Smyth, G. K., and Verbyla, A. P. (1999). Double generalized linear models: approximate REML and diagnostics. In *Proceedings of the 14th International Workshop on Statistical Modelling*, Graz, Austria, July 19-23, 1999, H. Friedl, A. Berghold, G. Kauermann (eds.), Technical University, Graz, Austria, pages 66-80.

Double Generalized Linear Models: Approximate REML and Diagnostics

Gordon K. Smyth[1] and Arūnas P. Verbyla[2]

[1] Department of Statistics and Demography, University of Southern Denmark, 5230 Odense M, Denmark

[2] BiometricsSA, The University of Adelaide/South Australian Research and Development Institute, PMB 1, Glen Osmond, SA 5064, Australia.

Abstract: This paper considers double generalized linear models, which allow the mean and dispersion to be modelled simultaneously in a generalized linear model context. Estimation of the dispersion parameters is based on a χ^2_1 approximation to the unit deviances, and the accuracy of the saddle-point approximation which underlies this is discussed. Approximate REML methods are developed for estimation of the dispersion, and these are related to the likelihood adjustment methods of McCullagh and Tibshirani (1990) and Cox and Reid (1987). The approximate REML methods can be implemented with very little added complication in a generalized linear model setting by adjusting the working vector and working weights. S-Plus functions for double generalized linear models are described. Through two data examples it is shown that the approximate REML methods are more robust than maximum likelihood, in the sense of being less sensitive to perturbations in the mean model.

Keywords: dispersion modelling; REML; generalized linear models; slippage models; adjusted profile likelihood

1 Introduction

Generalized linear models allow us to model responses which are not normally distributed, using methods closely analogous to linear methods for normal data (McCullagh and Nelder, 1989). They are more general than normal linear methods in that a mean-variance relationship appropriate for the data can be accommodated and in that an appropriate scale can be chosen for modelling the mean on which the action of the covariates is approximately linear. On the other hand, once the mean-variance relationship is specified, the variance is assumed known up to a constant of proportionality, the dispersion parameter. While generalized linear models continue to be extremely useful, the complexities often encountered in observed data and the possibilities opened by modern computing power ensure that there is a strong need now for even more flexible models. Modern requirements are for models which include random effects, non-parametric trends and non-homogenous dispersion. A comprehensive attack on many real problems in biomedical or environmental research would involve an integration of these and other components. In this paper we concentrate on non-homogeneous dispersion and the modelling of dispersion in terms in covariates.

It is well known that efficient estimation of mean parameters in regression depends on correct modelling of the dispersion. The loss of efficiency in using constant dispersion models when the dispersion is varying may be substantial. Modelling of the dispersion is also necessary to obtain correct standard errors and confidence intervals, as well as for many other applications such as prediction, estimation of detection limits or immunoassay (Carroll, 1987; Carroll and Rupert, 1988). In many studies, modelling the dispersion will be of direct interest in its own right, to identify the sources of variability in the observations.

Many authors have considered dispersion modelling for normal data, for example Aitkin (1987), Carroll (1987), Davidian and Carroll (1987), Carroll and Rupert, (1988). Smyth (1989) showed that similar methods could be used for a certain class of non-normal generalized linear models. In this paper we extend Smyth's (1989) methods to arbitrary generalized linear models by using the saddle-point approximation to the distribution of the responses.

Before dispersion modelling can take place, it is necessary to estimate the mean of the data accurately. For this reason, dispersion modelling takes place in the presence of a (possibly large) number of nuisance parameters. It is well known that maximum likelihood estimators for variance parameters in regression models are generally biased. For normal linear models it is common to use residual or restricted maximum likelihood (REML) instead of maximum likelihood to estimate parameters affecting the variances. REML maximizes the likelihood, not of the original observations, but of a set of zero mean contrasts. This has the effect of adjusting for available degrees of freedom, and produces estimators which are at least approximately unbiased.

The generalization of REML to non-normal models is not obvious, as zero mean contrasts do not generally exist. Several general methods of adjusting likelihood methods for nuisance parameters have been proposed, including Cox and Reid (1987), McCullagh and Tibshirani (1990) and Smyth and Verbyla (1996), which reduce to REML for normal linear models. In this paper we use the approach of McCullagh and Tibshirani (1990) to adjust the score vector and information matrices for leverage effects. We find that this requires minimal modification to the standard computations in a generalized linear model context. We note that the adjustments agree with Cox and Reid (1987) to second order, but not with the saddle-point conditional likelihood given by Smyth and Verbyla (1996).

Verbyla (1993) shows that REML estimators in normal linear regression enjoy a hitherto unappreciated robustness property, of being less sensitive

than the maximum likelihood estimators to perturbations in the model. This property supports the notion that REML can be considered more reliable than maximum likelihood in small samples. We show, through two data examples, that our adjusted likelihood methods also enjoy this property in this more general context.
Section 2 of this paper introduces double generalized linear models, in which the mean and the dispersion are modelled simultaneously. The saddle-point approximation and its accuracy is discusses in Section 3. Section 4 discusses generalizations of REML to non-normal models. The application to double generalized linear models is set out in Section 5, and two data examples are worked through in Section 6. S-Plus functions to fit double generalized linear models are also described. The paper finishes with a summary and pointers to software availability.

2 Double Generalized Linear Models

Suppose we observe independent responses y_i, $i = 1, \ldots, n$, together with covariate vectors $\mathbf{x}_i$ and $\mathbf{z}_i$, and possibly unequal weights w_i. Generalized linear models assume that the density of y_i can be written in the form

$$f(y; \mu_i, \phi/w_i) = a(y, \phi/w_i) \exp[\frac{w_i}{\phi}\{y\theta_i - \kappa(\theta_i)\}]$$

for suitable functions κ and a (McCullagh and Nelder, 1989). Here $\mu_i = E(y_i) = \dot{\kappa}(\theta_i)$, and $\operatorname{var} y_i = (\phi/w_i)V(\mu_i)$, where $V(\mu_i) = \ddot{\kappa}(\theta_i)$ is a known function. The function V is called the variance function, and captures the mean-variance relationship for the data. The dispersion parameter, ϕ, can be interpreted as the variability in y_i once dependence of the variance on the mean and weights has been taken into account.
Following Jørgensen (1987), we say that y_i follows an *exponential dispersion model* with mean μ_i and dispersion ϕ/w_i, and write $y_i \sim \mathrm{ED}(\mu_i, \phi/w_i)$. At first sight, its seems somewhat restrictive to assume such a specific distributional form for y. However Jørgensen (1987) showed that κ can be any moment generating function, i.e., any distribution with a well defined moment generating function belongs to an exponential dispersion model.
Double generalized linear models provide a framework for modelling the dispersion in generalized linear models as well as the mean (Smyth, 1989). We assume that $y_i \sim \mathrm{ED}(\mu_i, \phi_i/w_i)$. Generalized linear models traditionally assume that the means μ_i can be modelled via link-linear relationship $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ where g is a known link function and $\boldsymbol{\beta}$ is a vector of unknown regression coefficients. Double generalized linear models assume a second link-linear predictor for the dispersion $h(\phi_i) = \mathbf{z}_i^T\boldsymbol{\lambda}$ where h is another known link function, and $\mathbf{z}_i$ is a vector of covariates affecting the dispersion. In principle, μ_i and ϕ_i could be quite general functions of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, and could include non-parametric trend terms. For simplicity however, we concentrate on the traditional link-linear relationships in this paper.

3 Estimation

3.1 The Dispersion Submodel

For our purposes it is more informative to re-write the density of y_i in the form

$$f(y;\mu,\phi) = b(y,\phi)\exp\{-\frac{1}{2\phi}d(y,\mu)\} \qquad (1)$$

where d is a distance measure between y and μ. For most distributions of interest, d can be obtained as $d(y,\mu) = 2w_i\{t(y,y) - t(y,\mu)\}$ where $t(y,\mu) = y\theta - \kappa(\theta)$. For normal data, d is the squared residual $w_i(y-\mu_i)^2$ and ϕ is the variance. The family of densities defined by (1) for different d has been intensively studied by Jørgensen (1997), and is in a sense the most general distributional form for y for which μ can be interpreted as a location parameter and ϕ as a dispersion parameter. The saddle-point approximation states that $b(y,\phi) \approx \{2\pi\phi V(y)\}^{-1/2}$ as $\phi \to 0$, the relative error being $O(\phi)$ (Jørgensen, 1997, page 103). This is appreciably more accurate than the normal approximation to the density $f(y;\mu,\phi)$, which has additive error of $O(\phi^{1/2})$ (Barndorff-Nielsen and Cox, 1989).

Write $d_i = d(y_i,\mu_i)$. Direct computation of the moment generating function by integrating $\exp d(y_i,\mu_i)$ times the saddle-point density shows that $d_i \sim \phi_i\chi^2_1$ approximately as $\phi_i \to 0$, the convergence being $O(\phi_i)$. Since the χ^2_1 distribution is a special case of the gamma distribution, this suggests an iterative scheme for estimating $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ simultaneously. Given any working value for $\boldsymbol{\lambda}$, we can estimate $\boldsymbol{\beta}$ using an ordinary generalized model for the y_i with weights w_i/ϕ_i. Given any working value for $\boldsymbol{\beta}$, we can estimate $\boldsymbol{\lambda}$ using a gamma generalized linear model for the d_i. We call the gamma generalized linear model, used to estimate $\boldsymbol{\lambda}$ for fixed $\boldsymbol{\beta}$, the *dispersion submodel.* The dispersion submodel has its own dispersion, parameter, which is 2. This estimation scheme, which alternates between estimating $\boldsymbol{\beta}$ for fixed $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}$ for fixed $\boldsymbol{\beta}$, works particularly well because $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are orthogonal parameters.

3.2 Accuracy of the Saddle-Point Approximation

The saddle-point approximation which underlies the dispersion submodel is fundamental for generalized linear model theory. Apart from supporting the estimation scheme above, it is this theorem which asserts that the deviance residuals should be approximately normal, and that the residual deviance should follow approximately a chisquare distribution on the residual degrees of freedom (Jørgensen, 1997). It is therefore of considerable interest to obtain reliable guidelines regarding the accuracy of the approximation.

The saddle-point approximation is exact when the y_i are normal or inverse-Gaussian. In these cases the unit deviances d_i are exactly $\phi_i\chi^2_1$. In other cases we use the following rule of thumb to judge the accuracy of the approximation. Let $\tau_i = \phi_i V(y_i)/(y_i - \text{boundary})^2$ where "boundary" represents the boundary of the support of y. For a gamma or Poisson distribution, the only boundary is at zero. For a binomial distribution with n trials, there are boundaries at zero and n. In the definition of τ_i, we take the closest boundary to y_i. We will take the saddle-point approximation to be satisfactory when $\tau_i \leq 1/3$ for all i. When this condition in satisfied we have good reason to treat the deviance residuals as normal and to use the gamma dispersion submodel outline in the previous subsection.

This rule of thumb has both heuristic and theoretical justifications. We describe the heuristic first. It is well known that unimodal distributions are often approximately normal when the mean is more than two or three standard deviations from the boundary of the distribution. This rule works well for the binomial and Poisson distributions for example. Knowing that the accuracy of the saddle-point approximation depends on y and ϕ but not on μ, we consider a distribution with mean equal to the observed response y. Then $\sqrt{\tau_i}$ measures the number of standard deviations separating the mean from the boundary of the distribution. Since $\sqrt{\tau_i} < 1/3$ would usually be sufficient for normality, and since the saddle-point error is of the order of the normal approximation error squared, we take $\tau_i < 1/3$ as the cutoff.

The theoretical justification applies to generalized linear models with power variance functions, $V(y) = \mu^p$ for some p. For $p \geq 1$ the only boundary is at zero, and $\tau_i = \phi_i y_i^{p-2}$ is the squared coefficient of variation. Following Jørgensen (1997), we call the distributions with power variance functions *Tweedie models*, in honor of Tweedie (1984). This family includes many distributions of interest, including the normal, Poisson, gamma and inverse-Gaussian distributions. For Tweedie models with $p \geq 1$, it can be shown that the relative error of the saddle-point approximation to $f(y; \mu, \phi)$ is in fact a non-decreasing function of τ. For the gamma distribution for example, the saddle-point approximation consists of replacing $\Gamma(1/\phi_i)$ with its Stirling's formula approximation. In this case $\tau_i = \phi_i < 1/3$ ensures that the saddle-point relative error is less than 2.8%. For the Poisson distribution, the saddle-point approximation replaces $y_i!$ with Stirling's formula. In this case $\tau_i = 1/y_i < 1/3$ ensures that the relative error is again less than 2.8%.

The binomial distribution is the most important example which does not belong to the Tweedie family. In this case $\tau_i < 1/3$ essentially requires that $y \geq 3$ and $n - y \geq 3$, which insures that the saddle-point approximation is at worst accurate to 4%.

When the y_i are gamma, the saddle-point approximation is not exact but can be modified to obtain exact results, as was shown by Smyth (1989). In that case the d_i follow exactly a *digamma distribution*, and the mean and variance function of the dispersion submodel can be modified to obtain exact maximum likelihood results. We neglect this refinement in this paper, as it is important only for gamma responses with moderate to large ϕ_i. Computer programs referred to in this paper however do compute the refinement when it is available.

4 Likelihood Adjustments

For a general weighted normal linear model,

$$\mathbf{y} \sim N(X\boldsymbol{\beta}, \Sigma(\boldsymbol{\lambda})) \tag{2}$$

with Σ depending on a vector $\boldsymbol{\lambda}$ of unknown parameters, the REML method is to estimate $\boldsymbol{\lambda}$ from the distribution of the fitted residuals. Let Q be any $n \times (n-k)$ matrix of rank $n-k$ orthogonal to X. Then Q spans the residual space of the linear model. REML estimation maximizes the likelihood of $Q^T\mathbf{y}$ instead of that of $\mathbf{y}$, i.e.,

$$Q^T\mathbf{y} \sim N(0, Q^T\Sigma Q)$$

This leads to an approximately unbiased estimator for $\boldsymbol{\lambda}$, and one which may be consistent even if k increases at the same rate as n. Estimation of $\boldsymbol{\lambda}$ is not affected by the specific choice of Q.

It is not obvious how the idea of REML can be extended to non-normal or non-linear models, because in general mean zero contrasts, such as those which make up the columns of Q, do not exist. There are however a number of general strategies designed to deal with nuisance parameters which agree with REML for normal linear models. In the general setting, we consider a log-likelihood function $\ell(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\lambda}) = \log L(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\lambda})$. We wish to estimate $\boldsymbol{\lambda}$ in the presence of the vector $\boldsymbol{\beta}$ of nuisance parameters.

Firstly, if we could specify completely priors for all the parameters, then inference for $\boldsymbol{\lambda}$ would proceed through the marginal posterior distribution for $\boldsymbol{\lambda}$. The posterior distribution for the parameters is

$$p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}) = \frac{L(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\lambda})p(\boldsymbol{\beta}, \boldsymbol{\lambda})}{p(\mathbf{y})}$$

where $p(\boldsymbol{\beta}, \boldsymbol{\lambda})$ is the joint prior of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, and $p(\mathbf{y})$ is the marginal distribution of $\mathbf{y}$. The marginal posterior distribution for $\boldsymbol{\lambda}$ alone is obtained by integrating out the nuisance parameters,

$$p(\boldsymbol{\lambda}|\mathbf{y}) = \int p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y})d\boldsymbol{\beta}$$

In the normal linear model (2), the posterior density $p(\boldsymbol{\lambda}|\mathbf{y})$ is proportional to the REML likelihood if the prior for $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is flat in the neighborhood of interest (Harville, 1974). This shows that REML can be viewed as the Bayesian principle of marginal inference.

In many cases, for example when priors are not available, we want to base inference entirely on the likelihood function. Let $\hat{\boldsymbol{\beta}}_\lambda$ be the maximum likelihood estimator of $\boldsymbol{\beta}$ for a given fixed value of $\boldsymbol{\lambda}$. The profile log-likelihood for $\boldsymbol{\lambda}$ is $\ell(\mathbf{y}; \hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\lambda})$. A number of methods have been proposed for modifying the profile log-likelihood to reduce its dependence on $\boldsymbol{\beta}$. These involve conditioning on a suitable statistic, such as $\hat{\boldsymbol{\beta}}_\lambda$, which is asymptotically sufficient for $\boldsymbol{\beta}$. Fundamental work on modified profile likelihoods is by Barndorff-Nielsen (1983). See also Barndorff-Nielsen and Cox (1994). Cox and Reid (1987) proposed a simplification of Barndorff-Nielsen's modified profile likelihood, which is applicable when $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ are orthogonal. Cox and Reid's approximate conditional log-likelihood is

$$\ell(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\lambda}) - \frac{1}{2} \log |\mathcal{J}_\beta(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\lambda})|$$

where $\mathcal{J}_\beta$ is the observed information matrix for $\boldsymbol{\beta}$. This reduces to the REML log-likelihood for normal linear models.

A third strategy is to directly correct the profile score function for non-zero expectation (McCullagh and Tibshirani, 1990). Let

$$U(\boldsymbol{\beta}, \boldsymbol{\lambda}) = E \frac{\partial \ell}{\partial \boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\lambda})$$

An approximately unbiased estimator of $\boldsymbol{\lambda}$ can be obtained from the estimating equation

$$\frac{\partial \ell}{\partial \boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\lambda}) - U(\hat{\boldsymbol{\beta}}_\lambda, \boldsymbol{\lambda}) = 0$$

Again, this approach leads to REML for the normal linear model.

5 Approximate REML for Double GLMs

Standard generalized linear model theory tells us that, for fixed $\boldsymbol{\lambda}$, maximum likelihood estimators of $\boldsymbol{\beta}$ can be obtained by solving the weighted least squares equation

$$X^T W_m X \boldsymbol{\beta} = X^T W_m \mathbf{z}_m \tag{3}$$

repeatedly, where W_m is a diagonal matrix of working weights

$$W_m = \operatorname{diag}\left(\frac{w_i}{\phi_i \dot{g}(\mu_i)^2 V(\mu_i)}\right)$$

and $\mathbf{z}_m$ is the vector of working responses $\dot{g}(\mu_i)(y_i - \mu_i) + g(\mu_i)$. At each iteration, the weight matrix W_m and working vector $\mathbf{z}_m$ are updated, and the equation is solved again for $\boldsymbol{\beta}$. This is known as iteratively reweighted least squares.

A similar weighted least squares equation exists for maximum likelihood estimator of $\boldsymbol{\lambda}$ given $\boldsymbol{\beta}$, namely

$$Z^T W_d Z \boldsymbol{\lambda} = Z^T W_d \mathbf{z}_d \tag{4}$$

where

$$W_d = \operatorname{diag}\left(\frac{1}{\dot{h}(\phi_i)^2 V_d(\phi_i)}\right)$$

and $z_{di} = \dot{h}(\phi_i)(d_i - \phi_i) + h(\phi_i)$. Full maximum likelihood estimation for all parameters can be obtain by alternating between of the iteration for $\boldsymbol{\beta}$ and the iteration for $\boldsymbol{\lambda}$ until overall convergence is obtained (Smyth, 1989 and 1996).

The least squares equation (3) for $\boldsymbol{\beta}$ has "Hat matrix"

$$H = W_m^{1/2} X (X^T W_m X)^{-1} X^T W_m^{1/2}.$$

We will write h_i for the diagonal elements of H, often known as leverages. By expanding $\hat{\boldsymbol{\beta}}$ about $\boldsymbol{\beta}$ in a Taylor series expansion, it can be shown that

$$E\{d(y_i, \hat{\mu}_i)\} = \phi_i(1 - h_i) + O(n^{-2}) \tag{5}$$

where $\hat{\mu}_i$ is μ_i evaluated at $\hat{\boldsymbol{\beta}}_\lambda$. Here we are assuming the Fisher information increases at the same rate as the sample size n, i.e., the minimum eigenvalue of $X^T W_m X$ is $O(n)$ as n increases. The result is even more accurate in the case that the generalized linear model uses a canonical link, for example reciprocal for the gamma distribution. In that case, the error in (5) is $O(n^{-3})$. In the special case of linear regression, (5) agrees with the well known result that $E\{(y_i - \hat{\mu}_i)^2\} = \sigma^2(1 - h_i)$ where σ^2 is the variance of the y_i. This suggests that we modify the working vector in the dispersion submodel from that given in the previous paragraph to to $z_{di}^* = \dot{h}(\phi_i)\{d_i - (1 - h_i)\phi_i\} + h(\phi_i)$. An approximately unbiased estimator of $\boldsymbol{\lambda}$ can be obtained by solving (4) with $\mathbf{z}_d^*$ in place of $\mathbf{z}$.

It is interesting to note that differentiating the Cox and Reid (1987) approximate conditional likelihood can also be shown to lead to (5). However the saddle-point approximate conditional likelihood given by Smyth and Verbyla (1996) gives a different but related expression, which collapses to (5) in the normal case.

Considerable further computation with Taylor series expansions also leads to an expression for the variance of $d(y_i, \hat{\mu}_i)$, from which an expression for the variance of the adjusted $\hat{\boldsymbol{\lambda}}$ can be obtained as in McCullagh and Tibshirani (1990). This leads to $\mathrm{var}(\hat{\boldsymbol{\lambda}}) \approx \mathcal{I}_{\lambda\lambda}^{-1}$ with $\mathcal{I}_{\lambda\lambda} = \frac{1}{2} Z^T W_d^* Z$ and

$$W_d^* = W_d - 2\mathrm{diag}\left(\frac{h_i}{\phi_i^2 \dot{h}(\phi_i)^2}\right) + H^2$$

Here H^2 represents the matrix (h_{ij}^2) where h_{ij} are the elements of the hat matrix H. (Note also that $h(\phi)$ represents the dispersion link function, which is unrelated to H and the h_i.)

This gives a very straightforward scheme for converting maximum likelihood estimation for $\boldsymbol{\lambda}$ to approximate REML. In the iteratively reweighted least squares update (4) for $\boldsymbol{\lambda}$, we simply change $\mathbf{z}_d$ to $\mathbf{z}_d^* = (z_{d1}^*, \ldots, z_{dn}^*)^T$ and the weight matrix W_d to W_d^*. This will ensure not only that the estimator $\hat{\boldsymbol{\lambda}}$ is approximately unbiased, but also that the dispersion submodel will give correct adjusted standard errors. In practice the matrix H^2 is expensive to calculate, so we approximate it with $\mathrm{diag}(h_i^2)$. This gives

$$W_d^* \approx W_d - 2\mathrm{diag}\left(\frac{h_i}{\phi_i^2 \dot{h}(\phi_i)^2} + h_i^2\right)$$

In many practical examples the dispersion link h is logarithmic. In that case, the expression for W_d simplifies considerably. Using $\dot{h}(\phi) = 1/\phi_i$ and $V_d(\phi_i) = \phi_i^2$ we have $W_d^* \approx \mathrm{diag}(1 - h_i^2)$.

6 Diagnostics and Examples

It is well known that REML estimation leads to estimators for the variance parameters which are more nearly unbiased than does maximum likelihood estimation. It is less well known that REML estimators are also more robust, in the sense of being less sensitive to perturbations of the data (Verbyla, 1993). This arises because of the allowance for effective degrees of freedom in the mean model. The REML estimators are less likely to follow an aberrant fitted value with a very high leverage value. In this section we show, through two data examples, that our approximate REML method also shares this property. While we demonstrate the principle for these two examples only, the robustness of the approximate REML likelihood to changes in the mean model can be expected to hold generally, because of the relationship of approximate REML with marginal likelihood. The likelihood must be, by definition, lower at the REML estimators than it is at the ML estimators. At the same time, the likelihood as the mean varies from the REML estimators must be greater than that as the mean varies from the ML estimators, since the likelihood integrated over all the mean values must be greater at the REML estimators. It follows that the maximum occurs at a relatively sharp peak of the likelihood, while the REML estimator is associated with a flatter plateau of high likelihood values.
The sensitivity of the REML estimators is investigated in this paper using the *mean slippage* or *shift outlier* model. The slippage model for an outlier at observation j is

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} + \zeta \mathbf{e}_j, \quad i = 1, \ldots, n$$

This perturbation of the mean model has the effect of introducing a unit leverage for the jth case.

6.1 Poison Experiment

Box and Cox (1964) describe an experiment involving 3 poisons and 4 treatments or antidotes. The experiment was conducted as a 3×4 factorial

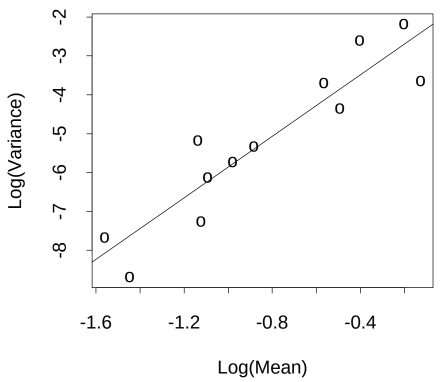

FIGURE 1. The sample mean-variance relationship for the poison experiment.

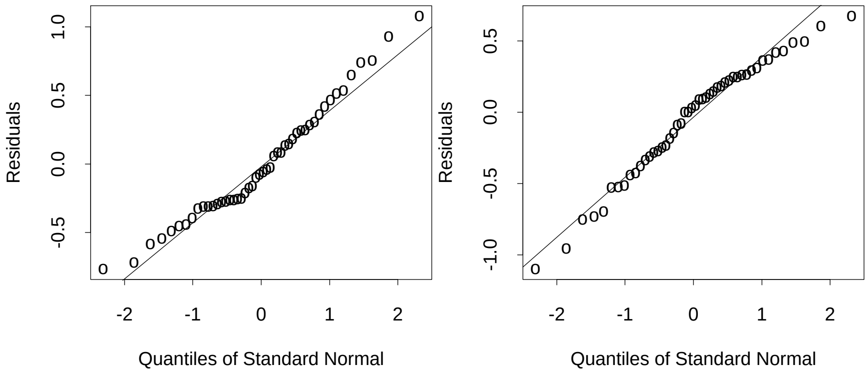

FIGURE 2. Normal probability plots of residuals for the Poison experiment. Left plot assumes reciprocal normal survival times and the right plot assumes a quartic variance function.

experiment with 4 replicates, and the response is the survival times of the animals.

A plot of the log-sample variance against log-sample mean for each poison-treatment combination gives a nearly linear trend with slope nearly equal to 4 (Figure 1). This strongly suggests $V(\mu) = \mu^4$. The approximate variance stabilizing transformation for this power variance function is the reciprocal transformation, and Box and Cox (1964) for this reason treated the reciprocal times as normally distributed. An alternative approach is to directly analyze the survival times using a generalized linear model with variance function $V(\mu) = \mu^4$.

Figure 2 shows the normal probability plot of the residuals from fitting a two-way interaction model with reciprocal normal times, and the normal probability plot of the residuals from a Tweedie generalized linear model for the survival times with power variance function $V(\mu) = \mu^4$. In this case the maximum value of $\tau = \hat{\phi}\hat{\mu}_i^{4-2}$ is 0.15, so we can be confident that the deviance residuals from the generalized linear model should be approximately normal. The reciprocal normal probability plot shows slight skewness to the right. The power variance function probability plot shows slight left skewness. Overall we feel that the probability plot for the power variance function is at least as satisfactory as that for the reciprocal normal model, and we proceed to analyze the data using this approach. This has the advantage of directly analyzing the observed responses on their own scale.

To fit a generalized linear model with power variance function, we use the S-Plus family function `tweedie`, written by one of the authors, which is available from the URL listed at the end of the paper. The command for fitting the quartic Tweedie generalized linear model is

```
glm(Time ~ Poison*Treatment,
family=tweedie(var.power=4,link.power=0))
```

Here `var.power = p` specifies the mean variance relationship $V(\mu) = \mu^p$ and `link.power = q` specifies a power link function, $\mu_i^q = \mathbf{x}_i^T\boldsymbol{\beta}$, with $q = 0$ indicating the logarithmic link.

A double generalized linear model is fitted using the S-Plus function `dglm`, for example

```
out <- dglm(Time~Poison*Treatment,~Poison,
family=tweedie(var.power=4,link.power=0),method="reml")
```

Here the first argument specifies a model formula for the mean submodel, and the second argument does the same for the dispersion submodel. Maximum likelihood estimation or approximate REML may be chosen through the `method` argument. The function `dglm` produces an object of class "dglm". There are special `summary` and `anova` methods written for the class. For example `summary(out)` will print estimated regression coefficients, standard errors and the overall likelihood for the fitted model, while `anova(out)` will print a table of likelihood ratio tests for the mean and dispersion submodels. Full programs and help file are available from the WWW site listed at the end of this paper.

In S terminology, the object class dglm is constructed so that it inherits from the classes "lm" and "glm". This means that any S-Plus function designed for linear models (lm objects) or generalized linear models (glm objects) can be applied to a dglm object with sensible results. Generic functions with methods for glm objects, such as `residuals(out)`, and functions with method for lm objects, such as `drop1(out)`, will produce results for the mean submodel. To treat the dispersion submodel as an ordinary generalized linear model, use `residuals(out$disp)`, `drop1(out$disp)` and so on.

For this data we find no mean model interaction between Poison and Treatment. In the dispersion model, we do find a main effect for Poison. The

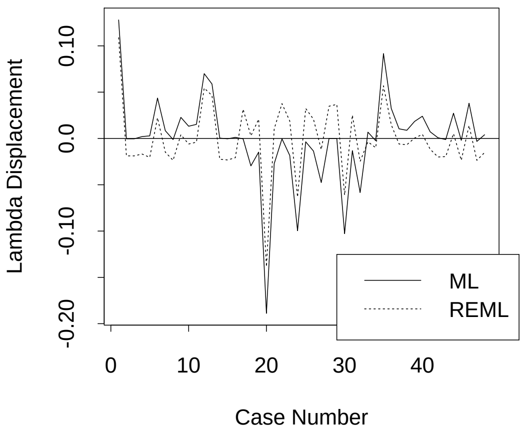

FIGURE 3. Displacements of the dispersion coefficient under mean slippage models for the Poison experiment.

estimated dispersion for Poison 2 is nearly three times as large as those for Poisons 1 and 3. The contrast for Poison 2 versus the average of the other two log-dispersions has a log-likelihood ratio test statistic of 5.96, which has P-value of 0.015 as a chi-square random variable on one degree of freedom. This is similar to the result found by Aitkin (1987), treating the survival times as reciprocal normal.

We now examine displacement of the dispersion parameter in response to mean slippage models. We fit the mean model `Time` $\sim$ `Poison+Treatment` and the dispersion model $\sim$ `Poison2`, where `Poison2` is the factor distinguishing the second poison from poisons 1 and 3. We compute the change in the coefficient for `Poison2` under slippage models, i.e., as the indicator vector $\mathbf{e}_j$ for case j is added to the mean model. The results are given in Figure 3. We see that the larger displacements are all reduced for REML estimation compared with ML estimation. This shows that the REML estimation of the dispersion model is less sensitive to perturbations of the mean model than is maximum likelihood estimation.

6.2 Blood CPK in Skiers

The data gives the blood CPK concentrations of skiers 12 hours into a cross country ski marathon (Zuliani et al, 1983). Leakage of the enzyme CPK into the blood is a common symptom of muscle stress. Figure 4 relates log-CPK concentrations to the age of each skier. This shows an approximately linear decreasing trend, and also decreasing variability, as age increases. Attempts to stabilize the variance by using a stronger transformation than logarithmic are unattractive because the relatively low observation for one skier of age 33 tends to become an outlier. Instead, we model the blood

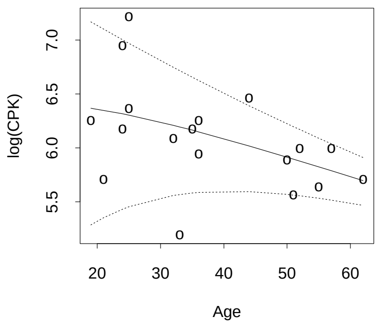

FIGURE 4. Log-blood CPK concentration versus age for skiers in a cross-country marathon. The solid line is the estimated median and the dotted lines are 0.1 and 0.9 quantiles of the response distribution under a gamma mean-dispersion model.

CPK concentrations directly as gamma with a log-link. The following is summary output from the function `dglm`.

```
Call: dglm(formula = CPK ~ Age, dformula =  ~ Age,
    family = tweedie(var.power = 2, link.power = 0), method = "reml")

Mean Coefficients:
                  Value  Std. Error    t value
(Intercept)  6.88658455 0.310400523  22.186124
        Age -0.01902809 0.006193807  -3.072115
(Dispersion Parameters for Tweedie family estimated as below )

    Scaled Null Deviance: 26.42384 on 17 degrees of freedom
Scaled Residual Deviance: 16.20617 on 16 degrees of freedom

Dispersion Coefficients:
                  Value Std. Error    t value
(Intercept)  0.41601037 1.09056097  0.3814646
        Age -0.06333238 0.02855352 -2.2180238
(Dispersion Parameter for Digamma family taken to be 2 )

    Scaled Null Deviance: 66.08189 on 17 degrees of freedom
Scaled Residual Deviance: 54.391 on 16 degrees of freedom

Minus Twice the Log-Likelihood: 248.676
Number of Alternating Iterations: 11
```

The output shows the REML t-value for Age in the dispersion model as -2.2, which has P-value 0.028 as an approximate standard normal random variable. The likelihood ratio test statistic for Age, which can be obtained

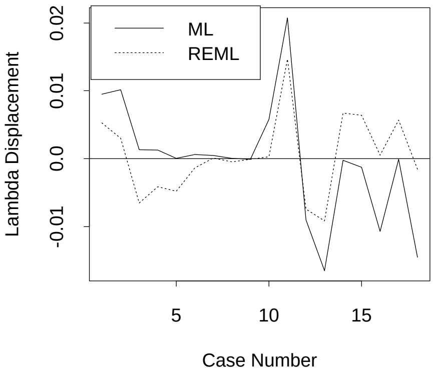

FIGURE 5. Blood CPK concentration: displacements of the coefficient for Age in the dispersion submodel under mean slippage models.

separately from `anova` output, is 6.04 with P-value 0.014. The fitted model for the dispersion is

$$\log \hat{\phi}_i = 0.416 - 0.0633\text{Age}$$

which means that the dispersion decreases from 0.46 at age 19 to 0.03 at age 62. The estimated response standard deviation $\hat{\phi}_i^{1/2}\hat{\mu}_i$ decreases by 89% from age 19 to age 62. Figure 4 includes the fitted mean and dispersion models.

Results for the displacement of the coefficient for Age in the dispersion submodel under mean slippage models are given in Figure 5. Again we see that the larger displacements are reduced under REML compared with ML. Software and documentation for the S-Plus functions used are available from the URL `http://www.maths.uq.edu.au/~gks/s/`.

References

Aitkin, M. (1987). Modelling variance heterogeneity in normal regression using GLIM. *Appl. Statist.*, **36**, 332–9.

Barndorff-Nielsen, O. E. (1983). On a formula for the distribution of a maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Barndorff-Nielsen, O. E. and Cox, D. R. (1989). *Asymptotic techniques for use in statistics.* Chapman and Hall.

Barndorff-Nielsen, O. E., and Cox, D. R. (1994). *Inference and Asymptotics.* London: Chapman and Hall.

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformation (with discussion). *J. R. Statistic. Soc. B*, **26**, 211–52.

Carroll, R. J. (1987). The effect of variance function estimating on prediction and calibration: an example. In *Statistical Decision Theory and Related Topics IV* (eds J. O. Berger and S. S. Gupta), vol. II. Heidelberg: Springer.

Carroll, R. J., and Rupert, D. (1988). *Transforming and Weighting in Regression*. London: Chapman and Hall.

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

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

Jørgensen, B. (1987). Exponential dispersion models (with discussion). *J. R. Statist. Soc. B*, **49**, 127–162.

Jørgensen, B. (1997). *The theory of dispersion models.* Chapman and Hall, London.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* London: Chapman and Hall.

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

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

Smyth, G. K. and Verbyla, A. P. (1996). A conditional approach to residual maximum likelihood estimation in generalized linear models. *J. Roy. Statist. Soc.* B, **58**, 565–572.

Tweedie, M. C. K. (1984). An index which distinguishes between some important exponential families. In *Statistics: Applications and New Directions. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference.* (Eds. J. K. Ghosh and J. Roy), pp. 579–604. Calcutta: Indian Statistical Institute.

Verbyla, A. P. (1993). Modelling variance heterogeneity: residual maximum likelihood and diagnostics. *J. Roy. Statist. Soc.*, B **55**, 493–508.

Zuliani, U., Mandras, A., Beltrami, G. F., Bonetti, A., Montani, G., and Novarini, A. (1983). Metabolic modifications caused by sport activity: effect in leisure-time cross-country skiers. *J. Sports Medicine and Physical Fitness*, **23**, 385–392.