Exact and Approximate REML for Heteroscedastic Regression[*]

Gordon K. Smyth[†]
Department of Mathematics, University of Queensland
Australia

A. Frederik Huele
Centre for Quantitative Methods B. V.
Eindhoven, The Netherlands

Arūnas P. Verbyla
BiometricsSA, University of Adelaide and
South Australian Research and Development Institute
Australia

12 September 2001

Abstract

Exact REML for heteroscedastic linear models is compared with a number of approximate REML methods which have been proposed in the literature, especially with the methods proposed by Lee & Nelder (1998) (LN98) and Smyth & Verbyla (1999) (SV99) for simultaneous mean-dispersion modelling in generalized linear models. It is shown that neither of the LN98 or SV99 methods reduces to REML in the normal linear case. Asymptotic variances and efficiencies are obtained for these and other estimators of the same general form. A new algorithm is suggested, similar to one suggested by Huele et al. (2000), which returns the correct REML estimators and an improved approximation to the standard errors. It is possible to obtain REML estimators by alternating between two generalized linear models but the final fitted generalized linear model objects will not return the correct standard errors for the variance coefficients. The true REML likelihood calculations therefore fit only partially into the double generalized linear model framework.

Keywords: Maximum likelihood; Modified profile likelihood; Residual maximum likelihood; Restricted maximum likelihood; Sandwich estimator.

[*] *Statistical Modelling* **1**, 161–175 (2001).

[†] Address for correspondence: Walter and Eliza Hall Institute of Medical Research, Post Office, Royal Melbourne Hospital, Australia 3050

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 considered is a slight generalization of that considered by Verbyla (1993) in which it was assumed that $g = \log$. An efficient numerical implementation of REML for this model has been described by Smyth (2002). Our purpose in this article is to compare exact REML with a number of approximate REML methods which have been proposed in the literature, especially with the methods proposed by Lee & Nelder (1998) (LN98) and Smyth & Verbyla (1999) (SV99) for simultaneous mean-dispersion modelling in generalized linear models. Since REML is strictly applicable only to normal linear models, the above heteroscedastic regression model is the most general model of the type considered by LN98 and SV99 for which exact REML methods are possible.

We show that neither of the LN98 or SV99 methods reduces to REML in the normal linear case. We obtain asymptotic variances and efficiencies for these and other estimators of the same general form. We consider the question of convergence of iterative estimation algorithms and show how to avoid the convergence problems reported by Nelder & Lee (1998), Huele (1998) and Huele & Engel (1998). We suggest a new algorithm which is a compromise of LN98 and SV99 and similar to one suggested by Huele et al. (2000). The new algorithm returns the correct REML estimators and an improved approximation to the standard errors.

The LN98 and SV99 methods were motivated by a desire to estimate the heteroscedastic regression model by way of two coupled generalized linear models, as in Nelder & Pregibon (1987), Aitkin (1987), Smyth (1989) and Nelder & Lee (1991). This approach yields significant advantages in terms of being able to use established generalized linear model diagnostics, concepts and software — see especially SV99 who created a *double generalized linear model* object in S-Plus for this purpose. The first generalized linear model is in this case just a linear regression with weights w_i/σ_i^2 and is used to estimate the mean coefficient vector $\boldsymbol{\beta}$. The second generalized linear model has as responses the squared residuals or squared standardized residuals from the first model and estimates the variance coefficient vector $\boldsymbol{\gamma}$. Aitkin (1987) and Smyth (1989) showed that maximum likelihood estimates for all parameters may be obtained by alternating between two generalized linear models in this way. This paper shows that REML estimation is unfortunately not as straightforward. Although it is possible to obtain REML estimators by alternating between two generalized linear models, the final fitted generalized linear model objects will not return the correct standard errors for the variance coefficients — in other words, the uncertainty in estimating the variances is not correctly assessed without further special purpose calculation. The true REML likelihood calculations therefore do not fit neatly into the double generalized linear model framework.

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. An up-to-date review of REML etimation can be found in McCulloch & Searle (2001).

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. The use of heteroscedastic regression is now common practice in industrial statistics for analyzing unreplicated experiments. 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). In Section 8 we consider an off-line screening experiment which examines the effect of factors on the tensile strength of welds. In this type of experiment the estimated mean model is used to choose factor settings to maximise the weld strength while the variance model is used to minimize the sensitivity of the production process to variation in uncontrolled factors.

In the next two sections we summarize the maximum likelihood and REML estimating equations for the heteroscedastic regression model. Implementation details for the scoring iteration to compute the REML estimates are discussed in Section 4. REML estimation is contrasted with various approximate REML methods in Section 5. The efficiences of the non-REML estimators are examined in Section 6 and the accuracy of various approximations to the REML standard errors are studied in Section 7. A data example is given in Section 8 and the paper concludes with recommendations in Section 9.

2 Maximum Likelihood Estimation

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 and information matrix for $\boldsymbol{\beta}$ are

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

and

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

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 two parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal. The score vector and information matrix for $\boldsymbol{\gamma}$ are

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

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, $\Sigma_d = \operatorname{var}\mathbf{d} = \operatorname{diag}(2\sigma_i^4)$, $G = \operatorname{diag}\{1/\dot{g}(\sigma_i^2)\}$, $W_d = G^2\Sigma_d^{-1}$ and Z is the $n \times q$ design matrix with ith row $\mathbf{z}_i^T$. The method of scoring for computing the maximum likelihood estimates yields

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

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

$$\boldsymbol{\gamma}_{k+1} = \left(Z^T W_d Z\right)^{-1} Z^T W_d^{-1} \mathbf{z}_d \qquad (2)$$

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 a 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.

Smyth (1989) and Smyth (1996) have considered algorithmic strategies for improving on the scoring iteration for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. It turns out the convergence properties of the iteration are improved if $\boldsymbol{\gamma}$ is updated as in (2) before $\boldsymbol{\beta}$ and then the current $\boldsymbol{\gamma}_{k+1}$ is used in (1) instead of $\boldsymbol{\gamma}_k$. Smyth (1996) calls this a *nested* iteration and shows that it improves both the global convergence properties of the iteration and the eventual rate of convergence as the iteration approaches the solution.

It is possible to compute the maximum likelihood estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ by alternately fitting a linear regression with responses y_i and a gamma generalized linear model with responses d_i. The linear regression updates $\boldsymbol{\beta}$ as in (1) while each iteration of the gamma generalized linear model fit corresponds to an update of the form (2). In this approach it is not necessary to iterate the gamma generalized linear model to convergence at each step as has been done by many authors with related models. This correponds to updating $\boldsymbol{\gamma}$ through (2) many times before returning to (1) and introduces a unnecessary inner iteration which can itself introduce convergence problems. The update (2) should instead be implemented as a one-step gamma generalized linear model fit initialized at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_k$, and this results in the nested iteration described above.

3 Residual Maximum Likelihood

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

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

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

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

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

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

where $Z^* = \Sigma_d^{-1/2}GZ$ and V is an $n\times n$ 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 standardized 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

$$\boldsymbol{\gamma}_{k+1} = \boldsymbol{\gamma}_k + \mathcal{I}_R^{-1}\mathbf{u}_R \tag{3}$$

with $\mathcal{I}_R$ and $\mathbf{u}_R$ computed at $\boldsymbol{\gamma}_k$ and $\hat{\boldsymbol{\beta}}(\boldsymbol{\gamma})$.

4 Implementation of REML Scoring

Smyth (2002) describes how to compute the REML estimators for the heteroscedastic regression model considered in this paper in $O(n)$ operations including computation of the REML log-likelihood ℓ_R and information matrix $\mathcal{I}_R$. The algorithm includes a convergence check in order to ensure that the REML likelihood increases at each iteration. 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

$$\boldsymbol{\gamma}_{k+1} = \boldsymbol{\gamma}_k + (A + \lambda I)^{-1}\mathbf{u}_R \tag{4}$$

is an ascent step for the REML likelihood. 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 $\mathbf{u}_R$. The damping parameter is increased as required during the iteration to prevent the REML likelihood from decreasing. If the scoring step (4) increases the REML likelihood at first try, then λ is decreased by a predetermined factor for the next iteration, so that λ can be expected to approach zero as the iteration homes in on the solution.

In principle any reasonable approximation A can be used for $\mathcal{I}_R$. If $\lambda = 0$ and $A = \mathcal{I}_R$ then the algorithm (3) is the same as REML scoring (4). Other reasonable choices for A are $A = Z^{*T} V_1 Z^*$ or $A = Z^{*T} V_2 Z^*$ with $V_j = \operatorname{diag}\{(1-h_{ii})^j\}$. Either of these choices has the effect of decreasing slightly the computational burden of each iteration but at the likely cost of incurring extra iterations compared with $A = \mathcal{I}_R$. The approximation $V \approx V_2$ has been used previously by Cook & Weisberg (1983), Verbyla (1993) and Smyth & Verbyla (1999). However in Section 7 we will show that $V \approx V_1$ often produces better results.

Huele et al. (2000) point out that the score equation $\mathbf{u}_R = 0$ can be solved by repeatedly fitting a gamma generalized linear model with responses $d_i/(1-h_{ii})$, link $g()$ and prior weights $1-h_{ii}$. The d_i and the h_{ii} are updated at each iteration from the updated $\boldsymbol{\gamma}$ by fitting a normal linear regression to the reponses y_i. This is similar to the strategy of LN98 except that LN98 do not specify the prior weights $1-h_{ii}$. Alternatively one could use responses $d_i + h_{ii}\sigma_i^2$ and prior weights unity, with again d_i and h_{ii} updated at each iteration as in Huele (1998) and Huele & Engel (1998). Both 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 than the weighted in this respect), although these may be computed from $\mathcal{I}_R$ at convergence of the iteration. Neither of these iterative strategies — alternating between a normal generalized model for y_i and a gamma generalized linear model for d_i — can be guaranteed to converge, although either could be modified to correct this using Levenberg-Marquardt damping as in (4) or otherwise. In his implementation, Huele (1998) iterates the gamma generalized linear model to convergence for each value of d_i and h_{ii}. This has the effect, rather than of updating d_i and h_{ii} at each iteration, of introducing an extra inner iteration for each fixed value of d_i and h_{ii}. This inner iteration is unnecessary and introduces convergence problems as Huele (1998) reports on page 30. The gamma step is better implemented as a one-step generalized linear model fit initialized at $\boldsymbol{\gamma} = \boldsymbol{\gamma}_k$. When correctly implemented, and modified to ensure convergence, the alternating double generalized linear model strategy with a one-step gamma step is equivalent to the modified REML scoring iteration (4) with $A = Z^{T*} V_1 Z^*$ (for the weighted gamma strategy) or $A = Z^{T*} Z^*$ (for the unweighted gamma strategy).

5 Comparison with Other Methods

The REML idea of eliminating the mean parameters from the likelihood by considering the marginal distribution of zero-mean contrasts of the responses is strictly applicable only to normal linear models. There have been many attempts to extend the REML idea to more general models. Some of these methods are equivalent to REML for the heteroscedastic linear model considered in this paper, for example those of Cox & Reid (1987), McCullagh & Tibshirani (1990) and Smyth & Verbyla (1996). Each of these papers describe estimation principles which reduce to REML in the normal linear case, but none of them give detailed guidance as to how to organize the computational and inference procedures for specific models.

Several authors have developed special-purpose approximate REML methods for dispersion modelling in generalized linear models. Letting μ_i denote the mean and σ_i^2 the dispersion for the ith observation, link-linear models are assumed for both the means

$$f(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{5}$$

where $f()$ is a known monotonic differentiable function, and the dispersions

$$g(\sigma_i^2) = \mathbf{z}_i^T \boldsymbol{\gamma}. \tag{6}$$

This is a generalization of the heteroscedastic regression model considered in this paper. In many cases it is possible to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ simultaneously by alternating between two generalized linear models, one with the original observations as reponses in which $\boldsymbol{\beta}$ is estimated with $\boldsymbol{\gamma}$ fixed and one with the unit deviances as responses in which $\boldsymbol{\gamma}$ is estimated with $\boldsymbol{\beta}$ fixed. We refer to such cases as double generalized linear models. Maximum likelihood estimation for double generalized linear models was introduced by Smyth (1989). Approximate inference using extended quasi-likelihood (EQL) was introduced by Nelder & Pregibon (1987) and Nelder & Lee (1991). Approximate REML methods for double generalized linear models were proposed by Smyth & Verbyla (1996), LN98 and SV99. Nelder & Lee (1992) also used a simple degree of freedom adjustment when using EQL with single samples.

LN98 suggest estimating $\boldsymbol{\gamma}$ by minimising[1]

$$\sum_{i=1}^{n} \frac{d_i}{(1-h_{ii})\sigma_i^2} + \log \sigma_i^2 \tag{7}$$

with the idea that this leads to estimates similar to the REML estimates. They suggest a specific algorithm, which is to repeatedly fit a gamma regression to responses $d_i/(1-h_{ii})$. This is motivated by the fact that $d_i/(1-h_{ii})$ is unbiased in the normal linear case for σ_i^2. This algorithm is similar to one of the REML algorithms described in the previous section except that LN98 do not specify the use of prior weights $1-h_{ii}$. The LN98 algorithm therefore does not reduce to REML in the normal linear case and fails to have one of the key REML characteristics that observations with large leverage in the mean model are downweighted. The LN98 algorithm actually does not mininize the objective function (7), in part because the derivatives of $1/(1-h_{ii})$ have been neglected. Neither the algorithm nor the objective function are equivalent to REML in the normal linear case. Even

[1] LN98 contains a misprint "maximize" in place of "minimize".

though their published algorithm is not equivalent to REML, LN98 and Nelder & Lee (1998) do in fact present correct REML estimates in their data examples. A personal communication from Nelder (1999) to Verbyla confirms that they have in fact implemented the gamma regression with prior weights $1 - h_{ii}$ as suggested independently by Huele et al. (2000). In this paper, for the purposes of comparison, the "LN98 algorithm" will refer to the published algorithm of LN98 which does not have prior weights rather than to the implemented algorithm while the same algorithm with prior weights $1 - h_{ii}$ will be referred to as the weighted algorithm of Huele et al. (2000).

SV99 suggest estimating $\boldsymbol{\gamma}$ by gamma regression of $d_i + h_{ii}\sigma_i^2$ on Z, a procedure which leads to correct REML estimators. They go on to modify the prior weights in order to improve the estimation of the standard errors. The modified prior weights have in general a complicated form but reduce to $(1-h_{ii})^2$ for normal linear models. Since REML would arise from unit prior weights it is apparent that the final SV99 algorithm is not equivalent to REML either.

We have had problems with making the LN98 algorithm converge on some data sets. Since the algorithm does not actually minimize the objective function (7) it is not possible to include a line search or Levenberg-Marquardt damping as in Section 4 to secure convergence. The same problem exists also for the SV99 algorithm. Since it is not easy to characterise what objective function is being maximized by the algorithms, it is not possible to modify the iteration step to ensure an increase of that objective function at each iteration. In our implementations of the LN98 and SV99 algorithms, we have stopped the iteration when the REML likelihood stops increasing. The resulting numerical estimators however are not exactly REML and it is very difficult to characterize their exact properties. Another problem is that LN98 appear to iterate the gamma generalized model to convergence between updates of d_i and h_{ii}. This inner iteration is unnecessary for reasons described earlier in this paper and may be a cause for the convergence problems reported in Nelder & Lee (1998).

A very general paper on variance modelling is Davidian & Carroll (1987) in which mean-variance models are considered without explicit specification of the response distribution. In our notation they suggest estimating $\boldsymbol{\gamma}$ by minimizing

$$\sum_{i=1}^{n} \frac{\{d_i - (1 - h_{ii})\sigma_i^2\}^2}{(1 - h_{ii})^2\sigma_i^4}. \tag{8}$$

This is related to the idea of approximating V with V_2 as above. However this approach treats the d_i as approximately normal and symmetric whereas their actual distribution is highly skew. It is therefore likely to be less efficient that any of the other methods considered here. Davidian & Carroll (1987) do not give a specific algorithm by which the minimization of (8) might be achieved.

6 Asymptotic Variances and Efficiences

The estimating equations for the REML, LN98 and SV99 estimators can be written in the form

$$\mathbf{u} = Z^T W_d(\mathbf{d} - \boldsymbol{\sigma}^{*2}) \tag{9}$$

where $\boldsymbol{\sigma}^{*2} = \{(1 - h_{ii})\sigma_i^2\}$ and $W_d = \mathrm{diag}\{w_{di}/[2\dot{g}(\sigma_i^2)\sigma_i^4]\}$ and where the w_{di} are functions of the σ_i^2, X and Z but not $\boldsymbol{\beta}$. If $w_{di} = 1$ then $\mathbf{u}$ is the REML score vector and the solution of $\mathbf{u} = 0$ is the REML estimator of $\boldsymbol{\gamma}$. The LN98 estimator correponds to $w_{di} = 1/(1 - h_{ii})$ while the SV99 estimator corresponds to $w_{di} = (1 - h_{ii})^2$.

Under standard regularity conditions the asymptotic variance of the estimator $\tilde{\boldsymbol{\gamma}}$ defined by (9) is given by the robust sandwich estimator

$$\mathrm{var}\,\tilde{\boldsymbol{\gamma}} \approx \left[E\frac{\partial \mathbf{u}}{\partial \boldsymbol{\gamma}^T}\right]^{-1} \mathrm{var}\,\mathbf{u} \left[E\frac{\partial \mathbf{u}^T}{\partial \gamma}\right]^{-1} \tag{10}$$

This expression is based on a first order Taylor series expansion of the estimating equation (9) at $\tilde{\boldsymbol{\gamma}}$ about the true value of $\boldsymbol{\gamma}$ (Huber, 1967, Liang & Zeger, 1986). When $w_{di} = 1$ the variance (10) reduces to $(\mathrm{var}\,\mathbf{u})^{-1} = \mathcal{I}_R^{-1}$ which is the usual inverse REML information matrix. Since $\mathbf{d}$ has covariance matrix V it is straightforward to compute that

$$E\frac{\partial \mathbf{u}}{\partial \boldsymbol{\gamma}^T} = \frac{1}{2} Z^T \mathrm{diag}\left\{\frac{w_{di}}{\dot{g}(\sigma_i^2)\sigma_i^2}\right\} V\,\mathrm{diag}\left\{\frac{1}{\dot{g}(\sigma_i^2)\sigma_i^2}\right\} Z$$

and

$$\mathrm{var}\,\mathbf{u} = \frac{1}{2} Z^T \mathrm{diag}\left\{\frac{w_{di}}{\dot{g}(\sigma_i^2)\sigma_i^2}\right\} V\,\mathrm{diag}\left\{\frac{w_{di}}{\dot{g}(\sigma_i^2)\sigma_i^2}\right\} Z$$

The variance (10) is minimized with respect to w_d by the REML values $w_{di} = 1$, because the inverse REML information matrix provides the lower variance bound for unbiased estimators based on the fitted residuals and their distribution. Numerical experiments show that neither of the LN98 or SV99 estimators is uniformly superior to the other, depending on the specific values of X, Z and $\boldsymbol{\gamma}$, and that both are usually not far from the REML estimator. The LN98 estimator is often more efficient than the SV99 estimator. However the SV99 estimator is preferable when there are one or more highly leveraged observations with large values of h_{ii}.

6.1 Multiple Regression

It is instructive to consider the case of ordinary multiple regression in which the variance is actually constant. The y_i are assumed normal with mean $\mathbf{x}_i^T\boldsymbol{\beta}$ and constant variance σ^2. The maximum likelihood estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where $\hat{\mu}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}}$ while the REML estimator is the well known sample variance

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} d_i = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The LN98 estimator for σ^2 is

$$\frac{1}{n} \sum_{i=1}^{n} d_i/(1 - h_{ii})$$

while the SV99 estimator is

$$\frac{1}{\sum_{i=1}^{n}(1 - h_{ii})^3} \sum_{i=1}^{n} (1 - h_{ii})^2 d_i.$$

It is easily seen that neither LN98 nor SV99 give the REML estimator, although the differences are often not great in practice. The LN98 method up-weights residuals with high leverages while SV99 heavily down-weights them. The Davidian & Carroll (1987) estimator is equivalent to LN98 in this case. All of the REML-type estimators are unbiased for σ^2.

The REML information for σ^2 is

$$\mathcal{I}_R = \frac{1}{2\sigma^4} \mathbf{1}^T V \mathbf{1} = \frac{n-p}{2\sigma^4}$$

If one uses the standard errors for $\boldsymbol{\gamma}$ from the gamma generalized linear model in the LN98 method then the information for $\boldsymbol{\gamma}$ is estimated to be the same as that for ML with known means, namely $n/(2\sigma^4)$. However the variance of the LN98 estimator is actually

$$\frac{2\sigma^4}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{(1 - h_{ii})(1 - h_{jj})} V_{ij} = \frac{2\sigma^4}{n} \left\{ 1 + \frac{1}{n} \sum_{i=1}^{n} \sum_{j \neq i} \frac{h_{ij}^2}{(1 - h_{ii})(1 - h_{jj})} \right\}$$

where the V_{ij} are the elements of V. The SV99 method estimates the information matrix to be

$$\frac{1}{2\sigma^4} \mathbf{1}^T V_2 \mathbf{1} = \frac{\sum_{i=1}^{n}(1 - h_{ii})^2}{2\sigma^4}$$

whereas the variance of the SV99 estimator is actually

$$\frac{2\sigma^4}{\{\sum_{i=1}^{n}(1 - h_{ii})^3\}^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (1 - h_{ii})^2 (1 - h_{jj})^2 V_{ij}$$

$$= \frac{2\sigma^4}{\sum_{i=1}^{n}(1 - h_{ii})^3} \left\{ 1 + \frac{1}{\sum_{i=1}^{n}(1 - h_{ii})^3} \sum_{i=1}^{n} \sum_{j \neq i} h_{ij}^2 (1 - h_{ii})^2 (1 - h_{jj})^2 \right\}$$

Variances for the LN98 and SV99 estimators for two example linear regressions (with $n = 5$ and $p = 2$) are given in Table 1. The variances are given relative to the lower variance bound achieved by the REML estimator. In the first example none

Table 1: True and estimated variances for the LN98 and SV99 estimators relative to REML for estimating σ^2 from two linear regressions with $n = 5$ and $p = 2$. The variances are divided by $2\sigma^4/(n-p)$.

	True Variance		Estimated Variance	
x	LN98	SV99	LN98	SV99
1 1 3 5 5	1.016	1.105	0.6	1.622
1 2 3 4 20	1.078	1.005	0.6	1.345

of the leverages h_{ii} are very extreme and LN98 is close to fully efficient while SV99 is not. In the second example one of the leverages h_{ii} is much larger than the others and SV99 is close to fully efficient while LN98 is not. In general SV99 becomes fully efficient as $\max h_{ii} \to 1$, meaning that one of the observations is associated with a extreme value for x. Both LN98 and SV99 estimators become fully efficient as $\max h_{ii} \to 0$, in which case the estimated mean model converges to the true model, or as the h_{ii} become equal, corresponding to a balanced design. The REML and SV99 estimators remain defined even if $h_{ii} = 1$ for some i while the LN98 do not.

Also given in Table 1 are the nominal variances for the estimators provided by the gamma generalize linear model fit, again relative to the variance of the REML estimator. The LN98 method consistently underestimates the true sampling variance because it uses the ML information bound. The SV99 method over-estimates the sampling variability in each of these examples.

7 Approximating the Information Matrix

Several authors have suggested approximating the matrix V in the computation of ℓ_R by the diagonal matrix V_2 which shares the same diagonal elements. Numerical experiments show that the approximation $Z^{T*}V_2Z^*$ may be greater or lesser than $Z^{T*}VZ^*$ depending on the values for X, Z and $\boldsymbol{\gamma}$. This can be understood through the following reasoning. Both V and V_2 have positive eigenvalues. Since the sum of the eigenvalues is equal to sum of the diagonal elements, the eigenvalues of V and V_2 have the same sum. Since of the sum of the squared eigenvalues is the sum of all the squared elements of the matrix, the squared eigenvalues of V have a greater sum than those of V_2. Since the eigenvalues of V have the same mean but a greater squared sum it follows that they are more varied that those of V_2 and in particular the largest eigenvalue of V is likely to be greater than that of V_2. We can write

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

where H^2 is the matrix with elements h_{ij}^2. It is shown by Smyth (2002) that the eigenvectors of H^2 lie in the span of the columns of X and the componentwise squares and products of these columns. Therefore the eigenvectors of V with largest eigenvalues are likely to be highly collinear with the columns of X or with their

Table 2: Estimates of $\boldsymbol{\gamma}$ for the Welding-Strength Data.

	REML	LN98	SV99
Intercept	-3.15886	-3.15891	-3.15889
C	-2.73543	-2.73544	-2.73528
H	-0.08589	-0.08602	-0.08582
I	3.33239	3.33259	3.33236

squares and products. We can conclude that $Z^{T*}V_2Z^*$ is likely to under-estimate $Z^{T*}VZ^*$ when the columns of Z^* lie in the span of the columns of X or their squares and products. On the other hand, if the columns of Z^* are uncorrelated with the columns of X or its squares and products then $Z^{T*}VZ^* = Z^{T*}\text{diag}(1 - 2h_{ii})Z^*$ which is uniformly smaller than $Z^{T*}V_2Z^*$. Hence using the diagonal V_2 in place of V will over-estimate the standard errors when the columns of Z are similar to those of X and will under-estimate them when Z is unrelated to X or to squares and products of its columns.

An alternative approximation is $Z^{T*}V_1Z^*$. This approximation has the virtue of giving exactly the correct result when $Z^* = \mathbf{1}$, i.e., when the variance is actually constant, a fact which follows from $\mathbf{1}^TV\mathbf{1} = \mathbf{1}^TV_1\mathbf{1} = n - p$. The approximation using V_1 gives consistently smaller standard errors than with V_2. The worst case for this approximation is when Z^* is unrelated to X. In that case V_1 has $1 - h_{ii}$ on the diagonal whereas $1 - 2h_{ii}$ would give the correct result. Hence in the worst case the approximation using V_1 is halfway between the REML and the ML information matrices. This worst case is unlikely to arise in practice though because both X and Z will generally include an intercept and therefore will be at least partially collinear. In practice the approximation based on V_1 is often better than the approximation based on V_2.

LN98 do not state how standard errors for the $\hat{\boldsymbol{\gamma}}$ are computed. Using standard errors output from the gamma generalized linear model used in their method would correspond to using the Fisher information $Z^{*T}Z^*$ in place of the REML information $Z^{*T}VZ^*$. The unweighted gamma regression used by Huele & Engel (1998) would give the same result. These standard errors would be under-estimated by a factor of about $\{(n-p)/n\}^{1/2}$. LN98 and Nelder & Lee (1998) appear to have recognized this problem and the standard errors given in their numerical examples appear to be based on the approximate information matrix $Z^{*T}V_1Z^*$. These are the same as would be given by the weighted gamma regression used by Huele et al. (2000). The SV99 algorithm approximates the REML information matrix by $Z^{*T}V_2Z^*$ and over-estimates the standard errors in most cases.

Table 3: Estimated covariance matrices for $\hat{\gamma}$ for the Welding-Strength Data. The REML estimate is the inverse of the REML information matrix ℓ_R. The ML estimate is the inverse of the Fisher information matrix $\frac{1}{2}Z^{*T}Z^*$. The other two estimates are the inverses of $Z^{*T}V_1Z^*$ and $Z^{*T}V_2Z^*$ respectively.

REML covariance matrix.

	Intercept	C	H	I
Intercept	0.691	-0.351	-0.357	-0.340
C	-0.351	0.676	0.174	-0.075
H	-0.357	0.174	0.697	-0.082
I	-0.340	-0.075	-0.082	0.681

ML covariance matrix.

	Intercept	C	H	I
Intercept	0.50	-0.25	-0.25	-0.25
C	-0.25	0.50	0.00	0.00
H	-0.25	0.00	0.50	0.00
I	-0.25	0.00	0.00	0.50

Approximation based on V_1.

	Intercept	C	H	I
Intercept	0.670	-0.335	-0.335	-0.335
C	-0.335	0.657	0.150	-0.068
H	-0.335	0.150	0.657	-0.069
I	-0.335	-0.068	-0.069	0.657

Approximation based on V_2.

	Intercept	C	H	I
Intercept	0.899	-0.450	-0.450	-0.450
C	-0.450	0.927	0.414	-0.220
H	-0.450	0.414	0.927	-0.222
I	-0.450	-0.220	-0.222	0.927

8 Example: Welding-Strength Data

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). We consider the factors B and C for the mean and factors C, H, I for the variance as found to be important by Bergman & Hynén (1997) using a graphical method of analysis. We consider a log-linear model for the variance following Lee & Nelder (1998) and Huele & Engel (1998). The model is

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

for the mean and

$$\log \sigma^2 = \gamma_0 + \gamma_C c + \gamma_H h + \gamma_I i$$

for the variance. Estimates were computed using the $(0, 1)$ factor coding of Nelder & Lee (1998) and Huele & Engel (1998). Nelder & Lee (1998) report divergence for this model but we were able to obtain convergence using a single-step gamma fit at each iteration. In this example there is considerable collinearity between the design matrices X and Z. The estimated values for $\boldsymbol{\gamma}$ returned by REML and by the LN98 and SV99 algorithms agree to 3 decimal places (Table 2) but there are noticeable differences in the estimated covariance matrices. Table 3 gives estimated covariance matrices for $\hat{\boldsymbol{\gamma}}$ and shows that the approximation based on V_1 is in this case preferable to that based on V_2. The Cramer-Rao lower bound considerably underestimates the variance of the REML estimator in this case. The weighted strategy of Huele et al. (2000) gives reasonably accurate standard errors based on V_1 in this case while those of SV99 are based on V_2 and are over-estimated and those of LN98 are based on the Cramer-Rao lower bound and are under-estimated.

9 Conclusions

Ideally the heteroscedastic model should be estimated using exact REML methods and Smyth (2002) describes how this can be done efficiently. However if it is desired to estimate the model using existing generalized linear model software, either in order to avoid the need for special purpose programming or in order to use the conceptual framework of double generalized linear models, then we recommend using responses $d_i/(1 - h_{ii})$ and prior weights $1 - h_{ii}$ in the dispersion model. This is a compromise between the algorithms of LN98 and SV99. Unlike the LN98 and SV99 algorithms, it returns the correct REML estimators. It does not in general return correct standard errors for likelihood values for $\boldsymbol{\gamma}$ without further calculation. However the standard error is correct in the simplest case when the variance is actually constant and the discussion of Section 7 suggests that the standard errors may often be approximately correct in other cases as well. Neither of the LN98 or SV99 algorithms can be recommended when the true REML estimators are just as easily obtained. The LN98 method is somewhat inefficient when there are highly leveraged observations in the mean model and lacks the robustness properties studied by Verbyla (1993). The SV99 appears to preserve the robustness properties but is even less efficient than LN98 for a wide range of models. All algorithms are nonlinear estimation procedures and should ideally include step-length modifications to ensure algorithmic convergence.

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. & 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.

Cox, D. R. & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *J. R. Statist. Soc.* B **49**, 1–39.

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.

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.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. 5th Berkeley Symp. Math. Statist. and Probab.*, volume 1, pages 221–233. University of California Press, Los Angeles, CA.

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.

Liang, K.-Y. & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

McCullagh, P. & Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *J. R. Statist. Soc.* B **52**, 325–344.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models.* Wiley, New York.

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. (1992). Likelihood, quasi-likelihood and pseudo-likelihood: some comparisons. *J. R. Statist. Soc.* B **54**, 273–284.

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.

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. (1996). Partitioned algorithms for maximum likelihood and other nonlinear estimation. *Statistics and Computing* **6**, 201–216.

Smyth, G. K. (2002). An efficient algorithm for REML in heteroscedastic regression. *Journal of Statistical and Graphical Computing* **11**, 1–12.

Smyth, G. K. & Verbyla, A. P. (1996). A conditional likelihood approach to REML in generalized linear models. *J. R. Statist. Soc.* B **58**, 565–572.

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.

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. *J. Roy. Statist. Soc. B* **55**, 493–508.