The Annals of Statistics
1993, Vol. 21, No. 2, 1071–1092

ITERATIVE WEIGHTED LEAST SQUARES ESTIMATORS[1]

BY JIAHUA CHEN AND JUN SHAO

University of Waterloo and University of Ottawa

In a heteroscedastic linear model, we establish the asymptotic normality of the iterative weighted least squares estimators with weights constructed by using the within-group residuals obtained from the previous model fitting. An adaptive procedure is proposed which ensures that the iterative process stops after a finite number of iterations and produces an estimator asymptotically equivalent to the best estimator that can be obtained by using the iterative procedure. Theoretical and empirical results of the performance of the adaptive estimator are presented.

1. Introduction. One of the most useful models in statistical applications is the following general linear model:

$$y_{ij} = x_{ij}'\beta + e_{ij}, \qquad j = 1, 2, \ldots, n_i, \qquad i = 1, 2, \ldots, k, \tag{1.1}$$

where y_{ij} are responses, x_{ij} are values of a p-dimensional covariate, x_{ij}' is the transpose of x_{ij}, β is a p-vector of unknown parameters and e_{ij} are random errors. Usually, the e_{ij} are assumed to be mutually independent, $E(e_{ij}) = 0$ for all i, j, and the variances $\operatorname{var}(e_{ij})$ exist but are unknown and not necessarily equal.

Let $\theta = g(\beta)$ be the parameter of interest, where g is a known function from R^p to R. If $\operatorname{var}(e_{ij}) = \sigma^2$ for all i, j, the customary estimator of θ is the ordinary least squares estimator (OLSE) given by

$$\hat{\theta}_o = g(\hat{\beta}_o), \qquad \hat{\beta}_o = \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} x_{ij}' \right)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij} y_{ij}.$$

In many situations the errors in the same group, e_{ij}, $j = 1, 2, \ldots, n_i$, have a common distribution. For example, $x_{ij} = x_i$ for all j and y_{ij}, $j = 1, 2, \ldots, n_i$, are replications. But the errors from different groups, e_{ij} and $e_{i'j'}$ with $i \neq i'$, may have different distributions. Hence $\sigma_i^2 = \operatorname{var}(e_{ij})$ are not necessarily the same. Under such a case, the OLSE may be improved by a weighted least squares estimator (WLSE)

$$\hat{\theta}_w = f(\hat{\beta}_w), \qquad \hat{\beta}_w = \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} w_i x_{ij} x_{ij}' \right)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} w_i x_{ij} y_{ij} \tag{1.2}$$

Received March 1991; revised March 1992.
[1]Research supported by the Natural Sciences and Engineering Research Council of Canada.
AMS 1991 *subject classifications*. Primary 62J05; secondary 60F05.
Key words and phrases. Asymptotic normality, combining groups, adaptive estimator, efficiency.

for some weights $w_i > 0$, $i = 1, 2, \ldots, k$. If σ_i^2 are known,

$$\tilde{\theta} = g(\tilde{\beta}), \qquad \tilde{\beta} = \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} \sigma_i^{-2} x_{ij} x_{ij}' \right)^{-1} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \sigma_i^{-2} x_{ij} y_{ij} \tag{1.3}$$

improves $\hat{\theta}_o$ and has some optimality property. Since σ_i^2 are unknown, we need to estimate σ_i^2 by $\hat{\sigma}_i^2$ and use $w_i = \hat{\sigma}_i^{-2}$ in (1.2).

In some situations σ_i^2 are related to the x_{ij} or $x_{ij}'\beta$, $j = 1, 2, \ldots, n_i$, through a smooth but unknown function h, that is, $\sigma_i^2 = h(x_{i1}, x_{i2}, \ldots, x_{in_i})$. In these situations σ_i^2 can be estimated consistently and the resulting WLSE is asymptotically as efficient as $\tilde{\theta}$ in (1.3) [Carroll (1982) and Davidian and Carroll (1987)]. However, there are situations where σ_i^2 is unrelated to the x_{ij} and therefore this approach cannot be applied. For example, in a "common mean" problem [Fuller and Rao (1978)],

$$y_{ij} = \mu + e_{ij}.$$

The σ_i^2 are unrelated to x_{ij} since $x_{ij} = 1$ for all i, j. More generally, in a "common regression coefficients" problem [Box and Tiao (1973), pages 478–479], where for each i,

$$y_{ij} = x_{ij}'\beta + e_{ij}, \qquad j = 1, 2, \ldots, n_i,$$

is a regression model and we wish to combine the data from k different communities (batches, days) to improve the accuracy of estimates, the heteroscedasticity in σ_i^2 is usually caused by the variation among different communities (batches, days). In these situations estimators of σ_i^2 such as the MINQUE and its modification were proposed and studied in the literature [Rao (1970), (1973)]. If k is fixed and $\min_{i \le k} n_i \to \infty$, then these estimators are consistent and the resulting WLSE is as efficient as $\tilde{\theta}$. We confine our attention to the situation.

$$\sup_i n_i < \infty \quad \text{and} \quad k \text{ is large},$$

which is particularly true if y_{ij}, $j = 1, 2, \ldots, n_i$, are replications [Fuller and Rao (1978)]. Rao (1973) proposed a modified MINQUE of σ_i^2, the within-group average of residual squares

$$v_i(\hat{\beta}) = \frac{1}{n_i} \sum_{i=1}^{n_i} \left(y_{ij} - x_{ij}'\hat{\beta} \right)^2, \tag{1.4}$$

where $\hat{\beta}$ is an estimator of β. The WLSE constructed by using v_i is better than the WLSE constructed by using the MINQUE or the within-group sample variance [Rao (1973) and Carroll and Cline (1988)].

Because of the existence of a large number of nuisance parameters σ_i^2, estimators of σ_i^2 are not consistent as $k \to \infty$ and therefore the corresponding WLSE is not as efficient as the "estimator" $\tilde{\theta}$ in (1.3). Hence it is still possible to improve the WLSE. A natural appraoch suggested by many researchers is to

use the following iterative procedure:

(1.5) Obtain an estimate of $\beta \rightarrow$ obtain estimators of $\sigma_i^2 \rightarrow$ feedback and repeat.

More precisely, the WLSE after the mth iteration is

$$(1.6)\quad \hat{\theta}_w^{(m)} = g\left(\hat{\beta}_w^{(m)}\right), \qquad \hat{\beta}_w^{(m)} = \left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} w_i^{(m)} x_{ij} x_{ij}'\right)^{-1} \sum_{i=1}^{k}\sum_{j=1}^{n_i} w_i^{(m)} x_{ij} y_{ij},$$

where

$$w_i^{(m)} = \left[v_i\left(\hat{\beta}_w^{(m-1)}\right)\right]^{-1}, \qquad i = 1, 2, \ldots, k,$$

for $m = 1, 2 \ldots$. The OLSE $\hat{\beta}_o$ can be used as an initial estimator $\hat{\beta}_w^{(0)}$, in which case $\hat{\beta}_w^{(1)}$ is the WLSE proposed in Fuller and Rao (1978).

This raises some interesting questions:

(a) What is the asymptotic distribution of $\hat{\theta}_w^{(m)}$ for each fixed m?
(b) Is there a finite integer m^* such that $\hat{\theta}_w^{(m^*)}$ is the best among $\hat{\theta}_w^{(m)}$, $m = 1, 2, \ldots$?
(c) If such an m^* exists but is unknown, when should we stop the iterative process?

These problems are studied in the present paper. In the case where σ_i^2 are related to the x_{ij} or $x_{ij}'\beta$, problems similar to (a)–(c) were studied in Carroll, Wu and Ruppert (1988), for the small sample case.

Our main findings are the following.

1. It is shown in Section 2 that under some regularity conditions, the WLSE given by (1.2) with $w_i = [v_i(\hat{\beta})]^{-1}$ given by (1.4) is asymptotically normal with mean θ. As a consequence of this result, the WLSE $\hat{\theta}_w^{(m)}$ for each fixed m is asymptotically normal with mean θ and its asymptotic variance can be explicitly obtained.
2. It is shown in Section 3 that under some conditions, there exists a finite integer m^* such that in terms of its asymptotic efficiencies, $\hat{\theta}_w^{(m^*)}$ is better than $\hat{\theta}_w^{(m)}$ for any $m \neq m^*$. This m^* depends on σ_i^2, $i = 1, 2, \ldots, k$, and therefore is unknown.
3. An adaptive procedure is proposed in Section 3. Three main features of this adaptive procedure are: (i) It ensures that for given data, the iterative process (1.5) stops after a finite number of iterations; (ii) the resulting adaptive estimator of θ has an optimality property, that is, it is better than any $\hat{\beta}_w^{(m)}$ and is as good as the optimal $\hat{\beta}_w^{(m^*)}$ if it exists and is unique; (iii) an estimate of the asymptotic variance of the adaptive estimator is obtained.
4. If $n_i \leq 2$ for all i, it is shown in Section 4 that asymptotically the WLSE cannot improve the OLSE. We consider combining some groups so that each new group contains at least three observations. The adaptive proce-

dure can then be applied to the data with combined groups. The resulting adaptive estimator improves the OLSE in some cases.

5. A simulation study is presented in Section 5. The performance of the adaptive estimator is shown to be adequate in the simulation study.

2. General asymptotic results. Denote the $t \times t$ identity matrix by I_t. Let

$$\mathbf{e}_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})',$$

$$\mathbf{e} = (\mathbf{e}_1', \mathbf{e}_2', \ldots, \mathbf{e}_k')',$$

$$u_i = \mathbf{e}_i'\mathbf{e}_i/n_i,$$

$$\tilde{\mathbf{e}}_i = \mathbf{e}_i/u_i,$$

$$\tilde{\mathbf{e}} = (\tilde{\mathbf{e}}_1', \tilde{\mathbf{e}}_2', \ldots, \tilde{\mathbf{e}}_k')',$$

$$V_i = \operatorname{cov}(\mathbf{e}_i),$$

$$V = \text{block diag}(V_1, V_2, \ldots, V_k),$$

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})',$$

$$X = (X_1', X_2', \ldots, X_k')'.$$

The following assumptions will be used.

ASSUMPTION A. There are positive constants, σ_o, σ_∞, c_o and c_∞ and positive integers n_o and n_∞ such that for all i and j,

$$\sigma_o I_{n_i} \le V_i \le \sigma_\infty I_{n_i}, \qquad c_o I_p \le k^{-1}(X'X),$$

$$n_o \le n_i \le n_\infty, \qquad x_{ij}'x_{ij} \le c_\infty.$$

ASSUMPTION B. $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_k$ are independent.

ASSUMPTION C. There are positive constants d and δ such that for all i, j,

$$E\left(e_{ij} \Big/ \sum_{j=1}^{n_i} e_{ij}^2\right) = 0,$$

$$E|e_{ij}|^{2+\delta} \le d, \qquad E\left(\sum_{j=1}^{n_i} e_{ij}^2\right)^{-(1+\delta)} \le d.$$

The first part of Assumption C reflects a certain degree of symmetry of the error distributions. Without this condition, the WLSE may be inconsistent [Carroll and Cline (1988), Theorem 3].

We first establish a general asymptotic result for a WLSE $\hat{\beta}_w$.

THEOREM 1. *Suppose that Assumptions* A, B *and* C *hold. Let* $\hat{\beta}_w$ *be given by* (1.2) *with* $w_i = [v_i(\hat{\beta})]^{-1}$, $v_i(\hat{\beta})$ *given by* (1.4) *with* $\hat{\beta}$ *satisfying*

$$\hat{\beta} - \beta = G_k^{-1}(A_k\tilde{\mathbf{e}} + B_k\mathbf{e}) + o_p(k^{-1/2}), \tag{2.1}$$

where $G_k = \sum_{i=1}^{k} Eu_i^{-1}X_i'X_i$, A_k *and* B_k *satisfy* $A_kA_k' = O(k)$ *and* $B_kB_k' = O(k)$. *Then*

$$\hat{\beta}_w - \beta = G_k^{-1}\left(\tilde{A}_k\tilde{\mathbf{e}} + \tilde{B}_k\mathbf{e}\right) + o_p(k^{-1/2}), \tag{2.2}$$

where

$$\tilde{A}_k = X' + 2H_kG_k^{-1}A_k,$$
$$\tilde{B}_k = 2H_kG_k^{-1}B_k$$

and

$$H_k = E\left(\sum_{i=1}^{k} n_i^{-1}X_i'\tilde{\mathbf{e}}_i\tilde{\mathbf{e}}_i'X_i\right).$$

REMARK. It follows from (2.2) that

$$\left[G_k^{-1}\Sigma_kG_k^{-1}\right]^{-1/2}\left(\hat{\beta}_w - \beta\right) \to_d N(0, I_p),$$

where $\Sigma_k = \operatorname{var}(\tilde{A}_k\tilde{\mathbf{e}} + \tilde{B}_k\mathbf{e})$. Result (2.2) was established in Shao (1989a) for the special case where $\hat{\beta} = \hat{\beta}_o$ and e_{ij} are mutually independent.

PROOF OF THEOREM 1. Let $v_i = v_i(\hat{\beta})$. Then

$$\hat{\beta}_w - \beta = \left(\sum_{i=1}^{k} v_i^{-1}X_i'X_i\right)^{-1}\sum_{i=1}^{k} v_i^{-1}X_i'\mathbf{e}_i. \tag{2.3}$$

Note that

$$\begin{aligned} v_i^{-1} &= u_i^{-1} - u_i^{-2}(v_i - u_i) + u_i^{-2}v_i^{-1}(v_i - u_i)^2 \\ &= u_i^{-1} + 2n_i^{-1}u_i^{-2}\sum_{j=1}^{n_i}\phi_{ij}e_{ij} - n_i^{-1}u_i^{-2}\sum_{j=1}^{n_i}\phi_{ij}^2 + u_i^{-2}v_i^{-1}(v_i - u_i)^2, \end{aligned} \tag{2.4}$$

where $\phi_{ij} = x_{ij}'(\hat{\beta} - \beta)$. We now show that

$$\sum_{i=1}^{k} u_i^{-2}v_i^{-1}(v_i - u_i)^2X_i'\mathbf{e}_i = o_p(k^{1/2}) \tag{2.5}$$

and

$$\sum_{i=1}^{k} n_i^{-1}u_i^{-2}\sum_{j=1}^{n_i}\phi_{ij}^2X_i'\mathbf{e}_i = o_p(k^{1/2}). \tag{2.6}$$

For $(2(1 + \delta))^{-1} < \alpha_1 < \alpha_2 < 1/2$, let

$$\mathbf{A}_k = \left\{\min_i \max_j |e_{ij}| \geq k^{-\alpha_1},\ \max_{i,j}|x_{ij}'(\hat{\beta} - \beta)| \leq k^{-\alpha_2}\right\}.$$

Let c be a generic constant. On the event $\mathbf{A}_k$,

$$v_i \geq \max_j |e_{ij} - x'_{ij}(\hat{\beta} - \beta)|^2/n_i$$

$$\geq (1 - k^{\alpha_1 - \alpha_2})^2 \max_j e_{ij}^2/n_i \geq (1 - k^{\alpha_1 - \alpha_2})^2 \sum_{j=1}^{n_i} e_{ij}^2/n_i^2 \geq cu_i$$

and therefore

$$u_i^{-2} v_i^{-1} (v_i - u_i)^2 \leq c(v_i - u_i)^2 u_i^{-3}.$$

Under Assumption A,

$$(v_i - u_i)^2 = n_i^{-2}\left[\sum_{j=1}^{n_i} \phi_{ij}^2 - 2\sum_{j=1}^{n_i} \phi_{ij} e_{ij}\right]^2$$

$$\leq \left[\sum_{j=1}^{n_i} \phi_{ij}^2 + 2\left(\sum_{j=1}^{n_i} e_{ij}^2\right)^{1/2}\left(\sum_{j=1}^{n_i} \phi_{ij}^2\right)^{1/2}\right]^2$$

$$\leq c\left[\|\hat{\beta} - \beta\|^4 + u_i\|\hat{\beta} - \beta\|^2\right],$$

where $\|\hat{\beta} - \beta\|^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta)$. Consequently, on the event $\mathbf{A}_k$,

$$\left|\sum_{i=1}^{k} u_i^{-2} v_i^{-1} (v_i - u_i)^2 X_i' \mathbf{e}_i\right| \leq c\left(\|\hat{\beta} - \beta\|^4 \sum_{i=1}^{k} u_i^{-5/2} + \|\hat{\beta} - \beta\|^2 \sum_{i=1}^{k} u_i^{-3/2}\right)$$

$$= o_p(k^{1/2}),$$

since $\|\hat{\beta} - \beta\| = O_p(k^{-1/2})$ and $\max_i u_i^{-1} \leq c \max_i \min_j e_{ij}^{-2} \leq ck^{2\alpha_1} = o_p(k)$.

Note that

$$P\{\mathbf{A}_k^c\} \leq P\left\{\max_{i,j} |x'_{ij}(\hat{\beta} - \beta)| \geq k^{-\alpha_2}\right\} + \sum_{i=1}^{k} P\left\{\max_j |e_{ij}| \leq k^{-\alpha_1}\right\}. \tag{2.7}$$

Since $\|\hat{\beta} - \beta\| = O_p(k^{-1/2})$, the first term on the right side of (2.7) goes to 0. The second term on the right side of (2.7) also goes to 0, since

$$P\left\{\max_j |e_{ij}| \leq k^{-\alpha_1}\right\} \leq P\{u_i \leq k^{-2\alpha_1}\} \leq P\{u_i^{-1} \geq k^{2\alpha_1}\}$$

$$\leq Eu_i^{-(1+\delta)} k^{-2\alpha_1(1+\delta)}$$

and $2\alpha_1(1 + \delta) > 1$. This proves $P(\mathbf{A}_k) \to 1$, and hence (2.5) holds. The proof of (2.6) is similar.

From (2.4)–(2.6),

$$\sum_{i=1}^{k} v_i^{-1} X_i' \mathbf{e}_i = \sum_{i=1}^{k} u_i^{-1} X_i' e_i + 2\left[\sum_{i=1}^{k} n_i^{-1} X_i' \tilde{\mathbf{e}}_i \tilde{\mathbf{e}}_i' X_i\right](\hat{\beta} - \beta) + o_p(k^{1/2}). \tag{2.8}$$

Following the argument in the proof of (2.5), we obtain that

$$k^{-1}\left(\sum_{i=1}^{k} v_i^{-1}X_i'X_i - G_k\right) \to_p 0,$$

which, together with Assumption A, implies that

$$k\left[\left(\sum_{i=1}^{k} v_i^{-1}X_i'X_i\right)^{-1} - G_k^{-1}\right] \to_p 0.$$

Also, it is easy to see that

$$k^{-1}\left(\sum_{i=1}^{k} n_i^{-1}X_i'\tilde{\mathbf{e}}_i\tilde{\mathbf{e}}_i'X_i - H_k\right) \to_p 0.$$

Hence the result follows from (2.3) and (2.8). □

Let $\{A_k^{(0)}\}$ and $\{B_k^{(0)}\}$ be the two sequences of $p \times (\sum_{i=1}^{k} n_i)$ matrices satisfying $(A_k^{(0)})(A_k^{(0)})' = O(k)$ and $(B_k^{(0)})(B_k^{(0)})' = O(k)$. For any positive integer m, let

$$A_k^{(m)} = \sum_{t=0}^{m-1} \left(2H_kG_k^{-1}\right)^t X' + \left(2H_kG_k^{-1}\right)^m A_k^{(0)}$$

and

$$B_k^{(m)} = \left(2H_kG_k^{-1}\right)^m B_k^{(0)},$$

where G_k and H_k are given in Theorem 1. Applying the δ-method and Theorem 1 repeatedly with $A_k = A_k^{(t-1)}$, $B_k = B_k^{(t-1)}$, $\hat{\beta} = \hat{\beta}_w^{(t-1)}$ and $\hat{\beta}_w = \hat{\beta}_w^{(t)}$, $t = 1, 2, \ldots, m$, we obtain the following result for the WLSE $\hat{\theta}_w^{(m)}$ given in (1.6).

THEOREM 2. *Assume the conditions in Theorem* 1 *and that* (2.1) *holds with* $\hat{\beta} = \hat{\beta}_w^{(0)}$. *Assume also that* g *is differentiable at* β. *Then for any fixed* m,

$$\left\{\nabla g(\beta)G_k^{-1}\Sigma_k^{(m)}G_k^{-1}[\nabla g(\beta)]'\right\}^{-1/2}\left(\hat{\theta}_w^{(m)} - \theta\right) \to_d N(0,1), \tag{2.9}$$

where $\Sigma_k^{(m)} = \mathrm{Var}(A_k^{(m)}\tilde{\mathbf{e}} + B_k^{(m)}\mathbf{e})$.

We now discuss some important special cases.

1. When $\hat{\beta}_w^{(0)} = \hat{\beta}_o$ (the OLSE), (2.1) is satisfied since

$$A_k^{(0)} = 0 \quad \text{and} \quad B_k^{(0)} = G_k(X'X)^{-1}X'.$$

Then for $m \geq 1$,

$$A_k^{(m)} = \sum_{t=0}^{m-1} \left(2H_kG_k^{-1}\right)^t X',$$

$$B_k^{(m)} = \left(2H_kG_k^{-1}\right)^m G_k(X'X)^{-1}X'.$$

2. If e_{ij}/σ_i, for $j = 1, 2, \ldots, n_i$; $i = 1, 2, \ldots, k$, are independent and identically distributed (i.i.d.), $n_i = n_o$ for all i and if for $j \neq j'$,

$$E(e_{ij}e_{ij'})/u_i^2 = 0, \tag{2.10}$$

then $H_k = n_o^{-1}G_k$, $G_k = \rho(n_o)X'V^{-1}X$, where $\rho(n_o) = \sigma_i^2 E u_i^{-1}$,

$$A_k^{(m)} = \sum_{t=0}^{m-1}\left(\frac{2}{n_o}\right)^t X' + \left(\frac{2}{n_o}\right)^m A_k^{(0)}$$

$$= \left(1 - \frac{2}{n_o}\right)^{-1}\left[1 - \left(\frac{2}{n_o}\right)^m\right]X' + \left(\frac{2}{n_o}\right)^m A_k^{(0)}$$

and

$$B_k^{(m)} = \left(\frac{2}{n_o}\right)^m B_k^{(0)}.$$

3. If e_{ij}/σ_i are i.i.d. and for $j \neq j'$,

$$Ee_{ij}e_{ij'}/u_i = 0, \tag{2.11}$$

then

$$\Sigma_k^{(m)} = A_k^{(m)}E(\tilde{\mathbf{e}}\tilde{\mathbf{e}}')\left(A_k^{(m)}\right)' + A_k^{(m)}\left(B_k^{(m)}\right)' + B_k^{(m)}\left(A_k^{(m)}\right)' + B_k^{(m)}V\left(B_k^{(m)}\right)'.$$

Note that (2.10) and (2.11) hold if e_{ij}/σ_i are i.i.d. with a symmetric distribution. Combining (1)–(3), we have the following result. The proof is straightforward and omitted.

THEOREM 3. *Suppose that the conditions in Theorem* 2 *and* (2.10) *and* (2.11) *hold. Assume the* e_{ij}/σ_i *are i.i.d.,* $n_i = n_o$ *for all* i *and* $\hat{\beta}_w^{(0)} = \hat{\beta}_o$. *Then for any* m,

$$\left(\tau_k^{(m)}/k\right)^{-1/2}\left(\hat{\theta}_w^{(m)} - \theta\right) \to_d N(0,1),$$

with

$$\tau_k^{(m)} = k\left[\frac{\varphi(m)}{\rho(n_o)}\nabla g(\beta)(X'V^{-1}X)^{-1}\nabla g(\beta)'\right.$$

$$\left. + \psi(m)\nabla g(\beta)(X'X)^{-1}(X'VX)(X'X)^{-1}\nabla g(\beta)'\right],$$

where

$$\varphi(m) = \left(1 - \frac{2}{n_o}\right)^{-2}\left(1 - \left(\frac{2}{n_o}\right)^m\right)^2 + 2\left(1 - \frac{2}{n_o}\right)^{-1}\left(1 - \left(\frac{2}{n_o}\right)^m\right)\left(\frac{2}{n_o}\right)^m$$

and

$$\psi(m) = \left(\frac{2}{n_o}\right)^{2m}.$$

3. The efficiency of the WLSE and an adaptive estimator. We now study the asymptotic efficiency of $\hat{\theta}_w^{(m)}$. For simplicity, we confine our attention to the situation where $n_i = n_o$ for all i and e_{ij}/σ_i are i.i.d. Let

$$b_k = k\left[\nabla g(\beta)(X'X)^{-1}(X'VX)(X'X)^{-1}\nabla g(\beta)'\right]$$

and

$$\tilde{b}_k = k\left[\nabla g(\beta)(X'V^{-1}X)^{-1}\nabla g(\beta)'\right].$$

Note that b_k/k and $\tilde{b}_k/k$ are the asymptotic variances of $\hat{\theta}_o$ and $\tilde{\theta}$, respectively. From Theorem 3, the asymptotic variance of $\hat{\theta}_w^{(m)}$ for a fixed m is

$$\tau_k^{(m)}/k = \left[\frac{\varphi(m)}{\rho(n_o)}\tilde{b}_k + \psi(m)b_k\right]\Big/k, \tag{3.1}$$

where $\varphi(m)$ and $\psi(m)$ are given in Theorem 3:

LEMMA 1. *When* $n_o \geq 3$,

(i) $\varphi(m)$ *is strictly increasing in* m.

(ii) *The function*

$$\Delta(m) = \frac{\varphi(m+1) - \varphi(m)}{\psi(m) - \psi(m+1)}$$

is strictly increasing in m *and* $\Delta(m) \to \infty$ *as* $m \to \infty$.

PROOF. (i) From the expression of $\varphi(m)$, we need only consider the function

$$\left[\left(1 - \left(\frac{2}{n_o}\right)^m\right)^2 + 2\left(1 - \frac{2}{n_o}\right)\left(1 - \left(\frac{2}{n_o}\right)^m\right)\left(\frac{2}{n_o}\right)^m\right]$$

$$= 1 - \frac{4}{n_o}x_m + \left(\frac{4}{n_o} - 1\right)x_m^2,$$

where $x_m = (2/n_o)^m$. Since

$$\frac{d}{dx}\left[1 - \frac{4}{n_o}x + \left(\frac{4}{n_o} - 1\right)x^2\right] = 2\left(\frac{4}{n_o} - 1\right)x - \frac{4}{n_o} < 0,$$

when $x \leq 2/3$, the function $1 - (4/n_o)x + ((4/n_o) - 1)x^2$ is strictly decreasing in x. This implies $\varphi(m)$ is strictly increasing in m.

(ii) Similarly, with $x_m = (2/n_o)^m$,

$$\Delta(m) = \left(1 - \frac{2}{n_o}\right)^{-2}\left[\frac{4(x_m - x_{m+1})}{n_o(x_m^2 - x_{m+1}^2)} + \left(1 - \frac{4}{n_o}\right)\right],$$

which is strictly increasing in m. Finally,

$$\lim_{m\to\infty} \frac{x_m - x_{m+1}}{(x_m^2 - x_{m+1}^2)} = \lim_{m\to\infty} \frac{1}{x_m + x_{m+1}} = \infty. \qquad \square$$

Using this lemma, we can compare the asymptotic efficiencies of $\hat{\theta}_w^{(m)}$, $m = 0, 1, 2, \ldots$. Throughout this section we assume $n_o \geq 3$. The case of $n_o = 2$ is treated in Section 4.

THEOREM 4. *Suppose that* $b = \lim_{k\to\infty} b_k$ *and* $\tilde{b} = \lim_{k\to\infty} \tilde{b}_k$ *exist. Then there exists an integer* m^* *such that*

$$\lim_{k\to\infty} \tau_k^{(m^*)} \Big/ \tau_k^{(m)} \leq 1$$

for all m *with equality holding at most for* $m = m^*$ *or* $m^* + 1$.

PROOF. Note that

$$\tau_k^{(m+1)} - \tau_k^{(m)} = \frac{\psi(m) - \psi(m+1)}{\rho(n_o)} \tilde{b}_k \left[\Delta(m) - \frac{\rho(n_o) b_k}{\tilde{b}_k}\right].$$

From Assumption A, $0 < b < \infty$ and $0 < \tilde{b} < \infty$. Since $\psi(m) - \psi(m+1) > 0$, whether or not $\lim_{k\to\infty} \tau_k^{(m+1)}/\tau_k^{(m)} \leq 1$ depends on the sign of

$$t(m) = \Delta(m) - \frac{\rho(n_o) b}{\tilde{b}}.$$

From Lemma 1, $t(m)$ can change sign only once as m increases. Thus, as m increases, the quantity

$$\lim_{k\to\infty} \left(\tau_k^{(m+1)}/\tau_k^{(m)} - 1\right)$$

will change sign only once. Since $\Delta(m) \to \infty$ as $m \to \infty$ (Lemma 1), there always exists a unique m^* such that

$$\lim_{k\to\infty} \frac{\tau_k^{(m^*)}}{\tau_k^{(m)}} \begin{cases} < 1 & \text{if } m < m^* \text{ or } m > m^* + 1 \\ \leq 1 & \text{if } m = m^* + 1. \end{cases}$$

This proves the theorem. $\square$

From Theorem 4, in terms of asymptotic variance, $\hat{\theta}_w^{(m^*)}$ is the optimal estimator of θ among $\hat{\theta}_w^{(m)}$, $m = 1, 2, \ldots$, and m^* is the optimal number of iterations. Note that m^* can be as small as 0. $\hat{\theta}_w^{(m^*+1)}$ is also optimal if $\lim_{k\to\infty}(\tau_k^{(m^*+1)}/\tau_k^{(m^*)}) = 1$, which does not occur frequently in practice.

In general, $b_k/\tilde{b}_k$ measures the degree of heteroscedasticity of the model (1.1), that is, $b_k/\tilde{b}_k$ is large when σ_i are very different and $b_k/\tilde{b}_k$ is close to 1 if σ_i are nearly the same ($b_k/\tilde{b}_k \geq 1$ by Jensen's inequality). Since $\Delta(m)$ is increasing as m increases, our result shows that the more different the σ_i are, the more iterations we need.

The optimal m^* is generally unknown and therefore we need to estimate it using the data. Let

$$a_k = \tilde{b}_k/\rho(n_o).$$

Then, by (3.1),

$$\tau_k^{(m)} = \varphi(m)a_k + \psi(m)b_k,$$

where $\varphi(m)$ and $\psi(m)$ are known and a_k and b_k are independent of m. Suppose that $\hat{a}_k$ and $\hat{b}_k$ are consistent estimators of a_k and b_k, respectively. That is,

$$\hat{a}_k - a_k \to_p 0 \quad \text{and} \quad \hat{b}_k - b_k \to_p 0. \tag{3.2}$$

Then $\hat{\tau}_k^{(m)} = \varphi(m)\hat{a}_k + \psi(m)\hat{b}_k$ is consistent for $\tau_k^{(m)}$ for each fixed m.

Let $\hat{r}_k = \hat{b}_k/\hat{a}_k$ and $r_k = b_k/a_k$. Then, by (3.2),

$$\hat{r}_k - r_k \to_p 0. \tag{3.3}$$

Define

$$\hat{m}_k = \min\{m: \Delta(m) \geq \hat{r}_k, m = 0, 1, 2, \ldots\}.$$

THEOREM 5. *Assume that* (3.3) *holds and* $\lim_{k\to\infty} r_k = r$. *Let* m^* *be as given in Theorem* 4. *Then*

$$\lim_{k\to\infty} P(\hat{m}_k = m^* \text{ or } \hat{m}_k = m^* + 1) = 1. \tag{3.4}$$

Furthermore, if m^* *is the unique optimum* ($\lim_{k\to\infty} \tau_k^{(m^*+1)}/\tau_k^{(m^*)} > 1$), *then*

$$\lim_{k\to\infty} P(\hat{m}_k = m^*) = 1. \tag{3.5}$$

PROOF. From Lemma 1 and the proof of Theorem 4,

$$\Delta(m^* - 1) < r \quad \text{and} \quad \Delta(m^* + 1) > r.$$

Then

$$P(\hat{m}_k \leq m^* - 1) \leq P(\Delta(m^* - 1) \geq \hat{r}_k) \to 0,$$

since $P(r - \varepsilon \geq \hat{r}_k) \to 0$ for $\varepsilon = r - \Delta(m^* - 1)$.

Similarly,

$$P(\hat{m}_k \geq m^* + 2) \leq P(\Delta(m^* + 1) < \hat{r}_k) \to 0.$$

This proves (3.4).

Now assume m^* is the unique optimum. Then $\Delta(m^*) > r$ and

$$P(\hat{m}_k = m^* + 1) \leq P(\Delta(m^*) < \hat{r}_k) \to 0,$$

which with (3.4) implies (3.5). □

Result (3.5) indicates that if we stop the iterative procedure (1.5) after $\hat{m}_k$ iterations, we always stop at the right time when k is large enough. This leads to the following adaptive estimator of θ:

$$\hat{\theta}_a = \hat{\theta}_w^{(\hat{m}_k)}.$$

This adaptive estimator has the desired optimality property, provided that the estimators $\hat{a}_k$ and $\hat{b}_k$ satisfying (3.2) can be found. Furthermore, an estimator of the accuracy of $\hat{\theta}_a$ is

$$\hat{\tau}_k/k,$$

with

$$\hat{\tau}_k = \varphi(\hat{m}_k)\hat{a}_k + \psi(\hat{m}_k)\hat{b}_k.$$

THEOREM 6. *Assume that* (3.2) *holds.*

(i) *As* $k \to \infty$,

$$P\left(\hat{\theta}_a = \hat{\theta}_w^{(m^*)} \text{ or } \hat{\theta}_a = \hat{\theta}_w^{(m^*+1)}\right) \to 1.$$

If in addition, m^* *is the unique optimum, then*

$$P\left(\hat{\theta}_a = \hat{\theta}_w^{(m^*)}\right) \to 1.$$

(ii) *Let* $\tau^* = \lim_{k\to\infty} \tau_k^{(m^*)}$. *Then*

$$\hat{\tau}_k - \tau^* \to_p 0.$$

PROOF. The results follow directly from (3.4) and (3.5). □

In practice when there is a computational limitation, one may force the iterative process (1.5) to stop after at most M steps, where M is an integer. That is,

$$\begin{aligned}\tilde{m}_k &= \min\{m: \Delta(m) \geq \hat{r}_k,\ m = 0, 1, 2, \ldots, M\} \\ &= M \quad \text{if } \Delta(M) \leq \hat{r}_k\end{aligned}$$

and

$$\tilde{\theta}_a = \hat{\theta}_w^{(\tilde{m}_k)}. \tag{3.6}$$

From the previous proofs, $\tilde{\theta}_a$ is the best among $\hat{\theta}_w^{(m)}$, $m = 0, 1, 2, \ldots, M$.

Estimators $\hat{a}_k$ and $\hat{b}_k$ satisfying (3.2) are needed to carry out this adaptive procedure. We suggest the use of the following consistent estimators:

$$\hat{a}_k = k \sum_{i=1}^{k} \sum_{j=1}^{n_i} (1 - h_{ij})\left(\hat{\theta}_w^{(i,j)} - \hat{\theta}_w^{(1)}\right)^2, \tag{3.7}$$

$$\hat{b}_k = k \sum_{i=1}^{k} \sum_{j=1}^{n_i} (1 - c_{ij})\left(\hat{\theta}_o^{(i,j)} - \hat{\theta}_o\right)^2, \tag{3.8}$$

where

$$h_{ij} = n_i^{-1} v_i^{-1} x_{ij}' \Big(\sum_{i=1}^k v_i^{-1} X_i' X_i\Big)^{-1} x_{ij}, \qquad v_i = v_i\big(\hat\beta_o\big),$$

$$c_{ij} = n_i^{-1} x_{ij}' (X'X)^{-1} x_{ij},$$

$$\hat\theta_w^{(i,j)} = g\big(\hat\beta_w^{(i,j)}\big),$$

$$\hat\beta_w^{(i,j)} = \hat\beta_w^{(1)} - (1 - h_{ij})^{-1} v_i^{-1} \left(\sum_{i=1}^k v_i^{-1} X_i' X_i \right)^{-1} x_{ij} \big(y_{ij} - x_{ij}' \hat\beta_w^{(1)} \big)$$

and

$$\hat\theta_o^{(i,j)} = g\big(\hat\beta_o^{(i,j)}\big), \qquad \hat\beta_o^{(i,j)} = \hat\beta_o - (1 - c_{ij})^{-1} (X'X)^{-1} x_{ij} \big(y_{ij} - x_{ij}' \hat\beta_o \big).$$

Note that $\hat b_k$ is a consistent (weighted) jackknife estimator of b_k, the asymptotic variance of $k^{1/2}(\hat\theta_o - \theta)$ [Shao (1989b)]. Also, $\hat a_k$ is asymptotically equivalent to

$$k \left[\nabla g(\beta) \left(\sum_{i=1}^k u_i^{-1} X_i' X_i \right)^{-1} \nabla g(\beta)' \right]$$

[Shao (1989b)] and hence $\hat a_k$ is consistent for $\tilde b_k / \rho(n_o) = a_k$ because of $E u_i^{-1} = \sigma_i^{-2} \rho(n_o)$.

In some cases we need to consider simultaneous estimation of $\theta = g(\beta)$, where g is a q-dimensional vector-valued function β, $1 \le q \le p$. For example, we may need to set a joint confidence region for several functions of β. When g is a vector-valued function, we may need different numbers of iterations for different components of g, which is undesirable unless we can reduce the problem to several one-dimensional problems (e.g., applying the Bonferroni method, we may obtain a joint confidence region for θ by taking the product of confidence intervals for components of θ). In the following we discuss some extensions of the one-dimensional results in Theorems 4–6 to the case of vector-valued g.

We shall use the same notation $\hat\theta_w^{(m)} = g(\hat\beta_w^{(m)})$, ∇g, a_k, b_k and $\tau_k^{(m)}$, but now g and $\hat\theta_w^{(m)}$ are q-vectors, ∇g is a $q \times p$ matrix and a_k, b_k and $\tau_k^{(m)}$ are $q \times q$ matrices. Using the delta method, one can show that the result in Theorem 3 still holds [with $N(0, 1)$ replaced by $N(0, I_q)$]. Since $\tau_k^{(m)}$ is a matrix, we need a scalar characteristic of $\tau_k^{(m)}$ to measure the efficiency of $\hat\theta_w^{(m)}$. For example, we may consider a quadratic loss function used frequently in simultaneous estimation:

$$L(\theta, d) = (d - \theta)' Q(\theta) (d - \theta),$$

where $Q(\theta)$ is a $q \times q$ positive definite matrix function continuous at the true parameter θ. The asymptotic risk function of $\hat\theta_w^{(m)}$ is

$$E_A L\big(\theta, \hat\theta_w^{(m)}\big) = E_A \big(\hat\theta_w^{(m)} - \theta\big)' Q(\theta) \big(\hat\theta_w^{(m)} - \theta\big) = \operatorname{trace}\big[Q(\theta) \tau_k^{(m)}/k\big]$$

$$= \big\{ \varphi(m) \operatorname{trace}\big[Q(\theta) a_k\big] + \psi(m) \operatorname{trace}\big[Q(\theta) b_k\big] \big\} \big/ k,$$

where E_A is the expectation under the asymptotic distribution of $\hat{\theta}_w^{(m)} - \theta$. The scalars trace$[Q(\theta)a_k]$ and trace$[Q(\theta)b_k]$ can be consistently estimated by trace$[Q(\hat{\theta}_o)\hat{a}_k]$ and trace$[Q(\hat{\theta}_o)\hat{b}_k]$, where $\hat{a}_k$ and $\hat{b}_k$ are still given by (3.7) and (3.8) (with the square replaced by the vector product). Then the results in Theorems 4–6 can be directly extended to this case. For example, we estimate the optimal number of iterations by

$$\hat{m}_k = \min\left\{ m\colon \Delta(m) \geq \frac{\operatorname{trace}\left[Q(\hat{\theta}_o)\hat{b}_k\right]}{\operatorname{trace}\left[Q(\hat{\theta}_o)\hat{a}_k\right]},\ m = 0, 1, 2, \ldots \right\}.$$

Another widely used measure of efficiency is $|\tau_k^{(m)}|$, the determinant of $\tau_k^{(m)}$, which is also called the generalized asymptotic variance of $\hat{\theta}_w^{(m)}$. We prove some results similar to those in Theorems 4–6.

THEOREM 7. *Suppose that* $a_k \to a > 0$ (*positive definite*) *and* $b_k \to b > 0$.

(i) *There exists an integer* m^* *such that* $\lim_{k\to\infty}(|\tau_k^{(m^*)}|/|\tau_k^{(m)}|) \leq 1$ *for all* m *with equality holding for at most finitely many* m*'s.*

(ii) *Let* $\hat{\tau}_k^{(m)} = \varphi(m)\hat{a}_k + \psi(m)\hat{b}_k$ *and* $\hat{m}_k$ *be a minimizer of* $|\hat{\tau}_k^{(m)}|$, *that is,* $|\hat{\tau}_k^{(\hat{m}_k)}| = \min_m |\hat{\tau}_k^{(m)}|$. *Then*

$$\lim_{k\to\infty} P(\hat{m}_k \in \mathbf{M}) = 1, \tag{3.9}$$

where $\mathbf{M} = \{m\colon \lim_{k\to\infty}(|\tau_k^{(m^*)}|/|\tau_k^{(m)}|) = 1\}$.

PROOF. (i) Since $a > 0$ and $b > 0$, there exists a nonsingular matrix Γ such that

$$\Gamma'[\varphi(m)a + \psi(m)b]\Gamma = \varphi(m)I_q + \psi(m)\Lambda,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_q)$ with $0 < \lambda_1 \leq \cdots \leq \lambda_q$. Then

$$\lim_{k\to\infty} |\tau_k^{(m)}| = |\Gamma|^2 |\varphi(m)I_q + \psi(m)\Lambda| = |\Gamma|^2 \prod_{t=1}^{q} [\varphi(m) + \psi(m)\lambda_t].$$

Note that

$$\begin{aligned}[\varphi(m+1) + \psi(m+1)\lambda_t] - [\varphi(m) + \psi(m)\lambda_t] \\ = [\psi(m) - \psi(m+1)][\Delta(m) - \lambda_t]\end{aligned}$$

and by Lemma 1, $\Delta(m) - \lambda_t$ can change sign only once as m increases. Thus

$$\frac{\varphi(m+1) + \psi(m+1)\lambda_t}{\varphi(m) + \psi(m)\lambda_t} \begin{cases} < 1 & \text{if } m < m(t) \\ > 1 & \text{if } m > m(t) + 1 \end{cases}$$

for an integer $m(t)$, $t = 1, \ldots, q$. Since $\lambda_1 \leq \cdots \leq \lambda_q$, $m(1) \leq \cdots \leq m(q)$. Hence

$$\lim_{k\to\infty} \frac{|\tau_k^{(m+1)}|}{|\tau_k^{(m)}|} \begin{cases} < 1 & \text{if } m < m(1) \\ > 1 & \text{if } m > m(q) + 1. \end{cases}$$

This proves that $\{\lim_{k\to\infty}|\tau_k^{(m)}|,\ m = 0, 1, 2, \ldots\}$ has a minimum at some m^* satisfying $m(1) \le m^* \le m(q) + 1$. Hence the result in (i) holds.

(ii) Since $\varphi(m)$ and $\psi(m)$ are bounded,

$$\hat{\tau}_k^{(m)} \to_p \varphi(m)a + \psi(m)b \quad \text{uniformly in } m.$$

This implies (3.9). □

Thus, in terms of the generalized asymptotic variance, $\hat{m}_k$ in (3.9) can be used to stop the iterative procedure.

4. Asymptotics for the case of $n_i \le 2$. One of the conditions required for the asymptotic results in the previous sections is that for any i,

$$E\left(\sum_{j=1}^{n_i} e_{ij}^2\right)^{-(1+\delta)} \le d \tag{4.1}$$

for some constants δ and d. This condition is satisfied for most error distributions if $n_i \ge 3$ [see Proposition 4.1 in Shao (1989a)]. Often, when $n_i \le 2$, $Eu_i^{-1} = \infty$ and hence (4.1) does not hold. Furthermore, Lemma 1 in Section 3 requires $n_o \ge 3$. Hence the results in the previous sections cannot be applied when $n_i \le 2$. The problem of having $Eu_i^{-1} = \infty$ can be avoided by considering

$$\tilde{u}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} e_{ij}^2 + k^{-1}$$

and

$$\tilde{v}_i = \tilde{v}_i(\hat{\beta}) = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_{ij} - x_{ij}'\hat{\beta}\right)^2 + k^{-1},$$

since $E\tilde{u}_i^{-1}$ exists for any n_i. However, the WLSE with weights $w_i = \tilde{v}_i^{-1}$ is not asymptotically better than the OLSE. More precisely, we have the following result.

THEOREM 8. *Suppose that Assumption* A *holds and* $n_i = n_o = 2$ *for all* i. *Suppose also that the* e_{ij}/σ_i *are i.i.d. with a symmetric density which is positive and continuous at* 0. *Let* $\hat{\beta}_w$ *be given by* (1.2) *with* $w_i = \tilde{v}_i^{-1}$ *and let* $\hat{\beta} = \hat{\beta}_o$ *be the OLSE. Then*

$$\hat{\beta}_w - \beta = \hat{\beta}_o - \beta + o_p(k^{-1/2}). \tag{4.2}$$

PROOF. Under the symmetry condition, when $i_1 \ne i_2$, $\mathbf{e}_{i_1}'X_{i_1}X_{i_2}'\mathbf{e}_{i_2}/\tilde{u}_{i_1}\tilde{u}_{i_2}$ has zero expectation. Let $f(x)$ be the density function of e_{ij}/σ_i. Since f is positive and continuous at 0,

$$E\tilde{u}_i^{-1} \sim \int \frac{cf(x)f(y)}{k^{-1} + x^2 + y^2}\,dx\,dy \sim \int_0^1\int_0^1 \frac{ck}{1 + kx^2 + ky^2}\,dx\,dy = O(\log k).$$

Therefore,

$$E\left(\sum_{i=1}^{k}\tilde{u}_i^{-1}X_i'\mathbf{e}_i\right)^2 = \sum_{i=1}^{k}E\mathbf{e}_i'X_iX_i'\mathbf{e}_i/\tilde{u}_i^2 = O\left(kE\left(e_{11}^2/\tilde{u}_1^2\right)\right) = O(k\log k),$$

where the second equality follows from the boundedness of X_i and σ_i. Consequently,

$$\sum_{i=1}^{k}\tilde{u}_i^{-1}X_i'\mathbf{e}_i = O_p\left((k\log k)^{1/2}\right) = o_p\left(k^{1/2}\log k\right) \tag{4.3}$$

and

$$\sum_{i=1}^{k}\tilde{v}_i^{-1}X_i'\mathbf{e}_i = \sum_{i=1}^{k}\left(\tilde{v}_i^{-1}-\tilde{u}_i^{-1}\right)X_i'\mathbf{e}_i + o_p\left(k^{1/2}\log k\right). \tag{4.4}$$

Note that

$$\begin{aligned}\sum_{i=1}^{k}\left(\tilde{v}_i^{-1}-\tilde{u}_i^{-1}\right)X_i'\mathbf{e}_i = {} & \sum_{i=1}^{k}\left(\tilde{u}_i\tilde{v}_i\right)^{-1}X_i'\mathbf{e}_i\mathbf{e}_i'X_i\left(\hat{\beta}-\beta\right)\\ & -\tfrac{1}{2}\sum_{i=1}^{k}\left(\tilde{v}_i\tilde{u}_i\right)^{-1}X_i'\mathbf{e}_i\left\|X_i\left(\hat{\beta}-\beta\right)\right\|^2.\end{aligned} \tag{4.5}$$

We show that the second term on the right side of (4.5) is negligible. Let α be the set of the i's satisfying

$$\max\{|e_{i1}|,|e_{i2}|\} \geq k^{-1/2}(\log k)^{1/4}. \tag{4.6}$$

Since $\max_{i,j}[x_{ij}'(\hat{\beta}-\beta)]^2 = O_p(k^{-1})$, we may focus on the event of $\max_{i,j}[x_{ij}'(\hat{\beta}-\beta)]^2 \leq k^{-1}(\log k)^{1/2}/4$. Then for each $i\in\alpha$, $\max_{i,j}[x_{ij}'(\hat{\beta}-\beta)]^2 \leq \max_j e_{ij}^2/4$ and

$$\begin{aligned}\tilde{v}_i &\geq k^{-1} + \max_j\left(y_{ij}-x_{ij}'\hat{\beta}\right)^2/2\\ &\geq k^{-1} + \left\{\max_j e_{ij}^2/2 - \max_{i,j}\left[x_{ij}'\left(\hat{\beta}-\beta\right)\right]^2\right\}\Big/2\\ &\geq k^{-1} + \max_j e_{ij}^2/8 \geq \tilde{u}_i/16.\end{aligned}$$

Also

$$\begin{aligned}E\tilde{u}_i^{-3/2} &\sim \int\frac{cf(x)f(y)}{\left(k^{-1}+x^2+y^2\right)^{3/2}}\,dx\,dy\\ &\sim \int_0^1\int_0^1\frac{c\sqrt{k}}{\left(1+kx^2+ky^2\right)^{3/2}}\,d\left(\sqrt{k}\,x\right)d\left(\sqrt{k}\,y\right) = O(k^{1/2}).\end{aligned}$$

Hence

$$\left\|X_i\left(\hat{\beta}-\beta\right)\right\|^2 \times\left\|X_i'\mathbf{e}_i\left(\tilde{u}_i\tilde{v}_i\right)^{-1}\right\| \leq O_p(k^{-1})\tilde{u}_i^{-3/2} = O_p(k^{-1/2}) \tag{4.7}$$

uniformly for $i \in \alpha$. Let α^c be the complement of α. Since $[x'_{ij}(\hat{\beta} - \beta)]^2 \le 2(y_{ij} - x'_{ij}\hat{\beta})^2 + 2e_{ij}^2$,

$$\begin{aligned}(4.8)\qquad \left\| X_i(\hat{\beta} - \beta) \right\|^2 \times \left\| X_i' \mathbf{e}_i (\tilde{u}_i \tilde{v}_i)^{-1} \right\| &= O_p\big(\|\mathbf{e}_i\|/\tilde{u}_i + \|\mathbf{e}_i\|/\tilde{v}_i\big) \\ &\le O_p\big(k\|\mathbf{e}_i\|\big) = O_p\big(k^{1/2}(\log k)^{1/4}\big)\end{aligned}$$

for $i \in \alpha^c$. Let N be the number of elements in α^c. Then N is a random variable with binomial distribution having parameters (k, p_k). The density of e_{ij}/σ_i being continuous at 0 implies that

$$p_k = P\big\{|e_{i1}| < k^{-1/2}(\log k)^{1/4},\ |e_{i2}| < k^{-1/2}(\log k)^{1/4}\big\} = O\big(k^{-1}(\log k)^{1/2}\big).$$

Therefore, $N = O_p((\log k)^{1/2})$ and

$$\begin{aligned}(4.9)\qquad &\sum_{i=1}^{k} (\tilde{u}_i \tilde{v}_i)^{-1} X_i' \mathbf{e}_i \left\| X_i(\hat{\beta} - \beta) \right\|^2 \\ &\quad = \sum_{i \in \alpha} O_p(k^{-1/2}) + \sum_{i \in \alpha^c} O_p\big(k^{1/2}(\log k)^{1/4}\big) \\ &\quad = O_p(k^{1/2}) + O_p\big(k^{1/2}(\log k)^{3/4}\big) = o_p\big(k^{1/2}(\log k)\big).\end{aligned}$$

From (4.3), (4.4), (4.5) and (4.9), we get

$$(4.10)\qquad \sum_{i=1}^{k} \tilde{v}_i^{-1} X_i' \mathbf{e}_i = \sum_{i=1}^{k} (\tilde{u}_i \tilde{v}_i)^{-1} X_i' \mathbf{e}_i \mathbf{e}_i' X_i (\hat{\beta} - \beta) + o_p\big(k^{1/2} \log k\big).$$

Similarly, for $i \in \alpha$,

$$\begin{aligned}\left\| \sum_{i \in \alpha} \big(\tilde{u}_i^{-1} - \tilde{v}_i^{-1}\big) X_i' X_i \right\| &\le \sum_{i \in \alpha} \left[\left\| X_i(\beta - \hat{\beta}) \right\|^2 + 2\big|\mathbf{e}_i' X_i(\hat{\beta} - \beta)\big| X_i' X_i \right] / \tilde{u}_i^2 \\ &\le c\|\hat{\beta} - \beta\|^2 \sum_{i \in \alpha} \tilde{u}_i^{-2} + c\|\hat{\beta} - \beta\| \sum_{i \in \alpha} \tilde{u}_i^{-3/2} = O_p(k),\end{aligned}$$

since $E\tilde{u}_i^{-2} = O(k)$ and $E\tilde{u}_i^{-3/2} = O(k^{1/2})$. We hence obtain

$$\begin{aligned}\left\| \sum_{i=1}^{k} \big(\tilde{u}_i^{-1} - \tilde{v}_i^{-1}\big) X_i' X_i \right\| &= O_p(k) + \left\| \sum_{i \in \alpha^c} \big(\tilde{u}_i^{-1} - \tilde{v}_i^{-1}\big) X_i' X_i \right\| \\ &= O_p(k) + O_p\big(k(\log k)^{1/2}\big) = o_p(k \log k),\end{aligned}$$

because each of $\tilde{u}_i$ or $\tilde{v}_i$ are larger than k^{-1}. From $E\tilde{u}_i^{-2} = O(k)$, we get

$$E\left\| \sum_{i=1}^{k} \big(\tilde{u}_i^{-1} - E\tilde{u}_i^{-1}\big) X_i' X_i \right\|^2 \le c \sum_{i=1}^{k} E\big(\tilde{u}_i^{-1} - E\tilde{u}_i^{-1}\big)^2 = O(k^2),$$

$$\left\| \sum_{i=1}^{k} \big(\tilde{u}_i^{-1} - E\tilde{u}_i^{-1}\big) X_i' X_i \right\| = O_p(k) = o_p(k \log k).$$

Clearly,

$$\left\| \sum_{i=1}^{k} E\tilde{u}_i^{-1} X_i' X_i \right\| \geq ck \log k.$$

Therefore,

$$\hat{\beta}_w - \beta = \left(\sum_{i=1}^{k} \tilde{v}_i^{-1} X_i' X_i \right)^{-1} \sum_{i=1}^{k} \tilde{v}_i^{-1} X_i' \mathbf{e}_i$$

$$= \left(\sum_{i=1}^{k} E\tilde{u}_i^{-1} X_i' X_i \right)^{-1} \sum_{i=1}^{k} (\tilde{u}_i \tilde{v}_i)^{-1} X_i' \mathbf{e}_i \mathbf{e}_i' X_i (\hat{\beta}_o - \beta) + o_p(k^{-1/2}).$$

Using the same technique as that in establishing (4.10), we can show that $(\tilde{u}_i \tilde{v}_i)^{-1} \mathbf{e}_i \mathbf{e}_i'$ in the preceding equation can be replaced by $E\tilde{u}_i^{-1} I_2$. This proves (4.2). □

Result (4.2) indicates that when $n_i = 2$ for all i, the asymptotic variance of $\hat{\beta}_w^{(m)}$ is the same as that of $\hat{\beta}_o$ if $\hat{\beta}_o$ is used as the initial estimator. Therefore, asymptotically $\hat{\beta}_w^{(m)}$ does not improve $\hat{\beta}_o$ for any $m \geq 1$. Despite this asymptotic result, $\hat{\beta}_w^{(m)}$ and $\hat{\beta}_o$ may have different performances for fixed k (see the simulation result in Section 5).

Note that Theorems 1 and 2 do not require that the within-group errors have the same variance. Hence we may combine some groups with $n_i \leq 2$ so that all the new groups contain at least three observations and then apply the adaptive procedure.

In some cases there is auxiliary information about variance pattern. For example, we may consider some physical background: The variances may vary with time, community, batch, and so on. A general rule for combining groups is to combine groups with similar variances since combining groups with the same variance generally increases the efficiency of the WLSE.

Since the adaptive procedure picks the OLSE if the OLSE is better than the WLSE, the adaptive estimator with combined groups is asymptotically as good as the OLSE and improves the OLSE when the heteroscedasticity is severe. In Section 5 we show by simulation that with a correct combining method, the adaptive estimator with combined groups significantly improves the OLSE and the WLSE with noncombined groups.

Another method for improving the OLSE when n_i are very small is the empirical Bayes method. Assuming that the σ_i is random, Hooper (1990) proposed the use of empirical Bayes estimators (EBE). He showed (by asymptotic theory and by simulation) that the EBE improves the Fuller and Rao WLSE $\hat{\theta}_w^{(1)}$, provided that we can correctly specify the type of distribution of the random σ_i. Similarly, the method of combining groups requires some auxiliary information about the σ_i. In the case where $n_i \leq 2$, it is actually hard

to improve the OLSE without having any auxiliary information about the variances.

5. Simulation results. We study by simulation the finite sample bias and root mean square error (rmse) of the OLSE, one-step WLSE [WLSE(1)], two-step WLSE [WLSE(2)] and the adaptive estimator (AWLSE). To illustrate, we force the iterative process to stop after at most two steps so that the adaptive estimator given by (3.6) with $M = 2$ is used. We consider a common mean problem

$$y_{ij} = \mu + e_{ij}, \qquad j = 1, 2, \ldots, n_o, i = 1, 2, \ldots, k, \tag{5.1}$$

where $k = 40$, $n_o = 3$ or 4 and e_{ij}/σ_i are i.i.d. random variables with common $N(0, 1)$ distribution. Without loss of generality, we take $\mu = 0$ in our simulation. Four models with different variance patterns are considered:

MODEL 1 (Homoscedastic model). $\sigma_i = 1$ for all i.

MODEL 2.

i	1–10	11–20	21–30	31–40
σ_i	1	1.5	2.25	3.375

MODEL 3.

i	1–10	11–20	21–30	31–40
σ_i	0.5	1	2	4

MODEL 4. $\{\sigma_i - 0.1, i = 1, \ldots, 40\}$ is a random sample from Beta(0.5, 0.5).

Models 1–4 are arranged in increasing order of heteroscedasticity. To examine the performances of various estimators in the case where $n_o = 2$, we also consider the following model:

MODEL 5. Model (5.1) but $n_o = 2$, $k = 80$. $\{\sigma_i - 0.1, i = 1, \ldots, 40\}$ and $\{\sigma_i - 0.1, i = 41, \ldots, 80\}$ are two ordered random samples from Beta(0.5, 0.5).

Under Model 5, in addition to the OLSE and WLSE, we also consider the estimators based on the data with combined groups and denote the one-step, two-step and adapted estimators by CWLSE(1), CWLSE(2) and CAWLSE, respectively. Two combining methods are used:

1. Randomly combining every two groups;
2. Combining the ith group with the $(40 + i)$th group, $i = 1, 2, \ldots, 40$.

Note that for the second method, we actually assume that we know that the variances for the ith group and the $(40 + i)$th group are close. This allows us

to see what is the best we can do by combining groups. On the other hand, the first method (random combining) is the most inefficient combining method and hence we can see the worst.

The results in Table 1 are based on 10,000 repetitions. In addition to the biases and rmse's of various estimators, Table 1 also shows the number of times (frequency) that AWLSE equals OLSE, WLSE(1) and WLSE(2) in 10,000 repetitions.

The following is a summary of the simulation results:

1. The performance (in terms of rmse and bias) of the adaptive estimator AWLSE is generally good and is always close to the best estimator among the OLSE, one-step WLSE and two-step WLSE. In all cases, the biases are negligible, due to the fact that the OLSE and WLSE are unbiased when the errors have symmetric distributions.
2. In terms of the rmse, the improvement of the adaptive estimator over the OLSE can be as large as 48% when the heteroscedasticity is severe.
3. Except for Model 2 with $n_o = 3$, the adaptive procedure picks the winner with a high probability (in many cases the probability is over 90%).
4. When $n_o = 2$ (Model 5), the WLSE(1) improves the OLSE in terms of rmse (about 20%). This does not conflict with Theorem 8 which is an asymptotic ($k \to \infty$) result. Also, one of the conditions in Theorem 8 is that the ratio $\gamma_k = (\max_i \sigma_i^2)/(\min_i \sigma_i^2)$ is bounded. In Model 5, however, we find $\gamma_k = 118$ and $\gamma_k/k = 1.5$ which is not small. Generally speaking, the WLSE(1) may still improve the OLSE when $n_o = 2$ provided γ_k/k is not relatively small.
5. In Model 5, when an inefficient combining method (randomly combining) is used, the performance of the CWLSE is almost the same as that of the WLSE. On the other hand, when the groups are combined with some prior knowledge (the second method), the CWLSE and CAWLSE have very good performances: They improve the OLSE in terms of rmse (about 45%) and also improve the WLSE.

TABLE 1
Simulation results b = bias $\times 10^4$, r = root mean square error $\times 10^2$, f = frequency distribution of AWLSE

Model 1				
$(n_o, k) = (3, 40)$	OLSE	WLSE(1)	WLSE(2)	AWLSE
b	−4.6118	−3.8018	−3.8918	−4.4189
r	9.1800	10.1963	11.587	9.3039
f	9028	660	312	
$(n_o, k) = (4, 40)$	OLSE	WLSE(1)	WLSE(2)	AWLSE
b	−7.7597	−7.8325	−7.5796	−7.1013
r	7.8718	8.7922	9.7372	7.9627
f	9222	703	75	

TABLE 1
(*Continued*)

Model 2				
$(n_o, k) = (3, 40)$	OLSE	WLSE(1)	WLSE(2)	AWLSE
b	2.6376	−2.0820	−4.1923	−4.7963
r	20.2841	17.8672	18.4310	19.0425
f	3193	4326	2481	
$(n_o, k) = (4, 40)$	OLSE	WLSE(1)	WLSE(2)	AWLSE
b	−11.1978	−10.6913	−9.9395	−13.6378
r	17.5123	14.6337	15.1137	14.9461
f	549	6954	2506	
Model 3				
$(n_o, k) = (3, 40)$	OLSE	WLSE(1)	WLSE(2)	AWLSE
b	8.9063	−0.0209	−3.2834	−1.8743
r	21.0282	14.0737	12.4129	12.5488
f	1	200	9799	
$(n_o, k) = (4, 40)$	OLSE	WLSE(1)	WLSE(2)	AWLSE
b	−8.1214	−7.5647	−6.5526	−6.4433
r	18.2416	10.6878	9.4770	9.4825
f	0	5	9995	
Model 4				
$(n_o, k) = (3, 40)$	OLSE	WLSE(1)	WLSE(2)	AWLSE
b	1.5179	−0.4538	−1.5582	−1.1074
r	6.6878	4.1942	3.5880	3.6581
f	5	112	9883	
$(n_o, k) = (4, 40)$	OLSE	WLSE(1)	WLSE(2)	AWLSE
b	−7.8392	−2.7950	−0.6531	−0.6434
r	5.7416	3.1513	2.7081	2.7100
f	0	3	9997	
Model 5				
Noncombining				
$(n_o, k) = (2, 80)$	OLSE	WLSE(1)	WLSE(2)	
b	−2.7945	−3.4378	−3.6127	
r	5.6685	4.5148	4.0397	
Combining method (1)				
$(n_o, k) = (2, 80)$	OLSE	CWLSE(1)	CWLSE(2)	CAWLSE
b	−2.7945	−1.6845	−0.4240	−1.4609
r	5.6685	4.2820	4.2597	4.3298
f	163	4623	5214	
Combining method (2)				
$(n_o, k) = (2, 80)$	OLSE	CWLSE(1)	CWLSE(2)	CAWLSE
b	−2.7945	−2.7571	−2.4675	−2.5271
r	5.6685	3.4309	3.1166	3.1293
f	0	74	9926	

Acknowledgments. The authors would like to thank an Associate Editor and two referees for their helpful comments and suggestions.

REFERENCES

BOX, G. E. P. and TIAO, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, MA.

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

CARROLL, R. J. and CLINE, D. B. H. (1988). An asymptotic theory for weighted least-squares with weights estimated by replication. *Biometrika* **75** 35–43.

CARROLL, R. J., WU, C. F. J. and RUPPERT, D. (1988). The effect of estimating weights in weighted least squares. *J. Amer. Statist. Assoc.* **83** 1045–1054.

DAVIDIAN, M. and CARROLL, R. J. (1987). Variance function estimation. *J. Amer. Statist. Assoc.* **82** 1079–1091.

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.

HOOPER, P. M. (1990). Iterative weighted least squares estimation in heteroscedastic linear regression models. Preprint.

RAO, C. R. (1970). Estimation of heteroscedastic variances in linear models. *J. Amer. Statist. Assoc.* **65** 161–172.

RAO, J. N. K. (1973). On the estimation of heteroscedastic variances. *Biometrics* **29** 11–24.

SHAO, J. (1989a). Asymptotic distribution of the weighted least squares estimator. *Ann. Inst. Statist. Math.* **41** 365–382.

SHAO, J. (1989b). Jackknifing weighted least squares estimators. *J. Roy. Statist. Soc. Ser. B* **51** 139–156.

DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF WATERLOO
WATERLOO, ONTARIO N2L 3G1
CANADA

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF OTTAWA
OTTAWA, ONTARIO K1N 6N5
CANADA