the data only through $|r|$, so that $g(x^T\theta) = 1$ by convention and

$$\tilde{\ell}_{\theta,\sigma} = -\frac{x\, \dot{h}(r)}{\sigma\, h(r)}\ , \quad \text{and} \quad I_* = S_W^{-1}\ .$$

In effect, for the homoscedastic case that either g or $x^T\theta$ is constant, the information bound is the usual homoscedastic information.

5. INFORMATION BOUNDS AT THE NORMAL DISTRIBUTION

It is worth noting the following fact. Suppose we know $h(\cdot)$ a priori, e.g., we assume that the data follow a normal distribution. There is no extra information involved in knowing $h(\cdot)$ exactly if we already know it is symmetrically distributed about zero as in (1.1). Thus, even assuming normality, the information bounds (3.7) and (4.5) are unchanged.

6. ACHIEVING THE INFORMATION BOUNDS

For normally distributed observations and simple linear regression, the bound (3.7) is achieved by the estimator introduced by Carroll (1982). For mean heteroscedasticity in simple linear regression or any other model where the map $x \to x^T\theta$ is one-to-one, $x = E(x|x^T\theta)$ so that the second term in (4.5) vanishes. In this case, an estimator introduced by Carroll (1982) has been shown to achieve the information bounds when the data are normally distributed.

We now sketch our reasons for believing that estimators can be constructed which achieve the information bounds (3.7) and (4.5). We are presently working on finding precise conditions under which the following arguments are technically correct. Our starting point is the one-step construction used in Theorem 3.1 of Bickel (1982) and generalized somewhat in an unpublished paper by W.M. Huang. As outlined by Huang and Bickel in his conditions GR(iv) and H', the three essential steps are (1) that a root-N consistent estimator of θ exists; (2) that consistent estimates of the information bound (3.7) or (4.5) exist; (3) we can consistently estimate the optimal score function (3.2) or (4.4) rather well. The first step is easy, because least squares and simple M-estimates are already root-N consistent. For predictor-heteroscedasticity, the second and third steps should not be too hard to verify by using the kernel estimate of $g(\cdot)$ proposed by Carroll (1982) and a kernel estimate of $\dot{h}/h$ as in, for example, Lemma 4.1 of Bickel (1982). The second and third steps should hold for mean-heteroscedasticity as well, but are likely to be much harder technically. The reason is that in (4.4) and (4.5), we need to estimate not only g and $\dot{h}/h$, but also $\dot{g}/g$ and $E(x|x^T\theta)$.

If the distribution of $\{x_i\}$ is discrete rather than continuous as assumed in (1.2), there are problems of identifiability since many different functions g will fit the variance function at each support point x. Our belief is that, for this case, no real asymptotic improvement will be possible over ordinary weighted least squares unless the function g is more tightly specified.

The issue of robustness raised by Carroll & Ruppert (1982b) is still an important one. Generalized least squares, which has asymptotic variance S_w, is typically rather robust against small deviations in the model, e.g., the variance not quite a function of the mean, say but rather depending on x in a slightly different fashion. Our guess is that the same cannot be said of any estimator achieving the information bound.

We have calculated the possible asymptotic improvement over generalized least squares for normally distributed data design and means of Jobson & Fuller (1980) for the special model

$$E\,Y_i = x_i^T\theta$$

$$\mathrm{Var}(Y_i) = \sigma^2(x_i^T\theta)^\alpha\,.$$

The improvement tended to be monotonically increasing in the coefficient of variation, becoming noticable only when the average coefficient of variation exceeded 0.40 . In our experience, nearly normally distributed heteroscedastic data typically have average coefficients of variation not exceeding 0.30 . While our calculations are too fragmentary to make any general conclusions, they do suggest that when the form of the variance function is unknown, the simple smoothing techniques of Carroll (1982) will often be nearly asymptotically efficient.

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

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

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

Carroll, R.J. & Ruppert, D. (1982b). 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. & Ruppert, D. (1983). Robust estimators for random coefficient regression models. In Contributions to Statistics: Essays in Honour of Norman L. Johnson, P.K. Sen, editor. North Holland.

Hildreth, C. & Houck, J.P. (1968). Some estimates for a linear model with random coefficients. Journal of the American Statistical Association 63, 584-595.

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.

Johansen, S. (1983). Some topics in regression. Scandanavian Journal of Statistics 10, 161-194.

Matloff, N., Rose, R. & Tai, R. (1984). A comparison of two methods for estimation optimal weights in regression analysis. Journal of Statistical Computation and Simulation, to appear.

Pritchard, D.J., Downie, J. & Bacon, D.W. (1977). Further consideration of heteroscedasticity in fitting kinetic models. Technometrics 19, 227-236.

TABLE 1

Comparisons based on $|\ddagger|^{1/3}$ of various asymptotic covariances

α	σ	Generalized Least Squares	α known	α unknown	Form of Variance Function Unknown
0.0	0.5	.024	.024	.024	.024
	1.0	.096	.096	.096	.096
0.5	0.5	.921	.918	.919	.919
	1.0	3.681	3.643	3.660	3.66
1.0	0.5	29.06	22.18	25.03	25.39
	1.0	116.25	55.89	76.43	80.59