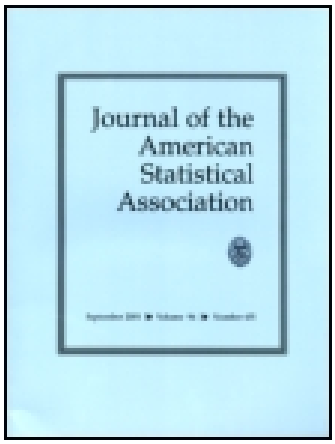

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Iterative Weighted Least Squares Estimation in Heteroscedastic Linear Models

Peter M. Hooper [a]

[a] Department of Statistics and Applied Probability , The University of Alberta , Edmonton , Canada , T6G 2G1

Published online: 20 Dec 2012.

To cite this article: Peter M. Hooper (1993) Iterative Weighted Least Squares Estimation in Heteroscedastic Linear Models, Journal of the American Statistical Association, 88:421, 179-184

To link to this article: http://dx.doi.org/10.1080/01621459.1993.10594309

Iterative Weighted Least Squares Estimation in Heteroscedastic Linear Models

PETER M. HOOPER*

This article addresses the problem of choosing weights for iterative weighted least squares estimation in heteroscedastic linear models. An asymptotically optimal method for determining weights at each iteration is derived under a Bayesian model for the variances. The method uses a compromise between model-based and model-free variance estimates. Consider a heteroscedastic linear regression model in which responses are grouped so that the variance is constant within each group. Let β denote the vector of regression parameters and let θ denote a vector of parameters determining a prior distribution for the variances. Iterative weighted least squares estimators are defined as follows. Given estimates $\hat{\beta}$ and $\hat{\theta}$, calculate a weight for the ith group as a function of $\hat{\theta}$, the values in the ith group of the predictor variables, and the average of the squared residuals from the estimated mean responses in the ith group. Given weights, calculate the weighted least squares estimate $\hat{\beta}$ and a new estimate $\hat{\theta}$. Continue until $\hat{\beta}$ converges. We derive the asymptotically optimal weight function under an inverse gamma model for the variances. The resulting weights have a simple form. At each iteration the inverse weight for the ith group is a weighted average of the average squared residual and a variance estimate based on the inverse gamma model.

KEY WORDS: Empirical Bayes; Robustness; Variance estimation.

1. INTRODUCTION

Consider a heteroscedastic linear model with responses arranged in k groups and variances constant within each group:

$$y_{ij} = \mathbf{x}_{ij}^T\beta + e_{ij}, \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, k, \quad (1.1)$$

where the $\mathbf{x}_{ij}$ are known $p \times 1$ vectors, β is an unknown $p \times 1$ vector of parameters, and the errors e_{ij} have mean 0 and unknown variances σ_i^2. The groups might be formed by arranging the $\mathbf{x}_{ij}$ into small clusters. Groups of size $n_i = 1$ are allowed. A popular strategy for estimating β is to first obtain estimates $\hat{\sigma}_i^2$ of the variances and then apply weighted least squares (WLS) using weights $1/\hat{\sigma}_i^2$. This strategy is motivated by the fact that the weights $1/\sigma_i^2$ are optimal when the errors are normally distributed. When groups are determined by replicates, so that $\mathbf{x}_{ij} = \mathbf{x}_i$ and $n_i \geq 2$, the simplest choice for $\hat{\sigma}_i^2$ is the sample variance $(n_i - 1)^{-1} \sum (y_{ij} - \bar{y}_i)^2$. Carroll and Cline (1988) showed that this choice yields highly inefficient estimates for β when the number of replicates n_i is small, the typical situation. Thus more sophisticated methods of variance estimation are usually needed.

One approach models the variance as a function of the mean and/or the covariates (see Carroll and Ruppert 1988). Another approach uses variance estimates of the form

$$\hat{v}_i(\tilde{\beta}) = n_i^{-1} \sum (y_{ij} - \mathbf{x}_{ij}^T\tilde{\beta})^2, \quad (1.2)$$

where $\tilde{\beta}$ is an estimate of β. We will refer to these approaches as model-based and model-free, where in this context "model" refers to a model for the variances. Both approaches use a model for the means. Among the model-based variance estimates, we include those based on the model assuming that all variances are equal.

Fuller and Rao (1978) studied the WLS estimator $\hat{\beta}$ with weights $\hat{w}_i = 1/\hat{v}_i(\tilde{\beta})$, where $\tilde{\beta}$ is the ordinary least squares (OLS) estimator. Shao (1992) considered a modification of $\hat{\beta}$ that replaces the $\hat{v}_i(\tilde{\beta})$ with empirical Bayes estimates of the variances. These two-step estimators perform well when the variances are not too dispersed but are less efficient under more severe heteroscedasticity. This is due to the reduced efficiency of the OLS estimator resulting in poorer variance estimators. Under severe heteroscedasticity, efficiency is improved by iterating the procedure: Replace $\tilde{\beta}$ by $\hat{\beta}$ in (1.2) to obtain new weights, then compute a new WLS estimator $\hat{\beta}$. The maximum likelihood estimator (MLE)—under a normal-theory model where nothing is assumed about the σ_i^2—is obtained by continuing iteration until $\hat{\beta}$ converges. Unfortunately, the MLE is much less efficient than the two-step estimators under mild to moderate heteroscedasticity.

The poor performance of the MLE seems to be the result of a feedback effect caused by small changes in $\hat{v}_i$ near 0. Variances for some groups are underestimated, and these groups are given greater weight in the subsequent estimate of β, tending to produce even smaller variance estimates in the next step. The present work began with an attempt to improve the MLE. Simulation studies suggested that the feedback effect is reduced by shrinking large weights toward 0 at each iteration; for example, by adding a positive constant to each variance estimate. This led to the related idea of combining model-free and model-based variance estimates. Weights of the form $\hat{w}_i = (n_i + \hat{\gamma})/(n_i\hat{v}_i + \hat{\gamma}\hat{\tau}_i)$, where $\hat{\tau}_i$ is a model-based variance estimate, are shown to be asymptotically optimal under a Bayesian model for the variances.

Carroll and Ruppert (1982) recommended using model-based variance estimators combined with M estimators to obtain iterative weighted estimators that are both efficient and robust. Our weighted estimators should have similar properties. M estimators can be calculated by iterative WLS with weights related to the ψ function (see Street, Carroll, and Ruppert 1988). A key idea in robust estimation is that observations appearing to be less reliable (outliers) should be given smaller weight in the estimate. More generally we

* Peter M. Hooper is Associate Professor, Department of Statistics and Applied Probability, The University of Alberta, Edmonton, Canada, T6G 2G1. Part of this work was conducted during a sabbatical leave at the Department of Statistics and Actuarial Science, University of Waterloo. The research was partially supported by the Natural Sciences and Engineering Research Council of Canada. The author is grateful to two referees for suggestions that led to substantial improvements in the results.

can attempt to form groups of observations thought to be equally reliable a priori and then downweight groups that appear less reliable given the data. The weights defined previously behave like $(1 + n_i/\hat{\gamma})/\hat{\tau}_i$ when $\hat{v}_i$ is small and like $(1 + \hat{\gamma}/n_i)/\hat{v}_i$ when $\hat{v}_i$ is large, and thus exploit the variance model to improve efficiency while providing robustness against outliers.

This article is organized as follows. The asymptotic theory is developed in Section 2, and the optimal estimators are derived in Section 3. The asymptotic variances of several estimators are compared in Section 4, and some finite-sample comparisons are reported in Section 5. All results are based on an assumption that errors are distributed symmetrically about 0. The effects of departures from this assumption are briefly considered in Section 6. Proofs are given in the Appendix.

2. ASYMPTOTIC THEORY

In this section we derive the asymptotic distribution of a general iterative WLS estimator within an empirical Bayes framework. The derivation is similar to that of Theorem 3 in Carroll and Cline (1988). Let $\mathbf{X}_i$ be the $n_i \times p$ matrix with jth row x_{ij}^T and write (1.1) in the form

$$\mathbf{y}_i = \mathbf{X}_i\beta + \mathbf{e}_i. \tag{2.1}$$

Define the *average squared errors* as $v_i = n_i^{-1}\|\mathbf{e}_i\|^2 = n_i^{-1} \times \sum e_{ij}^2$.

In the asymptotic theory we model the $\mathbf{X}_i$ as independent and identically distributed random matrices. This is a convenient device for specifying the behavior of the sequence $\{\mathbf{X}_i\}$. Our main assumptions concern the conditional distribution of $\mathbf{e}_i$ given $\mathbf{X}_i$. No assumptions are made about the distribution of $\mathbf{X}_i$ beyond a few regularity conditions. The theory is applicable in applications with nonrandom $\mathbf{X}_i$.

We assume that $\{(\mathbf{y}_i, \mathbf{X}_i, n_i, \mathbf{e}_i, v_i), i = 1, 2, \cdots\}$ is a sequence of independent and identically distributed random vectors. To avoid writing subscripts, let $(\mathbf{y}, \mathbf{X}, n, \mathbf{e}, v)$ denote a random vector with the same distribution. Let both n and $\operatorname{tr} \mathbf{X}^T\mathbf{X}$ be bounded above. The conditional error distribution is assumed to satisfy

$$E\{\mathbf{e}|\mathbf{X}, v\} = 0 \quad \text{and} \quad E\{\mathbf{e}\mathbf{e}^T|\mathbf{X}, v\} = v\mathbf{I}_n. \tag{2.2}$$

Condition (2.2) holds if the conditional distribution of $\mathbf{e}$ given $\mathbf{X}$ is invariant under the group of permutations and sign changes of the entries in $\mathbf{e}$.

Although it is possible to develop the theory purely in terms of the average squared errors, it is helpful to introduce random variables σ_i^2, distributed jointly with $(\mathbf{X}_i, \mathbf{e}_i)$ so that $\mathcal{L}(e_{ij}|\mathbf{X}_i, \sigma_i^2)$ has mean 0 and variance σ_i^2. We can then interpret the model from a Bayesian perspective as incorporating prior information about the conditional variances σ_i^2. There are in general many ways to define the σ_i^2; for example, $\sigma_i^2 = E\{e_{ij}^2|\mathbf{X}_i\}$, assuming this is finite. If the conditional distribution of e_{ij} given $\mathbf{X}_i$ is a scale mixture of normals then we define σ_i^2 so that the conditional distribution of e_{ij} given $(\mathbf{X}_i, \sigma_i^2)$ is normal; that is, we have $\mathcal{L}(\mathbf{e}|\mathbf{X}, \sigma^2) = N_n(0, \sigma^2\mathbf{I}_n)$ and hence $\mathcal{L}(v|\mathbf{X}, \sigma^2) = (\sigma^2/n)\chi_n^2$. We will refer to this as a normal-mixture model. In a normal-mixture model the weights $1/\sigma_i^2$ are the gold standard for WLS. In the more general setting, however, the $(1/\sigma_i^2)$-WLS estimator can be less efficient than more robust estimators. In comparison, the weights $1/v_i$ always perform well under condition (2.2). To motivate the choice of weights in this context, it seems preferable to regard the model-free variance estimate (1.2) as an estimate of v_i rather than of σ_i^2.

We consider weights that are functions of $\hat{v}_i$, $\mathbf{X}_i$, and $\hat{\theta}$, where θ is a $q \times 1$ parameter vector determined by the conditional distribution of v given $\mathbf{X}$. Let w be a real-valued function defined on the set $(0, \infty) \times \mathcal{M} \times R^q$, where $\mathcal{M}$ is the set of matrices with p columns. We assume that w is continuously differentiable in its first and third arguments and denote the partial derivatives by w' and $\mathbf{w}_\theta$. For conciseness we also let w, w', and $\mathbf{w}_\theta$ denote the random variables and the random vector $w(v, \mathbf{X}, \theta)$, $w'(v, \mathbf{X}, \theta)$, and $\mathbf{w}_\theta(v, \mathbf{X}, \theta)$. We assume that $E\{|w|\}$, $E\{vw^2\}$, $E\{v|w'|\}$, and $E\{v\|\mathbf{w}_\theta\|\}$ are finite and that $E\{w\mathbf{X}^T\mathbf{X}\}$ is nonsingular. The restriction $\Pr\{n \geq 3\} = 1$ is usually required for the function $w = 1/v$ because, under a normal-mixture model, $E\{1/v|n\}$ is infinite for $n \leq 2$. No such restriction is needed for weight functions such as (3.2), where w, vw, vw', and $v\|\mathbf{w}_\theta\|$ are bounded functions of $(v, \mathbf{X})$ for each θ.

For p vectors $\mathbf{b}$ and q vectors $\mathbf{a}$, write $\hat{v}_i(\mathbf{b}) = n_i^{-1}\|\mathbf{y}_i - X_i\mathbf{b}\|^2$ and $\hat{w}_i(\mathbf{b}, \mathbf{a}) = w(\hat{v}_i(\mathbf{b}), \mathbf{X}_i, \mathbf{a})$. Let $\hat{\mathbf{w}}_{i\beta}(\mathbf{b}, \mathbf{a})$ and $\hat{\mathbf{w}}_{i\theta}(\mathbf{b}, \mathbf{a})$ denote the vectors of partial derivatives of $\hat{w}_i(\mathbf{b}, \mathbf{a})$ with respect to $\mathbf{b}$ and $\mathbf{a}$. We assume the following smoothness conditions: For each $c_1 > 0$ there exists $c_2 > 0$ such that

$$E[\sup\{|\hat{w}_i(\mathbf{b}, \mathbf{a}) - \hat{w}_i(\beta, \theta)| : \|\mathbf{b} - \beta\|^2 + \|\mathbf{a} - \theta\|^2 \leq c_2\}] \leq c_1,$$

$$E[\sup\{\|\mathbf{e}_i\|\,\|\hat{\mathbf{w}}_{i\beta}(\mathbf{b}, \mathbf{a}) - \hat{\mathbf{w}}_{i\beta}(\beta, \theta)\| : \|\mathbf{b} - \beta\|^2 + \|\mathbf{a} - \theta\|^2 \leq c_2\}] \leq c_1.$$

and

$$E[\sup\{\|\mathbf{e}_i\|\,\|\hat{\mathbf{w}}_{i\theta}(\mathbf{b}, \mathbf{a}) - \hat{\mathbf{w}}_{i\theta}(\beta, \theta)\| : \|\mathbf{b} - \beta\|^2 + \|\mathbf{a} - \theta\|^2 \leq c_2\}] \leq c_1. \tag{2.3}$$

Let $\tilde{\beta}_k$ and $\tilde{\theta}_k$ be estimators based on $(\mathbf{y}_i, \mathbf{X}_i)$, $i = 1, 2, \ldots, k$, and assume that both are $\sqrt{k}$-consistent; that is, $(\tilde{\beta}_k, \tilde{\theta}_k) = (\beta, \theta) + O_p(1/\sqrt{k})$. Define

$$\mathbf{A} = E\{w\mathbf{X}^T\mathbf{X}\},$$

$$\mathbf{B} = E\{vw^2\mathbf{X}^T\mathbf{X}\},$$

$$\mathbf{C} = E\{(-2/n)vw'\mathbf{X}^T\mathbf{X}\},$$

and

$$\hat{\beta}_k = \left(\sum_{i=1}^k \hat{w}_i(\tilde{\beta}_k, \tilde{\theta}_k)\mathbf{X}_i^T\mathbf{X}_i\right)^{-1} \sum_{i=1}^k \hat{w}_i(\tilde{\beta}_k, \tilde{\theta}_k)\mathbf{X}_i^T\mathbf{y}_i. \tag{2.4}$$

Theorem 1. Given assumptions (2.1)–(2.4), we have

$$\sqrt{k}(\hat{\beta}_k - \beta) = \mathbf{A}_k^{-1}\{\mathbf{b}_k + \mathbf{C}_k\sqrt{k}(\tilde{\beta}_k - \beta) + \mathbf{D}_k\sqrt{k}(\tilde{\theta}_k - \theta)\}, \tag{2.5}$$

where $\mathbf{A}_k \to_p \mathbf{A}$, $\mathbf{C}_k \to_p \mathbf{C}$, $\mathbf{D}_k \to_p 0$, and $\mathbf{b}_k = (1/\sqrt{k}) \times \sum \hat{w}_i(\beta, \theta)\mathbf{X}_i^T\mathbf{e}_i$. Furthermore, if $\hat{\beta}_k = \tilde{\beta}_k$ and $\mathbf{A} - \mathbf{C}$ is nonsingular, then

$$\sqrt{k}(\hat{\beta}_k - \beta) \to_d N_p(0, (\mathbf{A} - \mathbf{C})^{-1}\mathbf{B}(\mathbf{A} - \mathbf{C})^{-1}). \tag{2.6}$$

We apply Theorem 1 to iterative weighted estimators defined as follows. The algorithm requires specification of upper bounds c_β and c_θ on the number of times β and θ are estimated.

Algorithm 1.

Set count $c = 0$ and compute preliminary estimates $\tilde{\beta}$ and $\tilde{\theta}$.
Do until $c = c_\beta$ or $\hat{\beta}$ converges:
  Compute weights $\hat{w}_i = w(\hat{v}_i(\tilde{\beta}), \mathbf{X}_i, \tilde{\theta})$, $i = 1, \ldots, k$.
  Compute the weighted estimate $\hat{\beta} = (\sum \hat{w}_i \mathbf{X}_i^T \mathbf{X}_i)^{-1} \times \sum \hat{w}_i \mathbf{X}_i^T \mathbf{y}_i$.
  Replace $\tilde{\beta}$ by $\hat{\beta}$.
  Set $c = c + 1$.
  If $c < c_\theta$, then replace $\tilde{\theta}$ with a new estimate $\hat{\theta}$.
Repeat.

Theorem 1 shows that the first-order behavior of $\hat{\beta}$ does not depend on c_θ. If w is a function of $\mathbf{X}$ and θ only, so that $w' = 0$ and $\mathbf{C} = 0$, then the first-order behavior of $\hat{\beta}$ is also the same for all $c_\beta \geq 1$. This is a well-known result (see, for example, Carroll and Ruppert (1988, th. 2.1). Carroll, Wu, and Ruppert (1988) showed that the second-order behavior for $c_\beta = 1$ differs from that for $c_\beta \geq 2$. If $\mathbf{C} \neq 0$, then the first-order behavior of $\hat{\beta}$ can depend strongly on c_β. When applying (2.6) we will assume without proof that the algorithm converges for $c_\beta = \infty$. The following argument indicates the rate of convergence when k is large. Starting with a $\sqrt{k}$-consistent estimator $\hat{\beta}^{(0)}$, define $\hat{\beta}^{(m)}$ by (2.4) with $\tilde{\beta} = \hat{\beta}^{(m-1)}$, $m = 1, 2, \ldots$. Apply (2.5) m times and, in each application, replace $\mathbf{A}_k$, $\mathbf{C}_k$, and $\mathbf{D}_k$ by the limits $\mathbf{A}$, $\mathbf{C}$, and 0. We obtain $\sqrt{k}(\hat{\beta}^{(m)} - \beta) \cong (\mathbf{A} - \mathbf{C})^{-1}\mathbf{b}_k + (\mathbf{A}^{-1}\mathbf{C})^m\{\sqrt{k}(\hat{\beta}^{(0)} - \beta) - (\mathbf{A} - \mathbf{C})^{-1}\mathbf{b}_k\}$, so the rate of convergence depends on the eigenvalues of $\mathbf{A}^{-1}\mathbf{C}$.

It can be shown that asymptotic normality (2.6) holds conditionally, given $\{(v_i, \mathbf{X}_i)\}$, subject to reasonable conditions on this sequence. In the expression for the covariance matrix, expectations are replaced with limits of averages, such as $\mathbf{A} = \lim(1/k) \sum w(v_i, \mathbf{X}_i, \theta)\mathbf{X}_i^T \mathbf{X}_i$. This conditional framework is useful for constructing confidence intervals; that is, standard errors obtained by substituting sample quantities for $\mathbf{A} - \mathbf{C}$ and $\mathbf{B}$ can be interpreted as estimates of variability conditioned on $\{(v_i, \mathbf{X}_i)\}$. Assumptions about the conditional distribution of v given $\mathbf{X}$ help motivate the choice of weights, but coverage probabilities of confidence intervals should be robust against departures from these assumptions. Classical normal-theory inference for $k = 1$ is conditioned on $(\mathbf{X}, v)$; that is, the null distributions of the t and F statistics are determined by the fact that, given $(\mathbf{X}, v)$, the error vector $\mathbf{e}$ is distributed uniformly on a sphere of radius $\sqrt{nv}$. Cohen and Sackrowitz (1989) studied the problem of combining interblock and intrablock information to test hypotheses about treatment effects. Their model is equivalent, after an orthogonal transformation, to that considered here but with normal errors, $k = 2$, and n_i fairly large. They obtained exact tests by conditioning on $(\mathbf{X}_1, v_1, \mathbf{X}_2, v_2)$.

A point of potential confusion should be mentioned. If the variance is modeled as a function of the mean, then we take θ to consist of β as well as other parameters. In the algorithm two different estimates of β can then appear in the expression for $\hat{w}_i$: the one in $\hat{v}_i$ that is updated in each cycle and the other in $\hat{\theta}$ that is held fixed after c_θ cycles.

3. OPTIMAL WEIGHTS

The optimal weight function is determined by the conditional distribution of v given $\mathbf{X}$. If the conditional distribution is assumed to lie in a parametric family indexed by a parameter vector θ, then Theorem 1 shows that asymptotic optimality is not affected when θ is replaced by a $\sqrt{k}$-consistent estimator. We assume initially that the conditional distribution is known and suppress the argument θ in the weight function.

The following regularity conditions are assumed. Suppose that the conditional distribution of v given $\mathbf{X}$ is absolutely continuous with density $p(v|\mathbf{X})$, that for each $\mathbf{X}$ the density is positive for all v, and that the density is continuously differentiable with respect to v. Suppose that w and p satisfy $E\{|vw(v, \mathbf{X})(\partial/\partial v)\log p(v|\mathbf{X})|\} < \infty$, $vw(v, \mathbf{X})p(v|\mathbf{X}) \to 0$ as $v \to 0$ or ∞ for each $\mathbf{X}$. Recall the usual partial ordering on covariance matrices: $\mathbf{\Sigma}_1 \leq \mathbf{\Sigma}_2$ if $\mathbf{a}^T\mathbf{\Sigma}_1\mathbf{a} \leq \mathbf{a}^T\mathbf{\Sigma}_2\mathbf{a}$ for all vectors $\mathbf{a}$.

Theorem 2. Under the assumptions of Theorem 1 and the regularity conditions stated previously, the weight function w minimizing the asymptotic covariance matrix in (2.6) is $w_0(v, \mathbf{X}) = n^{-1}\{(n-2)/v - 2(\partial/\partial v)\log(p(v|\mathbf{X}))\}$, and the minimum covariance matrix is $(\mathbf{A} - \mathbf{C})^{-1}\mathbf{B}(\mathbf{A} - \mathbf{C})^{-1} = [E\{vw_0^2\mathbf{X}^T\mathbf{X}\}]^{-1}$.

The optimal weight function can also be derived via maximum likelihood estimation. For a given weight function w, the iterative WLS estimate minimizes $\sum l(\hat{v}_i(\hat{\beta}), \mathbf{X}_i)$, where $l(v, \mathbf{X}) = \int^v nw(u, \mathbf{X})\, du$. The l function corresponding to w_0 is (up to an additive constant) $l_0(v, \mathbf{X}) = (n-2)\log(v) - 2\log(p(v|\mathbf{X}))$. If the conditional distribution of $\mathbf{e}$ given $(\mathbf{X}, v)$ is the uniform distribution on the sphere of radius $\sqrt{nv}$, then $\sum l_0(v_i, \mathbf{X}_i)$ is (up to an additive constant) -2 times the conditional log-likelihood, given $\mathbf{X}_1, \ldots, \mathbf{X}_k$.

The following result gives a simple expression for the optimal weight function under a normal-mixture model.

Theorem 3. If the assumptions of Theorem 2 hold and $\mathcal{L}(v|\mathbf{X}, \sigma^2) = (\sigma^2/n)\chi_n^2$, then the optimal weight function is $w_0(v, \mathbf{X}) = E\{1/\sigma^2|v, \mathbf{X}\}$.

Within the family of normal-mixture models, the inverse gamma distribution for the conditional variances leads to an easily interpreted weight function. Suppose that we have

$$\mathcal{L}(1/\sigma^2|\mathbf{X}) = (\gamma\tau)^{-1}\chi_\gamma^2$$

$$\text{and hence} \quad \mathcal{L}(v|\mathbf{X}) = \tau F_{n,\gamma}, \quad (3.1)$$

where γ and τ are positive parameters, with τ a function of $\mathbf{X}$ and/or $\mathbf{X}\beta$. Applying either Theorem 2 or Theorem 3 yields the optimal weight function

$$w_0(v, \mathbf{X}) = \frac{n + \gamma}{nv + \gamma\tau}. \quad (3.2)$$

The inverted weight $1/\hat{w}_i$ is then a weighted average of the model-free variance estimate $\hat{v}_i$ and the model-based variance estimate $\hat{\tau}_i$.

The parameters $\log(\tau)$ and $1/\gamma$ are location and dispersion parameters for the conditional distribution of $\log(\sigma^2)$ given

$\mathbf{X}$. We will derive estimators under the log-linear model, $\log(\tau_i) = \mathbf{u}_i^T\boldsymbol{\eta}$, where $\boldsymbol{\eta}$ is an unknown $r \times 1$ parameter vector and $\mathbf{u}_i$ is an $r \times 1$ vector determined by $\mathbf{X}_i$ and/or $\mathbf{X}_i\boldsymbol{\beta}$. The simplest example is the constant-variance model: $\log(\tau_i) = \eta$. A second example is the power-of-mean model: $\log(\tau_i) = \eta_1 + \eta_2\log(n_i^{-1} \sum \mathbf{x}_{ij}^T\boldsymbol{\beta})$.

Parameters for the constant-variance model can be estimated by the method of Hui and Berger (1983), a combination of method of moments applied to the v_i and maximum likelihood. We prefer to apply the method of moments to the $\log(v_i)$, because this approach generalizes more easily to log-linear models and yields tractable estimators for all parameters. For $a > 0$ define $\mu_{LG}(a) = E\{\log g\}$ and $\sigma^2_{LG}(a) = \operatorname{var}\{\log g\}$, where $\mathcal{L}(g) = (1/a)\chi^2_a$. Note that μ_{LG} is an increasing function and σ^2_{LG} is a decreasing function, and that both tend to 0 as $a \to \infty$. Put $z_i = \log(v_i) - \mu_{LG}(n_i)$, $\mathbf{z} = (z_1, \ldots, z_k)^T$, and $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_k)^T$. We assume that $\mathbf{U}$ has full rank r and that the column space of $\mathbf{U}$ contains $\mathbf{1}_k = (1, \ldots, 1)^T$. Put $\mathbf{H} = (h_{ij}) = \mathbf{U}(\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T$ and $s_z^2 = (k-r)^{-1}\mathbf{z}^T(\mathbf{I}_k - \mathbf{H})\mathbf{z}$. Under (3.1) we have $E\{\mathbf{z}|\mathbf{X}_1, \ldots, \mathbf{X}_k\} = \mathbf{U}\boldsymbol{\eta} - \mu_{LG}(\gamma)\mathbf{1}_k$ and $E\{s_z^2 \,|\mathbf{X}_1, \ldots, \mathbf{X}_k\} = \sigma^2_{LG}(\gamma) + (k-r)^{-1} \sum (1 - h_{ii})\sigma^2_{LG}(n_i)$. Given an estimate $\hat{\boldsymbol{\beta}}$, we obtain estimates for γ, $\boldsymbol{\eta}$, and τ_i by replacing $\boldsymbol{\beta}$ with $\hat{\boldsymbol{\beta}}$ in the definitions of $\mathbf{z}$ and $\mathbf{U}$ and solving

$$\sigma^2_{LG}(\hat{\gamma}) = \min[\max\{s_z^2 - (k-r)^{-1} \times \sum (1 - h_{ii})\sigma^2_{LG}(n_i),\ \sigma^2_{LG}(\gamma_{ub})\},\ \sigma^2_{LG}(\gamma_{lb})],$$

$$\hat{\boldsymbol{\eta}} = (\mathbf{U}^T\mathbf{U})^{-1}\mathbf{U}^T(\mathbf{z} + \mu_{LG}(\hat{\gamma})\mathbf{1}_k), \quad \text{and} \quad \hat{\tau}_i = \exp(\mathbf{u}_i^T\hat{\boldsymbol{\eta}}), \tag{3.3}$$

where $0 < \gamma_{lb} < \gamma_{ub} < \infty$ are a priori lower and upper bounds on γ. To make the estimates less sensitive to small changes in $\hat{v}_i$ near 0, we add a small positive constant ε_k to $\hat{v}_i$ before computing the logarithm.

If $\hat{\boldsymbol{\beta}}$ is $\sqrt{k}$-consistent, then, under the general model (2.2) with additional regularity conditions, it can be shown that $\hat{\gamma}$ and $\hat{\boldsymbol{\eta}}$ are $\sqrt{k}$-consistent estimators for suitably defined parameters γ^* and $\boldsymbol{\eta}^*$. Theorem 1 can be applied with $\boldsymbol{\theta}$ consisting of γ^*, $\boldsymbol{\eta}^*$, and $\boldsymbol{\beta}$. If (3.1) holds, $\varepsilon_k = O(1/\sqrt{k})$ and $\gamma_{lb} < \gamma < \gamma_{ub}$, then $(\gamma^*, \boldsymbol{\eta}^*) = (\gamma, \boldsymbol{\eta})$.

The estimate $\hat{\gamma}$ can be interpreted as a measure of overdispersion in the $\hat{v}_i$ beyond that expected under the normal-theory log-linear variance model. Smaller values of $\hat{\gamma}$ result in greater emphasis on the model-free variance estimates $\hat{v}_i$ in the weights and hence greater robustness against deficiencies in the model. The finite upper bound on γ ensures a minimum level of robustness. The positive lower bound limits the sensitivity of the weights to small changes in $\hat{v}_i$ near 0.

4. COMPARISON OF ASYMPTOTIC VARIANCES

As a benchmark for comparison, we define the *average squared error estimator* $\hat{\boldsymbol{\beta}}_{ASE}$ by the weights $\hat{w}_i = 1/v_i$. If the v_i were actually known, then one would not necessarily use $\hat{\boldsymbol{\beta}}_{ASE}$, because in some situations $\boldsymbol{\beta}$ would be uniquely determined. The asymptotic covariance matrix of $\sqrt{k}\hat{\boldsymbol{\beta}}_{ASE}$ is $\boldsymbol{\Sigma}_{ASE} = [E\{v^{-1}\mathbf{X}^T\mathbf{X}\}]^{-1}$. If $E\{1/v\}$ is infinite, as in a normal-mixture model with $\Pr\{n \le 2\} > 0$, then $\hat{\boldsymbol{\beta}}_{ASE}$ usually converges at a rate faster than $1/\sqrt{k}$. The *conditional variance estimator* $\hat{\boldsymbol{\beta}}_{CV}$, defined by the weights $\hat{w}_i = 1/\sigma_i^2$, provides a second benchmark useful under normal-mixture models. Using $E\{v|\mathbf{X}, \sigma^2\} = \sigma^2$, we obtain the asymptotic covariance matrix of $\sqrt{k}\hat{\boldsymbol{\beta}}_{CV}$ as $\boldsymbol{\Sigma}_{CV} = [E\{\sigma^{-2}\mathbf{X}^T\mathbf{X}\}]^{-1}$. If n is constant, v/σ^2 is distributed independently of $(\mathbf{X}, \sigma^2)$, and $\boldsymbol{\Sigma}_{ASE}$ is positive definite, then $\boldsymbol{\Sigma}_{CV} = E\{\sigma^2/v\}\boldsymbol{\Sigma}_{ASE}$. By Jensen's inequality, $E\{\sigma^2/v\} \ge 1$ with equality only if $\sigma^2 = v$, so $\hat{\boldsymbol{\beta}}_{ASE}$ is more efficient than $\hat{\boldsymbol{\beta}}_{CV}$. Under a normal-mixture model, we have $E\{\sigma^2/v\} = n/(n-2)$ for $n \ge 3$.

The OLS estimator is denoted by $\hat{\boldsymbol{\beta}}_{\text{OLS}}$. The asymptotic covariance matrix of $\sqrt{k}\hat{\boldsymbol{\beta}}_{\text{OLS}}$ is $\boldsymbol{\Sigma}_{\text{OLS}} = [E\{\mathbf{X}^T\mathbf{X}\}]^{-1}E\{v\mathbf{X}^T\mathbf{X}\}[E\{\mathbf{X}^T\mathbf{X}\}]^{-1}$.

The Fuller–Rao estimator $\hat{\boldsymbol{\beta}}_{FR}$ uses weights $\hat{w}_i = 1/\hat{v}_i(\hat{\boldsymbol{\beta}}_{\text{OLS}})$. An application of (2.5) shows that $\sqrt{k}\hat{\boldsymbol{\beta}}_{FR}$ has asymptotic covariance matrix $\boldsymbol{\Sigma}_{FR} = \boldsymbol{\Sigma}_{ASE}\{\boldsymbol{\Sigma}_{ASE}^{-1} + \mathbf{2C} + \mathbf{C}\boldsymbol{\Sigma}_{\text{OLS}}\mathbf{C}\}\boldsymbol{\Sigma}_{ASE}$, where $\mathbf{C} = E\{(2/n)v^{-1}\mathbf{X}^T\mathbf{X}\}$. If $n \ge 3$ is constant, then $\mathbf{C} = (2/n)\boldsymbol{\Sigma}_{ASE}^{-1}$ and $\boldsymbol{\Sigma}_{FR} = (1 + 4/n)\boldsymbol{\Sigma}_{ASE} + (4/n^2)\boldsymbol{\Sigma}_{\text{OLS}}$.

The conditional normal-theory MLE $\hat{\boldsymbol{\beta}}_{ML}$ is given by Algorithm 1 with $w = 1/v$ and $c_\beta = \infty$. We take $\hat{\boldsymbol{\beta}}_{\text{OLS}}$ as the initial estimator, but other estimators can also be used. Assuming that the iterative procedure converges to a consistent solution, the asymptotic distribution is given by (2.6) with $\mathbf{A} = \mathbf{B} = \boldsymbol{\Sigma}_{ASE}^{-1}$ and $\mathbf{C} = E\{(2/n)v^{-1}\mathbf{X}^T\mathbf{X}\}$. If $n \ge 3$ is constant, then $\sqrt{k}\hat{\boldsymbol{\beta}}_{ML}$ has asymptotic covariance matrix $\boldsymbol{\Sigma}_{ML} = \{n/(n-2)\}^2\boldsymbol{\Sigma}_{ASE}$. Note that $\hat{\boldsymbol{\beta}}_{ML}$ is more efficient than $\hat{\boldsymbol{\beta}}_{CV}$ when $E\{\sigma^2/v\} > \{n/(n-2)\}^2$.

The expressions for $\boldsymbol{\Sigma}_{FR}$ and $\boldsymbol{\Sigma}_{ML}$ may be compared with those for $\mathbf{S}_{EL}^{-1}$ and $\mathbf{S}_{ML}^{-1}$ in Theorem 4 and Corollary 3 of Carroll and Cline (1988). They assumed that n (=m in their notation) is constant and at least 3 and, in effect, that v/σ^2 is distributed independently of $(\mathbf{X}, \sigma^2)$. Their expressions use $\mathbf{S}_{\text{WLS}}^{-1} = \boldsymbol{\Sigma}_{CV}$ and the constants $\eta_{01} = E\{\sigma^2/v\}$, $\eta_{21} = 1/n$, and $\eta_{22} = \eta_{01}/n$. The identities for η_{21} and η_{22} follow from the symmetry of the error distribution. There is a small error in Carroll and Cline's Theorem 4(a): In the expression for $\mathbf{S}_{EL}^{-1}$, the term $4\eta_{22}\mathbf{S}_L^{-1}$ should be divided by m.

Let $\boldsymbol{\Sigma}_{EB}$ denote the asymptotic covariance matrix associated with the "empirical Bayes" weight function (3.2); that is, (2.6) with

$$\mathbf{A} - \mathbf{C} = E\left\{\frac{(n-2)v + \gamma\tau}{nv + \gamma\tau}\, w_0\mathbf{X}^T\mathbf{X}\right\} \quad \text{and}$$

$$\mathbf{B} = E\{vw_0^2\mathbf{X}^T\mathbf{X}\}. \tag{4.1}$$

If model (3.1) holds and n is constant, then the covariance matrices can be evaluated as follows:

$$\boldsymbol{\Sigma}_{CV} = [E\{\tau^{-1}\mathbf{X}^T\mathbf{X}\}]^{-1},$$

$$\boldsymbol{\Sigma}_{EB} = \{1 + 2/(n+\gamma)\}\boldsymbol{\Sigma}_{CV},$$

$$\boldsymbol{\Sigma}_{ML} = \{1 + 2/(n-2)\}\boldsymbol{\Sigma}_{CV} \quad \text{for} \quad n \ge 3, \tag{4.2}$$

$$\boldsymbol{\Sigma}_{\text{OLS}} = \{1 + 2/(\gamma-2)\}[E\{\mathbf{X}^T\mathbf{X}\}]^{-1} \times E\{\tau\mathbf{X}^T\mathbf{X}\}[E\{\mathbf{X}^T\mathbf{X}\}]^{-1} \quad \text{for} \quad \gamma > 2,$$

and

$$\boldsymbol{\Sigma}_{FR} = (1 + 2/n - 8/n^2)\boldsymbol{\Sigma}_{CV} + (4/n^2)\boldsymbol{\Sigma}_{\text{OLS}} \quad \text{for} \quad n \ge 3 \quad \text{and} \quad \gamma > 2.$$

If in addition τ is constant, then

$$\boldsymbol{\Sigma}_{\text{OLS}} = \{1 + 2/(\gamma-2)\}\boldsymbol{\Sigma}_{CV} \quad \text{for} \quad \gamma > 2,$$

and

$$\Sigma_{FR} = [1 + 2(n-2)/n^2 + 8/\{(\gamma-2)n^2\}]\Sigma_{CV} \quad \text{for} \quad n \geq 3 \quad \text{and} \quad \gamma > 2. \quad (4.3)$$

5. COMBINING INDEPENDENT ESTIMATES

The simplest application of WLS is the problem of estimating a common mean given samples from several populations with unknown variances; that is, $p = 1$ and $\mathbf{X} = \mathbf{1}_n$. Let $\hat{\beta}_{EB}$ be the empirical Bayes weighted estimator defined by Algorithm 1 with $c_\beta = \infty$, $c_\theta = 3$, and weight function (3.2). The parameters γ and τ are estimated using (3.3) with τ constant and with a priori bounds $(\gamma_{lb}, \gamma_{ub}) = (1, 10)$. Table 1 presents simulation results comparing the variances of $\hat{\beta}_{EB}$, $\hat{\beta}_{FR}$, and $\hat{\beta}_{ML}$ under model (3.1) when τ is constant; the total number of observations is 36; $(n, k) = (1, 36)$, (2, 18), (3, 12), (4, 9), (6, 6), and (9, 4); and γ varies from 100 to 1. In each case the variance is estimated from 3,000 replicates of $\hat{\beta}$ and is standardized by dividing by the variance of $\hat{\beta}_{CV}$.

The simulated variance of the Fuller–Rao estimator increases as γ decreases and fluctuates wildly when $\gamma = 1$, suggesting that the variance becomes infinite at this point. When $\gamma = 1$, the errors are Cauchy and the expected value of the OLS estimator fails to exist. For $n \geq 2$ the simulated variance of the conditional normal-theory MLE remains fairly constant as γ varies. The different behavior for $n = 1$ can be explained as follows. Here the solution to the iterative algorithm seems to depend strongly on the starting point; if $\hat{\beta}$ falls too close to an observation y_i, then in subsequent cycles $\hat{\beta}$ converges towards y_i. When γ is large, the OLS estimate tends to fall near the median observation and subsequent estimates tend to remain in this neighborhood. When γ is small, the estimate can be captured by an observation far from the median.

The simulated variances in Table 1 may be compared with the asymptotic variances (4.2) and (4.3) by plotting the ratio of standardized variances (simulated/asymptotic) against n and γ. For example, the ratio for $\hat{\beta}_{EB}$ with $n = 1$ and $\gamma = 100$ is $1.11/\{1 + 2/(1 + 100)\} = 1.09$. The plots, not shown here, reveal the following. The ratio for $\hat{\beta}_{EB}$ is typically greater than 1 and is close to 1 except when n and γ are both small. The ratio for $\hat{\beta}_{FR}$, with $n \geq 3$ and $\gamma > 2$, is always less than 1 and is smallest when n and γ are both small. The ratio for $\hat{\beta}_{ML}$ does not depend on γ and is approximately .75 for $n = 3$ and .95 for $n = 4$, 6, and 9. In summary, it appears that standard errors based on asymptotic theory tend to be slightly liberal for the empirical Bayes weighted estimator and conservative for the other two estimators.

The iterative algorithm for $\hat{\beta}_{EB}$ typically converges within five iterations and nearly always within fifteen. As suggested by (4.1) and the argument following Algorithm 1, convergence is slower when n and γ are small. The rate of convergence also depends on c_θ. In simulations with $c_\theta = \infty$, the algorithm often failed to converge for $n = 1$. Increasing c_θ seems to slightly improve the efficiency of $\hat{\beta}_{EB}$ for small γ but may also marginally reduce efficiency for large γ.

It is not surprising that comparisons in Table 1 mostly favor $\hat{\beta}_{EB}$, because the estimates were generated under a model where $\hat{\beta}_{EB}$ is approximately optimal. Under some models $\hat{\beta}_{EB}$ gives poor results; for example, when all but a few variances are large. Under various other models examined in simulation studies, however, $\hat{\beta}_{EB}$ usually performed about as well or better than both $\hat{\beta}_{FR}$ and $\hat{\beta}_{ML}$.

6. ASYMMETRIC ERROR DISTRIBUTIONS

The results in this article are based on assumption (2.2), which typically fails if the errors are not symmetrically distributed about 0. More research is needed on the effects of departures from this assumption, but a few comments can be made here. Carroll and Cline (1988, th. 3) showed that the conditional normal-theory MLE and the Fuller–Rao estimator usually are not consistent under asymmetric error distributions. The following discussion shows the lack of consistency of the empirical Bayes weighted estimator as well, although the effect of asymmetry in moderate-sized samples remains to be evaluated.

Table 1. Estimates of $\sigma^2_{EB}/\sigma^2_{CV}$, $\sigma^2_{FR}/\sigma^2_{CV}$, and $\sigma^2_{ML}/\sigma^2_{CV}$ Under Model (3.1)

		γ							
(n, k)		100	20	10	5	3	2	1.5	1
(1, 36)	EB	1.11	1.17	1.23	1.37	1.61	1.90	2.31	3.32
	FR	1.05	1.12	1.24	1.64	2.74	7.65	23.60	305.99
	ML	1.15	1.20	1.32	1.70	2.76	6.93	17.53	138.49
(2, 18)	EB	1.07	1.14	1.18	1.33	1.47	1.59	1.74	1.99
	FR	1.18	1.25	1.30	1.52	2.03	3.51	6.89	32.06
	ML	2.86	2.88	2.96	2.89	2.98	3.28	3.74	4.60
(3, 12)	EB	1.08	1.11	1.18	1.24	1.40	1.42	1.53	1.64
	FR	1.22	1.25	1.30	1.38	1.68	2.27	3.52	24.72
	ML	2.38	2.34	2.24	2.29	2.28	2.15	2.36	2.23
(4, 9)	EB	1.05	1.10	1.15	1.24	1.32	1.37	1.42	1.54
	FR	1.20	1.24	1.24	1.34	1.50	1.80	2.41	4.70
	ML	1.82	1.83	1.78	1.78	1.90	1.84	1.80	1.86
(6, 6)	EB	1.06	1.09	1.14	1.18	1.22	1.22	1.25	1.34
	FR	1.20	1.16	1.20	1.20	1.28	1.37	1.57	3.50
	ML	1.48	1.41	1.42	1.35	1.36	1.34	1.35	1.42
(9, 4)	EB	1.05	1.06	1.10	1.15	1.15	1.17	1.23	1.17
	FR	1.12	1.11	1.13	1.16	1.18	1.23	1.76	9,678.33
	ML	1.22	1.19	1.20	1.22	1.21	1.18	1.22	1.15

Let $\hat{\beta}_{EB}$ denote the iterative WLS estimator using the weight function (3.2). For simplicity we assume here that γ and the τ_i are known. Applying the comment following Theorem 2 shows that the estimate $\hat{\beta}_{EB}$ minimizes $(1/k) \times \sum (n_i + \gamma)\log(\|\mathbf{y}_i - \mathbf{X}_i\hat{\beta}_{EB}\|^2 + \gamma\tau_i)$. Under regularity conditions $\hat{\beta}_{EB}$ converges asymptotically to β^*, where β^* minimizes $E\{(n + \gamma)\log(\|\mathbf{y} - \mathbf{X}\beta^*\|^2 + \gamma\tau)\} = E\{(n + \gamma)\log(\|\tilde{\mathbf{e}} - \tilde{\mathbf{X}}(\beta^* - \beta)\|^2 + \gamma)\} + E\{(n + \gamma)\log(\tau)\}$ for $(\tilde{\mathbf{y}}, \tilde{\mathbf{X}}, \tilde{\mathbf{e}}) = \tau^{-1/2}(\mathbf{y}, \mathbf{X}, \mathbf{e})$. Suppose that $\tilde{\mathbf{e}}$ and $\tilde{\mathbf{X}}$ are conditionally independent given n, that the entries in $\tilde{\mathbf{e}}$ are independent and identically distributed, and that the first column of $\mathbf{X}$ is $\mathbf{1}_n$ (i.e., the model for the mean contains a constant term). If n and τ are constant and if $\tilde{\mathbf{X}} = \mathbf{1}_n\tilde{\mathbf{x}}^T$ (i.e., groups are determined by replicates), then $\beta_j^* = \beta_j$ for $j = 2, \ldots, p$; that is, slope parameters are estimated consistently but the intercept parameter usually is not. This conclusion is obtained by minimizing the conditional expectation given $\tilde{\mathbf{X}}$ and observing that the optimum value of $\tilde{\mathbf{x}}^T(\beta^* - \beta)$ can be taken as $\tau^{-1/2}(\beta_1^* - \beta_1)$, free of $\tilde{\mathbf{x}}$. Carroll and Welsh (1988) described related results for M estimators.

This conclusion can fail if either n or τ is allowed to vary. For example, suppose that τ is constant but $\mathbf{X}$ can take on the values [1:0] or $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. We can again choose $\beta^* - \beta$ to simultaneously minimize the conditional expectation for both values of $\mathbf{X}$, but typically neither component of the solution will equal 0. If τ varies with $\mathbf{X}$, then the conditional minimization argument fails, because $\tau^{-1/2}(\beta_1^* - \beta_1)$ is no longer free of $\tilde{\mathbf{x}}$. It appears that some linear functions of β may be estimated consistently but the coefficients may depend on τ. For example, if $p = 2$, $\mathbf{X}$ can take on the values [1:0] or [0:1], and τ is a function of the mean, $\tau(\beta_1) \neq \tau(\beta_2)$, then $\tau(\beta_1)^{-1/2}\beta_1 - \tau(\beta_2)^{-1/2}\beta_2$ is estimated consistently but $\beta_1 - \beta_2$ usually is not.

On a more positive note, the bias caused by asymmetry decreases as γ increases, because $\hat{\beta}_{EB}$ behaves more like OLS applied to the transformed data $(\tilde{\mathbf{y}}_i, \tilde{\mathbf{X}}_i)$. The following heuristic argument suggests that the bias also decreases as the group sizes n_i increase. Changing notation, write $\tilde{\mathbf{e}} = (\tilde{e}_1, \ldots, \tilde{e}_n)^T$ and $\tilde{v}_n = \|\tilde{\mathbf{e}}\|^2/n$, and suppose that $E\{\tilde{\mathbf{e}} \mid \tilde{v}_n, \tilde{\mathbf{X}}\} = E\{\tilde{e}_1 \mid \tilde{v}_n\}\mathbf{1}_n$ and that the $\tilde{e}_j$ are independent and identically distributed with mean 0. By Jensen's inequality we have, for each $m \geq 1$,

$$\begin{aligned} \operatorname{var} E\{\tilde{e}_1 \mid \tilde{v}_m\} &= E\{E[(E\{\tilde{e}_1 \mid \tilde{v}_m, \tilde{e}_{m+1}^2\})^2 \mid \tilde{v}_{m+1}]\} \\ &\geq E\{(E\{\tilde{e}_1 \mid \tilde{v}_{m+1}\})^2\} \\ &= \operatorname{var} E\{\tilde{e}_1 \mid \tilde{v}_{m+1}\}. \end{aligned}$$

Under symmetric error distributions, $E\{\tilde{e}_1 \mid \tilde{v}_n\}$ is identically 0. We conjecture that the bias will tend to increase as $\operatorname{var} E\{\tilde{e}_1 \mid \tilde{v}_n\}$ increases.

APPENDIX: PROOFS

Proof of Theorem 1. By Taylor's theorem there exist points $(\beta_{ik}^*, \theta_{ik}^*)$ on the line segment connecting $(\tilde{\beta}_k, \tilde{\theta}_k)$ and (β, θ) such that

$$\begin{aligned} \hat{w}_i(\tilde{\beta}_k, \tilde{\theta}_k) = \hat{w}_i(\beta, \theta) &+ \hat{\mathbf{w}}_{i\beta}(\beta_{ik}^*, \theta_{ik}^*)^T(\tilde{\beta}_k - \beta) \\ &+ \hat{\mathbf{w}}_{i\theta}(\beta_{ik}^*, \theta_{ik}^*)^T(\tilde{\theta}_k - \theta). \end{aligned}$$

Straightforward algebra yields (2.5) with

$$\mathbf{A}_k = k^{-1} \sum \hat{w}_i(\tilde{\beta}_k, \tilde{\theta}_k)\mathbf{X}_i^T\mathbf{X}_i,$$

$$\mathbf{C}_k = k^{-1} \sum \mathbf{X}_i^T\mathbf{e}_i\hat{\mathbf{w}}_{i\beta}(\beta_{ik}^*, \theta_{ik}^*)^T,$$

and

$$\mathbf{D}_k = k^{-1} \sum \mathbf{X}_i^T\mathbf{e}_i\hat{\mathbf{w}}_{i\theta}(\beta_{ik}^*, \theta_{ik}^*)^T.$$

The limits are obtained by using the smoothness conditions (2.3) to replace $(\tilde{\beta}_k, \tilde{\theta}_k)$ and $(\beta_{ik}^*, \theta_{ik}^*)$ by (β, θ), calculating $\hat{\mathbf{w}}_{i\beta}(\beta, \theta) = (-2/n_i)w'(v_i, \mathbf{X}_i, \theta)\mathbf{X}_i^T\mathbf{e}_i$ and applying (2.2). We obtain (2.6) from (2.5), Slutsky's theorem, and the fact that $\mathbf{b}_k$ is asymptotically $N_p(0, \mathbf{B})$.

Proof of Theorem 2. Integration by parts yields

$$\int_0^\infty w'(v, \mathbf{X})vp(v \mid \mathbf{X})\, dv = -\int_0^\infty w(v, \mathbf{X})\frac{\partial}{\partial v}\{vp(v \mid \mathbf{X})\}\, dv,$$

and it is then straightforward to show that $\mathbf{A} - \mathbf{C} = E[\{w + (2/n)vw'\}\mathbf{X}^T\mathbf{X}] = E\{vww_0\mathbf{X}^T\mathbf{X}\}$. Put $\mathbf{B}_0 = E\{vw_0^2\mathbf{X}^T\mathbf{X}\}$. We must prove that $(\mathbf{A} - \mathbf{C})^{-1}\mathbf{B}(\mathbf{A} - \mathbf{C})^{-1} \geq \mathbf{B}_0^{-1}$ or, equivalently, that $\mathbf{B}_0 - (\mathbf{A} - \mathbf{C})\mathbf{B}^{-1}(\mathbf{A} - \mathbf{C}) \geq 0$. This inequality follows from the fact that

$$\begin{bmatrix} \mathbf{B}_0 & \mathbf{A} - \mathbf{C} \\ \mathbf{A} - \mathbf{C} & \mathbf{B} \end{bmatrix} = E\left\{v\begin{bmatrix} w_0\mathbf{X}^T \\ w\mathbf{X}^T \end{bmatrix}[w_0\mathbf{X} : w\mathbf{X}]\right\}$$

is nonnegative definite symmetric.

Proof of Theorem 3. Put $t = 1/\sigma^2$ and let $Q(dt \mid \mathbf{X})$ denote the conditional distribution of t given $\mathbf{X}$. The conditional density of v given $(\mathbf{X}, t)$ is $p(v \mid \mathbf{X}, t) = \{\Gamma(n/2)(2/nt)^{n/2}\}^{-1}v^{(n-2)/2}\exp\{-ntv/2\}$. We have

$$\begin{aligned} \frac{\partial}{\partial v}p(v \mid \mathbf{X}) &= \frac{\partial}{\partial v}\int_0^\infty p(v \mid \mathbf{X}, t)Q(dt \mid \mathbf{X}) \\ &= \frac{n-2}{2v}p(v \mid \mathbf{X}) - \frac{n}{2}\int_0^\infty tp(v \mid \mathbf{X}, t)Q(dt \mid \mathbf{X}), \end{aligned}$$

and

$$\frac{\partial}{\partial v}\log p(v \mid \mathbf{X}) = \frac{n-2}{2v} - \frac{n}{2}E\{t \mid v, \mathbf{X}\},$$

and the result follows from Theorem 2.

[*Received January 1991. Revised April 1992.*]

REFERENCES

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., and Ruppert, D. (1982), "Robust Estimation in Heteroscedastic Linear Models," *The Annals of Statistics,* 10, 429–441.

Carroll, R. J., and Ruppert, D. (1988), *Transformation and Weighting in Regression,* New York: Chapman and Hall.

Carroll, R. J., and Welsh, A. H. (1988), "A Note on Asymmetry and Robustness in Linear Regression," *The American Statistician,* 42, 285–287.

Carroll, R. J., Wu, C. F. J., and Ruppert, D. (1988), "The Effect of Estimating Weights in Weighted Least Squares," *Journal of the American Statistical Association,* 83, 1045–1054.

Cohen, A., and Sackrowitz, H. B. (1989), "Exact Tests That Recover Interblock Information in Balanced Incomplete Blocks Designs," *Journal of the American Statistical Association,* 84, 556–559.

Fuller, W. A., and Rao, J. N. K. (1978), "Estimation for a Linear Regression Model With Unknown Diagonal Covariance Matrix," *The Annals of Statistics,* 6, 1149–1158.

Hui, S. L., and Berger, J. O. (1983), "Empirical Bayes Estimation of Rates in Longitudinal Studies," *Journal of the American Statistical Association,* 78, 753–760.

Shao, J. (1992), "Empirical Bayes Estimation of Heteroscedastic Variances," *Statistica Sinica,* 2, 495–518.

Street, J. O., Carroll, R. J., and Ruppert, D. (1988), "A Note on Computing Robust Regression Estimates Via Iteratively Reweighted Least Squares," *The American Statistician,* 42, 152–154.