An Efficient Algorithm for REML in Heteroscedastic Regression[*]

Gordon K. Smyth
Walter and Eliza Hall Institute of Medical Research
Australia 3050

Abstract

This paper considers REML (residual or restricted maximum likelihood) estimation for heteroscedastic linear models. An explicit algorithm is given for REML-scoring which yields the REML estimates together with their standard errors and likelihood values. The algorithm includes a Levenberg-Marquardt restricted step modification which ensures that the REML-likelihood increases at each iteration. This paper shows how the complete computation, including the REML information matrix, may be carried out in $O(n)$ operations.

Keywords: Residual maximum likelihood; Restricted maximum likelihood; Method of scoring; Levenberg-Marquardt restricted step method; Information matrix.

1 Introduction

This paper considers REML (residual or restricted maximum likelihood) estimation for heteroscedastic linear models. We suppose that the responses $y_1, \ldots, y_n$ are independent and that $y_i \sim N(\mu_i, \sigma_i^2/w_i)$ with

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

and

$$g(\sigma_i^2) = \mathbf{z}_i^T \boldsymbol{\gamma}$$

where the w_i are prior weights and $g()$ is a known monotonic differentiable function. Here $\mathbf{x}_i$ is a vector of covariates relevant for predicting the mean, $\mathbf{z}_i$ is a vector of covariates relevant for predicting the variance and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are vectors of regression coefficients. The model assumes that the variance of each y_i is not, given the covariates,

[*] *Journal of Graphical and Computational Statistics* **11**, 836–847 (December 2002).

functionally dependent on the mean μ_i. Responses which do show a marked mean-variance relationship can often be transformed to a suitable scale using a Box-Cox transformation (Box & Cox, 1964) after which analysis can proceed as in this paper, or else may be treated by more general variance modeling methods such as those of Davidian & Carroll (1987) or Smyth (1989).

Heteroscedastic regression models have an extensive literature going back to Park (1966), Rutemiller & Bowers (1968) and Harvey (1976). Aitkin (1987) considered a log-linear model for the variances and developed GLIM code for maximum likelihood estimation. Heteroscedastic regression has recently gained popularity in industrial statistics for analyzing unreplicated experiments, experiments for robust design and the analysis of process data. See for example Box & Meyer (1986a), Box & Meyer (1986b), Carroll & Ruppert (1988), Nair & Pregibon (1988), Nelder & Lee (1991), Chapter 10 of Myers & Montgomery (1995), Engel & Huele (1996), Bergman & Hynén (1997), Lee & Nelder (1998), Nelder & Lee (1998), Huele (1998) and Huele & Engel (1998).

REML is a method of estimating the variance parameters, in this case $\boldsymbol{\gamma}$, using a marginal likelihood function in which the mean parameters do not appear. This is achieved by considering the likelihood not of the y_i but of the set of all zero-mean contrasts of the y_i. REML estimation was introduced by Patterson & Thompson (1971) for normal random effects models. An extensive discussion was given by Harville (1977). There are various reasons for preferring REML over maximum likelihood for estimation of the variances. The most frequently quoted reasons are that the estimators are less biased and that an appropriate degree of freedom correction is produced in standard cases (Tunnicliffe Wilson, 1989). Other reasons are that REML is related to Bayesian marginal inference (Harville, 1974) and that REML is less sensitive to influential observations with high leverage in the mean model (Verbyla, 1993). Perhaps the strongest reason is that the REML score vector for the variance coefficients is unbiased, providing consistent estimators in situations where maximum likelihood estimators are inconsistent. A systematic study of REML for the log-linear variance model was undertaken by Verbyla (1993). Lee & Nelder (1998), Huele & Engel (1998), Smyth & Verbyla (1999), Huele et al. (2000) and Smyth et al. (2001) discuss how the REML estimator from Verbyla (1993) can be obtained by repeated fitting of generalized linear models, although this approach cannot be used to obtain standard errors.

The model considered in this paper is a slight generalization of that considered in Verbyla (1993), where g was assumed to be the exponential function. Verbyla (1993) did not give an algorithm for computing the estimates. Other authors, such as Carroll & Ruppert (1988), Nelder & Lee (1998), Huele (1998) and Huele & Engel (1998), have reported convergence problems trying to compute ML or REML estimators for heteroscedastic regression. This paper gives an explicit algorithm for REML-scoring which yields the REML estimates together with their standard errors and likelihood values. The algorithm includes a Levenberg-Marquardt restricted step modification (Fletcher, 1987, Section 5.2; Thisted, 1988, Section 4.5.3.3) in order to ensure that the REML-likelihood increases at each iteration. The algorithm is based on prototype Levenberg-Marquardt algorithms which are established in the numerical literature but which are not well known in the statistics literature. In particular the use of restricted step modifications for Fisher scoring has been very rare, one example being Osborne (1987). The implementation here includes new suggestions for the stopping criterion and the initialization of the damping parameter which utilize the Fisher scoring context.

Several authors, such as Cook & Weisberg (1983), Lee & Nelder (1998), Huele & Engel (1998) and Smyth & Verbyla (1999), have avoided the exact REML likelihood calculations because of the computational burden involved in computing the REML information matrix for large data sets. In each case they used a diagonal matrix to approximate a dense full-rank $n \times n$ matrix which appears in the REML information matrix. The effect of using such approximations has been examined in detail by Smyth et al. (2001). This paper gives a new decomposition for the correct REML information matrix which enables it to be computed in $O(n)$ operations thereby making an approximation unnecessary for most problems. This result allows the complete modified REML-scoring computation to be carried out in $O(n)$ operations using $O(n)$ storage.

Section 2 reviews maximum likelihood estimation for heteroscedastic regression and Section 3 sets out the basic REML calculations. Section 4 derives a decomposition for the REML information which allows $O(n)$ computation. An iterative algorithm for REML is outlined in Section 5. Section 6 considers a data example from industrial statistics.

2 Maximum Likelihood Estimation

Before considering REML estimation it is useful to set out the maximum likelihood calculations for comparison. This section summarizes maximum likelihood estimation for parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The results are a special case of results in Nelder & Pregibon (1987) and Smyth (1989).

The score vector for $\boldsymbol{\beta}$ is

$$U_\beta = X^T \Sigma_m^{-1} (\mathbf{y} - X\boldsymbol{\beta})$$

and the information matrix is

$$\mathcal{I}_\beta = X^T \Sigma_m^{-1} X$$

where $\Sigma_m = \operatorname{var} \mathbf{y} = \operatorname{diag}(\sigma_i^2 / w_i)$ and X is the $n \times p$ design matrix with ith row $\mathbf{x}_i^T$. The score vector and information matrix for $\boldsymbol{\gamma}$ are

$$U_\gamma = Z^T G \Sigma_d^{-1} (\mathbf{d} - \boldsymbol{\sigma}^2)$$

and

$$\mathcal{I}_\gamma = Z^T W_d^{-1} Z$$

where $\mathbf{d}$ is the n-vector of $d_i = w_i(y_i - \mu_i)^2$, $\boldsymbol{\sigma}^2 = E(\mathbf{d})$ is the n-vector of σ_i^2, $\Sigma_d = \operatorname{var} \mathbf{d} = \operatorname{diag}(2\sigma_i^4)$, $G = \operatorname{diag}\{1/\dot{g}(\sigma_i^2)\}$, $\dot{g}$ is the derivative of g, $W_d = G^2 \Sigma_d^{-1}$ and Z

is the $n \times q$ design matrix with ith row $\mathbf{z}_i^T$. The two parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal. The method of scoring for computing the maximum likelihood estimates yields

$$\boldsymbol{\beta}_{k+1} = \left(X^T \Sigma_m^{-1} X\right)^{-1} X^T \Sigma_m^{-1} \mathbf{z}_m$$

with $\mathbf{z}_m = \mathbf{y} - \boldsymbol{\mu} + X\boldsymbol{\beta}$ and

$$\boldsymbol{\gamma}_{k+1} = \left(Z^T W_d Z\right)^{-1} Z^T W_d^{-1} \mathbf{z}_d$$

with $\mathbf{z}_d = G^{-1}(\mathbf{d} - \boldsymbol{\sigma}^2) + Z\boldsymbol{\gamma}$. Here k indicates the kth iterate and the right-hand sides are evaluated at $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$. The scoring iteration for $\boldsymbol{\beta}$ has the form of weighted linear regression while the scoring iteration of $\boldsymbol{\gamma}$ is that for a gamma generalized linear model with responses d_i, link $g()$ and dispersion equal to 2.

3 Residual Maximum Likelihood

We now consider REML estimation for the heteroscedastic model. This section generalizes the results of Verbyla (1993) to an arbitrary link function for the variance. The REML estimator of $\boldsymbol{\gamma}$ is obtained by maximizing the marginal log-likelihood

$$\begin{aligned} \ell_R(\mathbf{y}; \boldsymbol{\gamma}) &= \ell(\mathbf{y}; \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma}), \boldsymbol{\gamma}) - \frac{1}{2} \log |X^T \Sigma_m^{-1} X| \\ &= -\frac{1}{2} \left(\log |\Sigma_m| + \mathbf{y}^T P \mathbf{y} + \log |X^T \Sigma_m^{-1} X| \right) \end{aligned}$$

where ℓ is the ordinary log-likelihood, $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$ is the conditional maximum likelihood estimator for $\boldsymbol{\beta}$ for given fixed $\boldsymbol{\gamma}$ and

$$P = \Sigma_m^{-1} - \Sigma_m^{-1} X (X^T \Sigma_m^{-1} X)^{-1} X^T \Sigma_m^{-1}.$$

The REML score vector for $\boldsymbol{\gamma}$ is

$$U_R = Z^T G \Sigma_d^{-1} (\mathbf{d} - \boldsymbol{\sigma}^{2*})$$

where $\boldsymbol{\sigma}^{2*}$ is the n-vector of $(1 - h_{ii})\sigma_i^2$ and the h_{ii} are the diagonal elements of

$$H = \Sigma_m^{-1/2} X (X^T \Sigma_m^{-1} X)^{-1} X^T \Sigma_m^{-1/2},$$

the hat matrix in the weighted regression for $\boldsymbol{\beta}$. The information matrix is

$$\mathcal{I}_R = Z^{*T} V Z^*$$

where $Z^* = \Sigma_d^{-1/2} G Z$ and V is the $n \times n$ matrix with diagonal elements $(1 - h_{ii})^2$ and off-diagonal elements h_{ij}^2, the h_{ij} being the elements of H. Here V is the covariance matrix of the squared residuals, $\Sigma_m^{-1}\mathbf{d}$, where $\mathbf{d}$ is evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$. The REML scoring iteration for $\boldsymbol{\gamma}$ is

$$\boldsymbol{\gamma}_{k+1} = \boldsymbol{\gamma}_k + \mathcal{I}_R^{-1} U_R \tag{1}$$

with $\mathcal{I}_R$ and U_R computed at $\boldsymbol{\gamma}_k$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$. This differs from the ordinary scoring iteration for $\boldsymbol{\gamma}$ in that $\boldsymbol{\sigma}^{2*}$ replaces $\boldsymbol{\sigma}^2$ in the score vector and the matrix V is inserted in the information matrix. The dense matrix V means that the REML scoring iteration cannot be written as an iteratively-reweighted least squares calculation and is therefore best treated as a general nonlinear iteration.

4 Computation of the Information Matrix

The REML information matrix involves the matrix V, which is a dense $n \times n$ matrix and is generally of full rank. Forming the matrix V explicitly therefore produces an $O(n^2)$ computation which is likely to be prohibitive for medium to large size data sets. Indeed several authors have avoided computing V by approximating it with a diagonal matrix (Cook & Weisberg, 1983; Huele & Engel, 1998; Smyth & Verbyla, 1999).

Although the off-diagonal elements of V are often of smaller order than the diagonal elements, there are $O(n^2)$ such elements so the effect on the information matrix of ignoring them is not generally negligible. Suppose that $h_{ij} = O(n^{-1})$ and that $n^{-1}\mathcal{I}_R$ has a positive definite limit, both of which are true under standard regularity conditions. Write $V_2 = \operatorname{diag} V$. Even though the off-diagonal elements of V are $O(n^{-2})$, $n^{-1}\mathcal{I}_R - n^{-1}Z^{*T}V_2Z^*$ does not converge to zero. Ignoring the off-diagonal elements of V produces relative errors in the elements of the $\mathcal{I}_R$ which do not tend to zero as n increases. Such errors will impact on the computed standard errors for $\hat{\boldsymbol{\gamma}}$ and will slow down convergence of the scoring algorithm for computing $\hat{\boldsymbol{\gamma}}$.

We show here that the correct REML information matrix can be computed in $O(n)$ operations for p and q constant, making diagonal approximation to V unnecessary for most problems. This is achieved by representing V as a diagonal matrix plus a matrix of low rank. The matrix V can be written

$$V = \operatorname{diag}(1 - 2h_{ii}) + H^2$$

where H^2 represents the matrix with elements h_{ij}^2. Recall that H is the "hat matrix", the matrix for the orthogonal projection operator onto the column space of $\Sigma_m^{-1/2}X$. Therefore H has p eigenvalues 1, where p is the rank of X, and $n - p$ eigenvalues 0. The nonzero eigenvalues of H correspond to eigenvectors which span the column space of $\Sigma_m^{-1/2}X$. We show now that H^2 has rank at most $p(p+1)/2$. Let $\mathbf{q}_1, \ldots, \mathbf{q}_p$ be an orthonormal basis for the range space of H. Any basis is sufficient, for example we can let the $\mathbf{q}_j$ be the columns of the Q-matrix from the QR-decomposition of $\Sigma_m^{-1/2}X$. Then the $\mathbf{q}_j$ are eigenvectors for H with eigenvalues equal to 1, so we can write

$$H = \sum_{a=1}^{p} \mathbf{q}_a \mathbf{q}_a^T$$

i.e.,

$$h_{ij} = \sum_{a=1}^{p} q_{a,i} q_{a,j}$$

where $q_{a,i}$, $i = 1, \ldots, n$, are the elements of $\mathbf{q}_a$. Squaring this expression gives

$$h_{ij}^2 = \sum_{a=1}^{p} q_{a,i}^2 q_{a,j}^2 + 2 \sum_{1 \le a < b \le p} q_{a,i} q_{a,j} q_{b,i} q_{b,j}$$

or in matrix terms

$$H^2 = \sum_{j=1}^{p} \mathbf{s}_j \mathbf{s}_j^T + 2 \sum_{k=1}^{p(p-1)/2} \mathbf{t}_k \mathbf{t}_k^T$$

where each $\mathbf{s}_j$ is a vector of the squares, $s_{j,i} = q_{j,i}^2$, and each $\mathbf{t}_k$ is a vector of products, $t_{k,i} = q_{a,i} q_{b,i}$ for $a < b$. This represents H^2 as the sum of $p + p(p-1)/2 = p(p+1)/2$ rank-one matrices. It follows that H^2 is of rank at most $p(p+1)/2$, less if $n < p(p+1)/2$ or if the $\mathbf{s}_j$ and $\mathbf{t}_k$ are not all linearly independent. The $\mathbf{t}_k$ can be computed for example from the pseudo code

```
$k = 0$
for $i = 1$ to $p - 1$ {
for $j = i + 1$ to $p$ {
    $k = k + 1$
    $\mathbf{t}_k = \mathbf{q}_i . \mathbf{q}_j$
}}
```

where $\mathbf{q}_i . \mathbf{q}_j$ represents component-wise vector multiplication.

The low rank representation for H^2 allows the information matrix $\mathcal{I}_R$ to be computed efficiently. Let S be the $n \times p$ matrix with columns $\mathbf{s}_j$ and let T be the $n \times p(p-1)/2$ matrix with columns $\mathbf{t}_k$. Let B be the $p(p+1)/2 \times q$ matrix $(S,\ 2T)^T Z^*$. Forming B requires $nqp(p+1)/2$ multiplications. Then

$$\mathcal{I}_R = Z^{*T} V Z^* = Z^{*T} \mathrm{diag}(1 - 2h_{ii}) Z^* + B^T B \tag{2}$$

The first term on the right-hand side requires $O(nq^2)$ flops while the second requires $O(pq^2)$. The extra computation involved in computing the correct REML information matrix rather than using a diagonal approximation for V is dominated by the formation of B.

Standard results for matrix multiplications (Golub & Van Loan, 1996) show that $\hat{\boldsymbol{\beta}}(\gamma)$ can be computed for fixed $\boldsymbol{\gamma}$ in $O(np^2)$ flops. Similarly each update (1) for $\boldsymbol{\gamma}$ would require $O(nq^2)$ flops given $\hat{\boldsymbol{\beta}}(\gamma)$ if V was approximated by a diagonal matrix. We can conclude that approximating V by a diagonal matrix produces an approximate REML scoring iteration requiring $O(np^2) + O(nq^2)$ flops. REML scoring using the correct matrix V and implemented as in (2) requires $O(np^2q) + O(nq^2)$ flops. The important result is that the computational burden, both in terms of flops and in terms of storage requirements, grows only linearly with the size of the data set for fixed p and q. This allows the REML information matrix to be computed for large data sets. In the author's implementation using the S-Plus language on a PC running Windows 98, evaluation of $\mathcal{I}_R$ involving explicit formation of the matrix V is limited by storage requirements to data sets with n no more than about 500. By comparison, evaluation of $\mathcal{I}_R$ using (2) causes no problems with $n = 100,000$ and p and q both small and requires only a few seconds of computer time.

5 Implementation of REML Scoring

Huele et al. (2000) have pointed out that the REML score equations $U_R = 0$ can be solved by repeatedly fitting a gamma generalized linear model with responses $d_i/(1 - h_{ii})$, link $g()$ and prior weights $1 - h_{ii}$, where the d_i and the h_{ii} are updated at each iteration from the updated $\boldsymbol{\gamma}$. Lee & Nelder (1998) describe a similar strategy, although without specifying the prior weights $1 - h_{ii}$. Alternatively one could use responses $d_i + h_{ii}\phi_i$ and prior weights unity, with again d_{ii} and h_{ii} updated at each iteration (Huele & Engel, 1998; Huele et al., 2000). Both of these strategies have the REML estimates for $\boldsymbol{\gamma}$ as a stationary value. On the other hand, neither generalized linear model gives correct standard errors for $\boldsymbol{\gamma}$ (the unweighted model being worse that the weighted in this respect), although these may be computed from $\mathcal{I}_R$ at convergence of the iteration.

An even more serious problem is that these iterative strategies often experience difficulties with convergence. One problem is that Lee & Nelder (1998) and Huele & Engel (1998) appear to iterate the gamma generalized linear model to convergence before updating $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$, so that $\boldsymbol{\gamma}$ is updated several times before the d_i and h_{ii} are recomputed. This is incorrect in that the d_i and h_{ii} are functions of $\boldsymbol{\gamma}$ in the REML score vector and these quantities must be updated with $\boldsymbol{\gamma}$ in order to solve the correct estimating equation. Iterating the generalized linear model to convergence also introduces an unnecessary inner iteration which can itself produce convergence problems. An appropriate implementation is to perform a single iteration of the gamma generalized linear model between updates of $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$, d_i and h_{ii}. The iteration is then a pseudo-Newton iteration for maximizing the REML likelihood.

REML scoring is a nonlinear iteration and cannot be guaranteed to converge without some sort of step restriction to ensure an increase in the likelihood at each iteration. Empirical experience with REML suggests that REML scoring often converges more slowly than the maximum likelihood algorithms discussed in Section 2 or Smyth (1989). The remainder of this section describes REML scoring with a simple Levenberg-Marquardt restricted step strategy. Let A be an approximation to the REML information matrix $\mathcal{I}_R$. The algorithm is based on the principle that $\lambda > 0$ can always be chosen sufficiently large so that

$$\boldsymbol{\gamma}_{k+1} = \boldsymbol{\gamma}_k + (A + \lambda I)^{-1} U_R \tag{3}$$

is an ascent step for the REML likelihood, where I is the $q \times q$ identity matrix. The parameter λ introduces Levenberg-Marquardt damping and has the effect of reducing the size of the $\boldsymbol{\gamma}$-step and rotating it slightly in the direction of U_R. The damping parameter is increased as required during the iteration to prevent the REML likelihood from decreasing. If the scoring step (3) decreases the REML likelihood at first try, then λ is decreased by a pre-determined factor for the next iteration. If A is chosen to be positive definite, then the algorithm is a generalization of the well-known Levenberg-Marquardt algorithm for nonlinear least squares, for which strong global convergence results exist (Fletcher, 1987, Sections 5.1, 5.2 and 6.1). The algorithm given here can therefore be expected to be globally convergent to a solution of the REML equations subject to fairly standard regularity conditions, most importantly that the eigenvalues of A are bounded away from zero. In practice, the algorithm converges unless rounding errors in floating point arithmetic make the desired level of precision impossible to achieve. In the latter case the iteration is terminated when λ becomes too large.

Initialization

1. Initialize d_i as the squared residuals from regression of $\mathbf{y}$ on X with weights w_i. Compute the leverages h_{ii} from this regression. (The leverages are usually best obtained from $h_{ii} = \sum_{j=1}^{p} q_{ij}^2$ where the q_{ij} are the elements of Q from the QR-decomposition of $\mathrm{diag}(w_i)X$.)

2. Compute $\boldsymbol{\gamma}_0$ as the regression coefficient vector from the linear regression of $g^{-1}\{d_i/(1-h_{ii})\}$ on Z with weights $1-h_{ii}$, and initialize $\sigma_i^2 = g^{-1}(\mathbf{z}_i^T \gamma_0)$. (Note that $\boldsymbol{\gamma}_0$ is defined even if some $h_{ii} = 1$ because any such point receives zero weight.)

3. Regress $\mathbf{y}$ on X with weights w_i/σ_i^2. Store the squared residuals as d_i and the leverages as h_{ii}. The QR-decomposition of the matrix $\Sigma_m^{-1/2} X$ will be produced as a by-product of this regression — store the upper triangular factor as R.

4. Compute the REML deviance (minus twice the log-likelihood) as

$$D = \sum_{i=1}^{n} \left\{ w_i d_i/\sigma_i^2 + \log(\sigma_i^2/w_i) + \log(2\pi) \right\} + 2 \log |R|.$$

   (Note that $|R|$ is the product of the diagonal elements of R as R is a triangular matrix.)

5. Initialize λ. I have used $\lambda = \mathrm{tr}A/q$, which is a simple crude estimator of the smallest eigenvalue of A.

Scoring Iteration

6. Using the Choleski decomposition of $A + \lambda I$ or otherwise, solve $(A + \lambda I)\boldsymbol{\delta} = U_R$ for $\boldsymbol{\delta}$ where A is an approximation to $\mathcal{I}_R$. Compute $\boldsymbol{\gamma}_{k+1} = \boldsymbol{\gamma}_k + \boldsymbol{\delta}$ and update the σ_i^2.

7. Regress $\mathbf{y}$ on X with weights w_i/σ_i^2. Update the d_i, the h_{ii} and R and compute the REML deviance D as in 4.

8. If D has decreased, then decrease λ by a suitable decrease factor and go on to the next step. If D has increased or has remained constant then increase λ by an increase factor and go back to 6. (I use increase factor 2 and decrease factor 10.)

9. If $\boldsymbol{\delta}^T U_R < \epsilon$ then accept convergence and go to 11. (I use $\epsilon = 10^{-5}$ by default.)

10. If $\lambda > m\,10^{16}$ where m is the maximum diagonal element of A then accept that further increase of the likelihood is limited by rounding error and go to 11. Otherwise return to 6.

Wrap-up

11. Compute standard errors for γ as the square-root diagonal elements of $\mathcal{I}_R$.

The use of the convergence criterion $\boldsymbol{\delta}^T U_R$ in step 9 above is justified by the fact that this quantity converges to $U_R^T \mathcal{I}_R U_R$, which has the form of a score test statistic for testing hypotheses about $\boldsymbol{\gamma}$. The criterion is equal to zero at $\hat{\boldsymbol{\gamma}}$ and is otherwise positive. Equally importantly, it is a well-scaled quantity for $\boldsymbol{\gamma}$ near $\hat{\boldsymbol{\gamma}}$ so that comparison with a constant cut-off value is meaningful.

In principle any reasonable approximation A can be used for $\mathcal{I}_R$. The algorithm above represents REML scoring if $A = \mathcal{I}_R$. Other reasonable choices for A are $A = Z^{*T} V_1 Z^*$ or $A = Z^{*T} V_2 Z^*$ where $V_j = \text{diag}\,(1 - h_{ii})^j$. The last approximation has been used by Cook & Weisberg (1983), Verbyla (1993) and Smyth & Verbyla (1999). Any diagonal approximation such as V_1 or V_2 has the effect of decreasing slightly the computational burden at each iteration compared with REML scoring but at the likely cost of incurring extra iterations. Smyth et al. (2001) argue that of the two approximations, V_1 often provides a better approximation to $\mathcal{I}_R$ than does V_2. An S-Plus function implementing REML scoring is available from www.statsci.org/~gks/s/. An R function is in the statmod package available from CRAN (www.r-project.org).

6 Example: Welding Strength

The data give the results of an off-line screening experiment for factors affecting welding quality conducted by the National Railway Corporation of Japan (Taguchi & Wu, 1980). The response variable is the observed tensile strength of the weld, one of several quality characteristics measured. There are nine two-level factors [A–I, following

Bergman & Hynén (1997)] in an unreplicated experiment of 16 runs. The data have been considered previously by Box & Meyer (1986a), Box & Meyer (1986b), Bergman & Hynén (1997), Nelder & Lee (1998) and Huele & Engel (1998). The limited amount of available data in relation to the number of explanatory factors means that reliable computation of REML estimates and standard errors requires special care.

We consider here three mean-dispersion models for which results are given Nelder & Lee (1998) and Huele & Engel (1998). The first is the model with all 9 main effects in the mean model and factors C, H and I in the dispersion model. Here C indicates welded material, H welding method and I preheating. The mean model is an ordinary additive linear model with 9 factors and the dispersion model is the log-linear model

$$\log \sigma^2 = \gamma_0 + \gamma_C c + \gamma_H h + \gamma_I i \tag{4}$$

Each indicator variable is coded as 1 for the high level of the factor and 0 for the low. Nelder & Lee (1998) give estimates and standard errors in their Table 3, and conclude from the standard errors for the dispersion model that factor H does not have a significant effect on the variance. Huele & Engel (1998) estimate the same model using a different fitting algorithm and get similar results for the mean model but different estimates and standard errors for the dispersion model. Both Nelder & Lee (1998) and Huele & Engel (1998) give standard errors for $\hat{\boldsymbol{\gamma}}$ based on approximations to $\mathcal{I}_R$. Unfortunately, computing the actual REML information matrix for this model shows that the dispersion model is singular. With all main effects in the mean model, the matrix V has three zero eigenvalues and is of rank 13 rather than rank 16. The design matrix Z for the 3-factor dispersion model has a range space with overlaps the null space of V, so that $\mathcal{I}_R$ has one zero eigenvalue. This has the effect that the γ_j are not identifiable and do not have unique REML estimators. The values for the $\hat{\gamma}_j$ at convergence are an accident of the fitting algorithm, which explains why Nelder & Lee (1998) and Huele & Engel (1998) get different results. The standard errors given for the $\hat{\gamma}_j$ in the two papers are entirely illusionary, as the "correct" standard errors obtained from $\mathcal{I}_R$ are actually infinite for all four dispersion parameters. For this model no diagonal approximation to V can produce an adequate approximation for $\mathcal{I}_R$.

The second model considered replaces the saturated main-effect model for the mean with

$$\mu = \beta_0 + \beta_B b + \beta_C c$$

where B indicates period of drying. The dispersion model is the same as before (4). This was the final model found by Bergman & Hynén (1997) using a graphical method of analysis. Carroll & Ruppert (1988) reported convergence problems with maximum likelihood estimation for this data and a similar model. Nelder & Lee (1998) found divergence when attempting to compute either the maximum likelihood or the REML estimators for this model. Only Huele & Engel (1998) have previously succeeded in computing REML estimates for this model. The algorithm given in Section 4 converges to the REML estimates without the need to increase λ at any point. With $A = \mathcal{I}_R$ the

Table 1: Estimates and standard errors for the welding-strength data model (4).

	Estimates	Standard errors		
	$\hat{\gamma}_j$	REML	V_1	V_2
Intercept	-3.15891	0.83131	0.81881	0.94816
C	-2.73544	0.82248	0.81048	0.96280
H	-0.08603	0.83509	0.81048	0.96280
I	3.33259	0.82502	0.81048	0.96280

algorithm requires 9 iterations to converge with $\epsilon = 10^{-5}$. With A derived from $V \approx V_1$ the algorithm requires 11 iterations to reach the same convergence criterion and with A derived from V_2 it requires 16 iterations. The Levenberg-Marquardt damping strategy is not actually needed for convergence in this case, but it also does not cost anything as the unmodified REML scoring algorithm requires the same number of iterations to reach the same precision.

The final estimates and standard errors are given in Table 1. The table gives the REML estimates for the dispersion model together with standard errors obtained from (i) the REML information matrix, (ii) approximating V with V_1 and (ii) approximating (iii) approximating V with V_2. The standard errors based on V_1 are slight under-estimates while those base on V_2 are about 16% too high. The final value for minus twice the REML log-likelihood is $D = 14.00547$. The REML estimates and standard errors agree with those given by Huele & Engel (1998) to 3 significant figures. Huele & Engel (1998) have presumably computed the full REML information at convergence of the iteration using explicit evaluation of V, as recommended in Section 3 of Huele et al. (2000).

Finally we consider the model

$$\mu = \beta_0 + \beta_B b + \beta_C c + \beta_I i \tag{5}$$

and

$$\log \sigma^2 = \gamma_0 + \gamma_C c + \gamma_I i \tag{6}$$

for which both Nelder & Lee (1998) and Huele & Engel (1998) give estimates. Here B indicates period of drying. Results are given in Table 2. Again the standard errors based on V_1 are slight under-estimates while those based on V_2 are over-estimates. Minus twice the REML log-likelihood is 14.14072. Nelder & Lee (1998) and Huele & Engel (1998) both give correct estimates to two or three decimal places. Neither Nelder & Lee (1998) nor Huele & Engel (1998) state explicitly how they have computed standard errors, but comparison with Table 2 shows that Nelder & Lee (1998) are using the approximation based on V_1 while Huele & Engel (1998) are using the correct REML information matrix.

Table 2: Estimates and standard errors for the welding-strength data model (6).

	Estimates	Standard errors		
	$\hat{\gamma}_j$	REML	V_1	V_2
Intercept	-3.06385	0.71992	0.71216	0.83372
C	-3.03748	0.83885	0.82624	0.96440
I	2.90415	0.84022	0.82598	0.96321

References

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Appl. Statist.* **36**, 332–339.

Bergman, B. & Hynén, A. (1997). Dispersion effects from unreplicated designs in the 2^{k-p} series. *Technometrics* **39**, 191–198.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformation (with discussion). *J. R. Statist. Soc.* B **26**, 211–252.

Box, G. E. P. & Meyer, R. D. (1986a). An analysis for unreplicated fractional factorials. *Technometrics* **28**, 11–18.

Box, G. E. P. & Meyer, R. D. (1986b). Dispersion effects from fractional designs. *Technometrics* **28**, 19–27.

Carroll, R. J. & Ruppert, D. (1988). *Transformations and weighting in regression.* Chapman and Hall.

Cook, R. D. & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70**, 1–10.

Davidian, M. & Carroll, R. J. (1987). Variance function estimation. *J. Amer. Statist. Ass.* **82**, 1079–1091.

Engel, J. & Huele, A. F. (1996). A generalized linear modeling approach to robust design. *Technometrics* **38**, 365–373.

Fletcher, R. (1987). *Practical Methods of Optimization.* John Wiley & Sons, Chichester, second edition.

Golub, G. H. & Van Loan, C. F. (1996). *Matrix Computations.* Johns Hopkins University Press, Baltimore, third edition.

Harvey, A. C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econometrica* **44**, 460–465.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika* **61**, 383–385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and related problems. *J. Am. Statist. Assoc.* **72**, 320–40.

Huele, A. F. (1998). *Statistical Robust Design.* PhD Thesis, Faculteit der Wiskunde, Informatica, Natuurkunde en Sterrenkunde, Korteweg-de Vries Instituut voor

Wiskunde, Amsterdam.

Huele, A. F. & Engel, J. (1998). Response to: joint modelling of mean and dispersion by nelder and lee. *Technometrics* **40**, 171–175.

Huele, A. F., Schoen, E. D., & Steeman, R. A. (2000). A note on REML estimation in the heteroscedastic linear model. Report 100286, CQM B. V., Eindhoven, The Netherlands.

Lee, Y. & Nelder, J. A. (1998). Generalized linear models for the analysis of quality-improvement experiments. *Canad. J. Statist.* **26**, 95–105.

Myers, R. H. & Montgomery, D. C. (1995). *Response Surface Methodology: Process Product Optimization using Designed Experiments.* Wiley, New York.

Nair, V. N. & Pregibon, D. (1988). Analyzing dispersion effects from replicated factorial experiments. *Technometrics* **30**, 247–257.

Nelder, J. A. & Lee, Y. (1991). Generalized linear models for the analysis of taguchi-type experiments. *Applied Stochastic Models and Data Analysis* **7**, 107–120.

Nelder, J. A. & Lee, Y. (1998). Letters to the editor: joint modelling of mean and dispersion. *Technometrics* **40**, 168–171.

Nelder, J. A. & Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika* **74**, 221–232.

Osborne, M. R. (1987). Estimating nonlinear models by maximum likelihood for the exponential family. *SIAM J. Sci. Statist. Comput.* **8**, 446–456.

Park, R. E. (1966). Estimation with heteroscedastic error terms. *Econometrica* **34**, 888.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **58**, 545–554.

Rutemiller, H. C. & Bowers, D. A. (1968). Estimation in a heteroscedastic regression model. *J. Amer. Statist. Ass.* **63**, 552–557.

Smyth, G. K. (1989). Generalized linear models with varying dispersion. *J. R. Statist. Soc.* B **51**, 47–60.

Smyth, G. K., Huele, A. F., & Verbyla, A. P. (2001). Exact and approximate REML for heteroscedastic regression. *Statistical Modelling* **1**, 161–175.

Smyth, G. K. & Verbyla, A. P. (1999). Adjusted likelihood methods for modelling dispersion in generalized linear models. *Environmetrics* **10**, 696–709.

Taguchi, G. & Wu, Y. (1980). *Introduction to Off-Line Quality Control.* Central Japan Quality Control Association, Nagoya, Japan.

Thisted, R. A. (1988). *Elements of Statistical Computing.* Chapman and Hall, New York.

Tunnicliffe Wilson, G. (1989). On the use of marginal likelihood in time series model estimation. *J. R. Statist. Soc.* B **51**, 15–27.

Verbyla, A. P. (1993). Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *Journal of the Royal Statistical Society, Series B* **55**, 493–508.