Leverage Adjustments for Dispersion Modelling in Generalized Nonlinear Models

Gordon K. Smyth
Department of Mathematics
University of Queensland, Australia

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

15 June 2001

Abstract

For normal linear models, it is generally accepted that residual maximum likelihood estimation is appropriate when covariance components require estimation. This paper considers generalized linear models in which both the mean and the dispersion are allowed to depend on unknown parameters and on covariates. For these models there is no closed form equivalent to residual maximum likelihood except in very special cases. The modified profile likelihood for the dispersion parameters is differentiated to obtain an adjusted score vector and adjusted information matrix. The expectation of the fitted deviances is obtained directly to show that the adjusted score vector is unbiased at least to $O(1/n)$. Exact results are obtained from simple samples. The results reduce to residual maximum likelihood estimation in the normal linear case. An asymptotic development is given which holds as the leverages in the mean model become small.

Keywords: Dispersion; Double generalized linear models; Maximum likelihood; Modified profile likelihood; Residual maximum likelihood; Saddlepoint approximation; gamma distribution; inverse-Gaussian distribution.

1 Introduction

This paper considers simultaneous mean-dispersion modelling for generalized linear models. Letting μ_i denote the mean and ϕ_i the dispersion for the ith observation, we typically assume a link-linear model for the mean

$$f(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{1}$$

where $\mathbf{x}_i$ is a vector of covariates relevant for predicting the mean and $\boldsymbol{\beta}$ is a vector of regression coefficients, and in addition an analogous link-linear model for the dispersion

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

where $\mathbf{z}_i$ is a vector of covariates relevant for predicting the dispersion and $\boldsymbol{\gamma}$ is another vector of regression coefficients. 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 or squared residuals as responses in which $\boldsymbol{\gamma}$ is estimated with $\boldsymbol{\beta}$ fixed. We refer to such cases as double generalized linear models. Most of the results of this paper are not in fact restricted to (1) or (2) but remain true for generalized nonlinear models for which the dependence of the μ_i on $\boldsymbol{\beta}$ and of the ϕ_i on $\boldsymbol{\gamma}$ may be quite general.

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 inference using double exponential familes was introduced by Diaconis & Efron (1985) and Efron (1986). Estimation using the methods of these papers produces dispersion estimators which tend to be too small because residuals are evaluated at the fitted means rather than at the true means. Various methods of adjusting the dispersion estimators for this bias were proposed by Smyth & Verbyla (1996), Lee & Nelder (1998) and Smyth & Verbyla (1999). In earlier papers Cordeiro & McCullagh (1991) considered bias correction for the maximum likelihood estimator of the dispersion for ordinary generalized linear models and Nelder & Lee (1992) used a simple degree of freedom adjustment when using EQL with single samples. The methods of Lee and Nelder are essentially ad hoc while the method of Smyth & Verbyla (1996) is applicable only when (1) defines a single-factor design or f is the canonical link. Smyth & Verbyla (1999) made use of an early version of the results of this paper without giving details. The aim of this paper is to derive an adjusted score vector and adjusted information matrix for the dispersion parameters $\boldsymbol{\gamma}$ and to give an asymptotic development which clarifies the properties of the earlier methods. The results of this paper agree with REML in the normal linear case. The results are also exact in the case of single samples from a gamma or inverse-Gaussian distribution.

Nelder & Pregibon (1987)'s EQL and Efron (1986)'s double exponential families are both approximations of the saddlepoint type as discussed by Barndorff-Nielsen & Cox (1979). The double generalized linear model idea of estimating the dispersions by way of a regression using the unit deviances assumes that the unit deviances are conditionally sufficient for the dispersions given the means, and this is true only when the saddlepoint approximation for the generalized linear model density is exact or exact after a re-normalization.

This restricts us to the normal, gamma or inverse-Gaussian families (Blaesild & Jensen, 1985, Smyth, 1989). The saddlepoint approximation is accurate for other distributional families whenever the dispersion is small (Jørgensen, 1998), and so our results flow on to other families whenever the variability is not too large. In particular our results hold in the contexts discussed by Nelder & Pregibon (1987) and Efron (1986).

Generalized linear models were introduced by Nelder & Wedderburn (1972) and are now one of the most useful methodologies for modelling categorical and non-normal data (McCullagh & Nelder, 1989). Generalized linear models automatically allow for dependence of the variance on the mean in the estimation process. The dispersion is the variance factor which remains after dependence on the mean is removed. The dispersion itself is traditionally treated as constant but there is now increasing recognition of the need to model it as a function of covariates. Rigby & Stasinopolous (1996) for example consider smoothing spline modelling of both the mean and the dispersion while Nair & Pregibon (1986), Nair & Pregibon (1988) and Bergman & Hynén (1997) consider dispersion effects in designed experiments.

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 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 parameters is unbiased, providing consistent estimators in situations where maximum likelihood estimators are inconsistent.

Double generalized linear models are a generalization of heteroscedastic regression models for normal data, and these 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. Engel & Huele (1996) applied a similar model to Taguchi-type experiments for robust design. A systematic study of REML for the log-linear variance model was undertaken by Verbyla (1993). Amongst other things, Verbyla (1993) provided a comparison of REML and maximum likelihood via regression diagnostics. Huele & Engel (1998) discussed the implementation of the REML estimator from Verbyla (1993) using a double generalized model.

There is also an active literature on variance modelling for nonlinear and nonparametric regression models. Davidian & Carroll (1987) give an excellent survey and a comparison of methods. They consider variance estimation and REML-type adjustments for very general models, including in principle double generalized models, although the emphasis in their paper is on applications where the responses are near normal. Hall & Carroll (1989) have considered the effect of estimation of the mean on the variance when both the mean and the variance are to be estimated nonparametrically.

Strictly speaking, REML is a method which applies only to normal linear models. The original derivation of REML for the normal linear model in terms of zero-mean contrasts does not generalize to double generalized linear models because zero-mean contrasts do not exist. Generalizations of REML to nonlinear or non-normal models are instead based on general methods for adjusting profile likelihoods for nuisance parameters. The profile likelihood is formed for the dispersion parameters $\boldsymbol{\gamma}$ and adjusted for estimation of the mean parameters $\boldsymbol{\beta}$. We refer to such adjustment methods here as approximate REML because their purpose is the same as for REML in the normal linear case and because they ideally include REML as a special case. Adjusted profile likelihoods have a large literature; see for example Kalbfleisch & Sprott (1970), Barndorff-Nielsen (1986) and Davison & Stafford (1998). In this paper we develop an approximate REML method for double generalized nonlinear models based on the approximate conditional likelihood for orthogonal parameters developed by Cox & Reid (1987).

Section 2 of this paper reviews REML estimation in the normal linear case and generalizes the results of Verbyla (1993). Sections 3 and 4 describe the distributional results underlying double generalized linear and nonlinear models. Section 5 reviews maximum likelihood estimation for these families and generalizes the results of Smyth (1989). In Section 6 we obtain a modified score vector and modified information matrix by differentiating the approximate conditional likelihood. The asymptotic orders of the approximations are given assuming small leverages. In Section 7 we given an expansion for the fitted deviances and show that the modified score vector is bias-adjusted (McCullagh & Tibshirani, 1990) to the appropriate order. In Section 8 we obtain exact results for the simple case in which both the mean and the dispersion are constant. This confirms the results of the earlier sections and also allows us to refine the earlier results to deal with the possibility of large leverages.

2 Heteroscedastic Regression

It is useful to summarize first the results for normal linear regression. Suppose that $y_i \sim N(\mu_i, \phi_i)$, $i = 1, \ldots, n$, are independent and that $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$ and the ϕ_i are functions of the parameter vector $\boldsymbol{\gamma}$. This section generalizes the results of Verbyla (1993) to an arbitrary model for the variances. The REML estimator of $\boldsymbol{\gamma}$ is obtained by maximizing the marginal log-likelihood

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

where ℓ is the ordinary log-likelihood, $\hat{\boldsymbol{\beta}}_\gamma$ is the conditional maximum likelihood estimator for $\boldsymbol{\beta}$ for given fixed $\boldsymbol{\gamma}$, X is the $n \times p$ design matrix with rows $\mathbf{x}_i^T$, $\Sigma_m = \mathrm{diag}(\phi_i)$ 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

$$U_R = \dot{\boldsymbol{\phi}}^T \Sigma_d^{-1} (\mathbf{d} - \boldsymbol{\phi}^*)$$

where $\dot{\boldsymbol{\phi}}$ is the $n \times q$ gradient matrix with elements $\partial\phi_i/\partial\gamma_j$, $\Sigma_d = \mathrm{diag}(2\phi_i^2)$, $\mathbf{d}$ is the vector of squared residuals $d_i = (y_i - \mu_i)^2$, $\boldsymbol{\phi}^*$ is the vector of $(1-h_{ii})\phi_i$ 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 REML information matrix for $\boldsymbol{\gamma}$ is

$$\mathcal{I}_R = \dot{\boldsymbol{\phi}}^T \Sigma_d^{-1/2} V \Sigma_d^{-1/2} \dot{\boldsymbol{\phi}}$$

where 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 Σ_d is the covariance matrix of $\mathbf{d}$ evaluated at the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ while V is the covariance matrix of the standardized vector $\Sigma_m^{-1}\mathbf{d}$ evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_\gamma$. The REML scoring iteration for $\boldsymbol{\gamma}$ is

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

with $\mathcal{I}_R$ and U_R computed at $\boldsymbol{\gamma}_k$ and $\hat{\boldsymbol{\beta}}_\gamma$.

If the ϕ_i satisfy (2) then

$$\dot{\boldsymbol{\phi}} = \mathrm{diag}\{1/\dot{g}(\phi_i)\} Z$$

which results in some simplification for U_R and $\mathcal{I}_R$. In that case it is possible to obtain the REML estimator for $\boldsymbol{\gamma}$ by repeatedly fitting a gamma regression model with responses $d_i/(1-h_{ii})$, link function g and prior weights $1-h_{ii}$, although this process does not produce in correct standard errors obtained from $\mathcal{I}_R$.

3 Exponential Dispersion Models

Following Jørgensen (1998) we define an exponential dispersion model $\mathrm{ED}(\mu, \phi)$ with mean μ and dispersion ϕ to be a family of distributions with density or probability mass function given by

$$p(y; \mu, \phi) = a(y, \phi) \exp\left[\frac{1}{\phi}\{y\theta - \kappa_m(\theta)\}\right] \tag{4}$$

where $a()$ and $\kappa()$ are known functions. The mean and variance of the distribution are μ and $\phi v_m(\mu)$ respectively where $\mu = \dot{\kappa}_m(\theta)$ and $v_m(\mu) = \ddot{\kappa}_m(\theta)$. A series of independent observations $y_1, \ldots, y_n$ is said to follow a generalized linear model when $y_i \sim \mathrm{ED}(\mu_i, \phi/w_i)$ for known weights w_i and the μ_i satisfy the link-linear model (1). In this paper we generally assume that the y_i follow a double generalized linear model, i.e., that $y_i \sim \mathrm{ED}(\mu_i, \phi_i/w_i)$ and the μ_i and ϕ_i satisfy (1) and (2). Almost all the results however apply to more general nonlinear models in which the μ_i and the ϕ_i may depend generally on $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ respectively.

Write $t(y,\mu) = y\theta - \kappa_m(\theta)$. The density (4) can be rewritten in the form

$$p(y;\mu,\phi) = b(y,\phi)\exp\left\{-\frac{1}{2\phi}d(y,\mu)\right\} \tag{5}$$

where $d(y,\mu) = 2\{\sup_{\mu'} t(y,\mu') - t(y,\mu)\}$ is the unit deviance and $b(y,\phi) = \sup_{\mu'} p(y;\mu',\phi)$. If y belongs to the domain of μ then the unit deviance is $d(y,\mu) = 2\{t(y,y) - t(y,\mu)\}$ and $b(y,\phi) = p(y;y,\phi)$. This is always true for the exponential dispersion models we will be concerned with (Jørgensen, 1998).

4 The Dispersion Submodel

If the μ_i are fixed then it is natural to use the weighted unit deviances $d_i = w_i d(y_i,\mu_i)$ as responses to estimate the ϕ_i. We assume in this paper that the d_i are sufficient for the ϕ_i given the μ_i. This implies that $p(y;\mu,\phi)$ is an exponential family density and that

$$b(y,\phi) = \{2\pi\phi v_m(y)\}^{-1/2}\rho(\phi) \tag{6}$$

for some function $\rho(\phi)$ (Smyth, 1989; Jørgensen, 1998). There are only three exponential dispersion models which have this form exactly: the normal, gamma and inverse-Gaussian families. For the normal and inverse-Gaussian families $\rho(\phi) = 1$. For the gamma family

$$\rho(\phi) = (2\pi)^{1/2}\frac{(-\alpha)^{-\alpha-1/2}e^{\alpha}}{\Gamma(-\alpha)}$$

where $\alpha = -1/\phi$. Here $\rho(\phi)$ is the ratio of $\Gamma(-\alpha)$ to its approximation using Stirling's formula. The range of applicability of (6) can be extended by the saddle-point approximation for exponential dispersion models which implies that

$$b(y,\phi) = \{2\pi\phi v_m(y)\}^{-1/2}\{1 + O(\phi)\} \tag{7}$$

as $\phi \to 0$ so that (6) is approximately true for any exponential dispersion model if ϕ is small (Jørgensen, 1998). Note that $\rho(\phi) \to 1$ as $\phi \to 0$.

From (5) and (6) the log-density may be written

$$\log p(y;\mu,\phi/w) = -\frac{1}{2}\log\{2\pi v_m(y)/w\} + \frac{1}{2}\{\alpha d - \kappa_d(\alpha)\} \tag{8}$$

with $\kappa_d(\alpha) = -\log(-w\alpha) - 2\log\rho\{-1/(w\alpha)\}$. This itself has the form of an exponential dispersion model for given μ with response d, canonical parameter α, cumulant function $\kappa_d()$ and dispersion parameter 2.

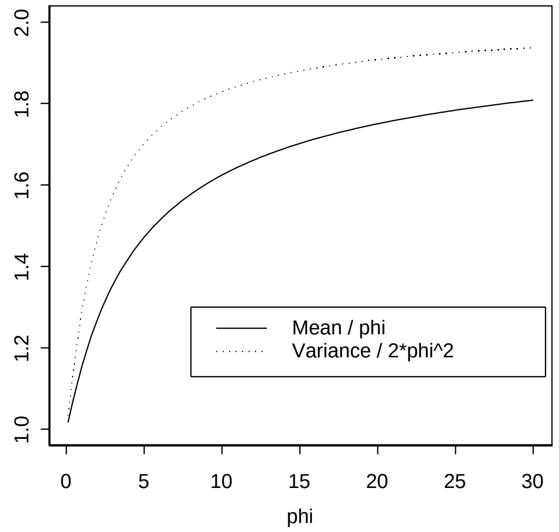

Figure 1: Mean and variance of the unit deviance $d(y,\mu)$ for the gamma distribution. The plot gives the mean and variance divided by ϕ and $2\phi^2$ respectively. The prior weight w is assumed to be unity.

The saddle-point approximation (7) implies that d_i/ϕ_i follows a χ^2_1-distribution if ϕ_i is small. This can be most easily seen from the moment generating function

$$\begin{aligned}
M(s) &= E\left\{\exp\left(s\frac{w}{\phi}d(Y,\mu)\right)\right\} \\
&= \int \exp\left\{s\frac{w}{\phi}d(y,\mu)\right\}\left\{\frac{w}{2\pi\phi v_m(y)}\right\}^{1/2}\exp\left\{-\frac{w}{2\phi}d(y,\mu)\right\}dy \\
&= (1-2s)^{-1/2}\int\left\{\frac{(1-2s)w}{2\pi\phi v_m(y)}\right\}^{1/2}\exp\left\{-\frac{(1-2s)w}{2\phi}d(y,\mu)\right\}dy \\
&= (1-2s)^{-1/2}
\end{aligned}$$

which is as for the χ^2_1-distribution. The saddlepoint approximation implies that $E(d_i)\approx\phi_i$ and $\mathrm{var}(d_i)\approx 2\phi_i^2$ for ϕ_i small. More exact results can be obtained from (8) which shows that $E(d_i)=\dot\kappa_d(\alpha_i)$ and $\mathrm{var}(d_i)=2\ddot\kappa_d(\alpha_i)$. We will write $\delta=\dot\kappa_d(\alpha)$ for the mean of d_i and $v_d(\delta)=\ddot\kappa_d(\alpha)$ so that $v_d()$ is the variance function for the dispersion submodel (8). For the normal and inverse-Gaussian distributions, d_i/ϕ_i is in fact exactly χ^2_1. For the gamma distribution

$$\delta=\dot\kappa_d(\alpha)=2w\{\log(-w\alpha)-\psi(-w\alpha)\}$$

and

$$v_d(\delta)=\ddot\kappa_d(\alpha)=2w^2\{1/(w\alpha)+\dot\psi(-w\alpha)\}$$

where $\psi()$ is the digamma function and $\dot\psi()$ is the trigamma function. Figure 1 shows that for ϕ_i large the mean and variance of the gamma disribution

d_i each become twice as large as the saddlepoint approximation would suggest. This shows that the usual χ^2_1 approximation for the deviances is not likely to be satisfactory for the gamma distribution for value of ϕ_i much greater than unity. (Such values correspond to gamma distributions which are more right skew than the exponential distribution.)

5 Maximum Likelihood

The likelihood computations for double generalized linear models are closely analogous to those for normal linear models. This section generalizes the results of Smyth (1989) to general mean and dispersion models. The score vector and information matrix for $\boldsymbol{\beta}$ are

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

and

$$\mathcal{I}_\beta = \dot{\boldsymbol{\mu}}^T \Sigma_m^{-1} \dot{\boldsymbol{\mu}}$$

where $\Sigma_m = \operatorname{var} \mathbf{y} = \operatorname{diag}\{\phi_i v_m(\mu_i)/w_i\}$ and $\dot{\boldsymbol{\mu}}$ is the gradient matrix of $\boldsymbol{\mu}$ with respect to $\boldsymbol{\beta}$. The score vector and information matrix for $\boldsymbol{\gamma}$ are

$$U_\gamma = \dot{\boldsymbol{\delta}}^T \Sigma_d^{-1} (\mathbf{d} - \boldsymbol{\delta})$$

and

$$\mathcal{I}_\gamma = \dot{\boldsymbol{\delta}}^T \Sigma_d^{-1} \dot{\boldsymbol{\delta}}$$

where $\mathbf{d}$ is the n-vector of $d_i = w_i d(y_i, \mu_i)$, $\boldsymbol{\delta} = E(\mathbf{d})$ is the n-vector of δ_i, $\Sigma_d = \operatorname{var} \mathbf{d} = \operatorname{diag}\{2 v_d(\delta_i)\}$ and $\dot{\boldsymbol{\delta}}$ is the gradient matrix of $\boldsymbol{\delta}$ with respect to $\boldsymbol{\gamma}$. The two parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal. Now $\partial\delta/\partial\phi = v_d(\delta)/\phi^2$ so we can rewrite the score vector and information matrix in terms of the ϕ_i as

$$U_\gamma = \dot{\boldsymbol{\phi}}^T \Sigma_\phi^{-1} (\mathbf{d} - \boldsymbol{\delta})$$

with $\Sigma_\phi = \operatorname{diag}(2\phi_i^2)$ and

$$\mathcal{I}_\gamma = \dot{\boldsymbol{\phi}}^T \operatorname{diag}\left\{ \frac{v_d(\delta_i)}{2\phi_i^4} \right\} \dot{\boldsymbol{\phi}}$$

where $\dot{\boldsymbol{\phi}}$ is the gradient matrix of $\boldsymbol{\phi}$ with respect to $\boldsymbol{\gamma}$.

If the μ_i and ϕ_i satisfy (1) and (2) the $\dot{\boldsymbol{\mu}} = X \operatorname{diag}\{1/\dot{f}(\mu_i)\}$ and $\dot{\boldsymbol{\phi}} = Z \operatorname{diag}\{1/\dot{g}(\mu_i)\}$. Then the method of scoring for computing the maximum likelihood estimates can be written as a pair of iteratively reweighted least squares calculations

$$\boldsymbol{\beta}_{k+1} = \left(X^T W_m^{-1} X \right)^{-1} X^T W_m^{-1} \mathbf{z}_m \tag{9}$$

with

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

and

$$(z_m)_i = \dot{f}(\mu_i)(y_i - \mu_i) + f(\mu_i)$$

and

$$\boldsymbol{\gamma}_{k+1} = \left(Z^T W_d Z\right)^{-1} Z^T W_d^{-1} \mathbf{z}_d \tag{10}$$

with

$$W_d = \operatorname{diag}\left\{\frac{v_d(\delta_i)}{2\dot{g}(\phi_i)^2 \phi_i^4}\right\}$$

$$(z_d)_i = \frac{\dot{g}(\phi_i)\phi_i^2}{v_d(\delta_i)}(d_i - \delta_i) + g(\phi_i).$$

The scoring iteration for $\boldsymbol{\beta}$ is as for a generalized linear model with weights w_i/ϕ_i. The scoring iteration of $\boldsymbol{\gamma}$ has the form of weighted regression of $\mathbf{z}_d$ on Z.

6 Modified Profile Likelihood

6.1 Modified Score Vector

The essential idea of REML estimation is to adjust the profile likelihood for the dispersion parameters in such a way that has it has the properties of a true likelihood and so that it does not depend on the mean parameters. The dispersion parameters $\boldsymbol{\gamma}$ are orthogonal to the mean parameters $\boldsymbol{\beta}$, so we use the approximate conditional profile likelihood defined by Cox & Reid (1987). Write $\ell(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma})$ for the log-likelihood function. The approximate conditional profile log-likelihood for $\boldsymbol{\gamma}$ is

$$\ell_R(\mathbf{y}; \boldsymbol{\gamma}) = \ell_C(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma})$$

where

$$\ell_C(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma}) = \ell(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma}) - \frac{1}{2}\log|-\ddot{\ell}_\beta(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma})|,$$

$\hat{\boldsymbol{\beta}}_\gamma$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ for the given value of $\boldsymbol{\gamma}$ and $-\ddot{\ell}_\beta$ is the observed information for $\boldsymbol{\beta}$. The approximate conditional profile likelihood is equal, apart from constant terms, to the REML likelihood in the normal linear case. We seek to estimate $\boldsymbol{\gamma}$ by maximizing ℓ_R.

The hat matrix for the mean submodel is

$$H = \Sigma_m^{-1/2}\dot{\boldsymbol{\mu}}(\dot{\boldsymbol{\mu}}^T\Sigma_m^{-1}\dot{\boldsymbol{\mu}})^{-1}\dot{\boldsymbol{\mu}}^T\Sigma_m^{-1/2}.$$

We assume that the elements h_{ij} of H are $O(n^{-1})$ as $n \to \infty$. This ensures that the maximum likelihood estimator of $\boldsymbol{\beta}$ is consistent and that the elements of $\mathcal{I}_\beta$ are $O(n)$. It is also ensures that the observed information is asymptotically equivalent to expected information,

$$-\ddot{\ell}_\beta = \mathcal{I}_\beta - \frac{\partial}{\partial\boldsymbol{\beta}}\left\{\dot{\boldsymbol{\mu}}^T\Sigma_m^{-1}\right\}(\mathbf{y} - \boldsymbol{\mu}) = \mathcal{I}_\beta\{I + O(n^{-1/2})\}$$

so we approximate $\mathcal{J}_\beta$ by $\mathcal{I}_\beta$ in ℓ_R.

The correction term can be formally differentiated as

$$\frac{\partial}{\partial \phi_i} \log |\mathcal{I}_\beta(\hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma})| = \frac{\partial}{\partial \mu_i} \log |\mathcal{I}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma})| \ \frac{\partial \hat{\mu}_i}{\partial \phi_i} + \frac{\partial}{\partial \phi_i} \log |\mathcal{I}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma})|$$

evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_\gamma$. Now $\partial \log |\mathcal{I}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma})|/\partial \mu_i = O(n^{-1})$, $\partial \hat{\mu}_i/\partial \phi_i = O(n^{-1/2})$ and $\partial \log |\mathcal{I}_\beta(\boldsymbol{\beta}, \boldsymbol{\gamma})|/\partial \phi_i = -h_{ii}/\phi_i$ so

$$\frac{\partial}{\partial \phi_i} \log |\mathcal{I}_\beta(\hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma})| = -\frac{h_{ii}}{\phi_i} + O(n^{-3/2}).$$

This leads to the modified score vector

$$\dot{\ell}_R = \frac{\partial \ell_C}{\partial \boldsymbol{\gamma}}(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}) = \dot{\boldsymbol{\phi}}^T \Sigma_\phi^{-1} (\mathbf{d} - \boldsymbol{\delta}^*) + O(n^{-1/2})$$

where $\boldsymbol{\delta}^*$ represents the leverage adjusted expectation vector with elements $\delta_i - h_{ii}\phi_i$.

This score vector agrees with that for the normal linear case given in Section 2. It differs from that suggested by Lee & Nelder (1998) who replace $\mathbf{d} - \boldsymbol{\delta}^*$ with the vector with elements $d_i/(1-h_{ii}) - \phi_i$. It agrees with the score vector given by Smyth & Verbyla (1999), although Smyth & Verbyla (1999) go on to introduce modified weights which change the estimator slightly. The behaviour of the Lee & Nelder (1998) and Smyth & Verbyla (1999) estimators has been examined in the normal linear case by Smyth et al. (2001).

6.2 Modified Information

To obtain the second derivative of ℓ_R we need

$$\ddot{\ell}_R(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}) = \frac{\partial^2 \ell_C}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\gamma}}(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}) \ \frac{\partial \hat{\boldsymbol{\beta}}_\gamma}{\partial \boldsymbol{\gamma}} + \frac{\partial^2 \ell_C}{\partial \boldsymbol{\gamma}^T \partial \boldsymbol{\gamma}}(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}).$$

The fitted leverages may be differentiated to get

$$\frac{\partial \hat{h}_{ii}}{\partial \phi_i} = \frac{\partial h_{ii}}{\partial \mu_i} \ \frac{\partial \hat{\mu}_i}{\partial \phi_i} + \frac{\partial h_{ii}}{\partial \phi_i} = -\frac{h_{ii}}{\phi_i} + \frac{h_{ii}^2}{\phi_i} + O(n^{-3/2})$$

and similarly

$$\frac{\partial \hat{h}_{ii}}{\partial \phi_j} = \frac{h_{ij}^2}{\phi_j} + O(n^{-5/2})$$

for $i \neq j$. This leads directly to

$$\frac{\partial^2 \ell_C}{\partial \boldsymbol{\gamma}^T \partial \boldsymbol{\gamma}}(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}) = -\dot{\boldsymbol{\phi}}^T \mathrm{diag}\left\{\frac{v_d(\delta_i)}{2\phi_i^4}\right\} \dot{\boldsymbol{\phi}} - \dot{\boldsymbol{\phi}}^T \Sigma_\phi^{-1/2} H^2 \Sigma_\phi^{-1/2} \dot{\boldsymbol{\phi}} + O(n^{-1/2}) + 0_E$$

where H^2 is the $n \times n$ matrix with elements h_{ij}^2 and 0_E represents terms with expectation zero.

The derivative of $\hat{\boldsymbol{\beta}}_\gamma$ is

$$\frac{\partial \hat{\boldsymbol{\beta}}_\gamma}{\partial \boldsymbol{\gamma}} = -\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} = -\mathcal{I}_\beta^{-1} \frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\gamma}^T} \{I + O(n^{-1/2})\},$$

so up to a factor of $1 + O(n^{1/2})$

$$\begin{aligned} & \frac{\partial^2 \ell_C}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\gamma}} (\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}) \frac{\partial \hat{\boldsymbol{\beta}}_\gamma}{\partial \boldsymbol{\gamma}} \\ = \;& -\dot{\boldsymbol{\phi}}^T \operatorname{diag}\left\{ \frac{w_i (y_i - \hat{\mu}_i)}{\phi_i^2 v(\hat{\mu}_i)} \right\} \dot{\boldsymbol{\mu}} \mathcal{I}_\beta^{-1} \dot{\boldsymbol{\mu}}^T \operatorname{diag}\left\{ \frac{w_i (y_i - \hat{\mu}_i)}{\phi_i^2 v(\hat{\mu}_i)} \right\} \dot{\boldsymbol{\phi}} \\ = \;& -\dot{\boldsymbol{\phi}}^T \operatorname{diag}\left\{ \frac{w_i^{1/2} (y_i - \hat{\mu}_i)}{\phi^{3/2} v(\hat{\mu}_i)^{1/2}} \right\} H \operatorname{diag}\left\{ \frac{w_i^{1/2} (y_i - \hat{\mu}_i)}{\phi_i^{3/2} v(\hat{\mu}_i)^{1/2}} \right\} \dot{\boldsymbol{\phi}} \end{aligned}$$

Now the standardized Pearson residuals $w_i^{1/2}(y_i - \hat{\mu}_i)/\phi_i^{1/2}$ have covariance matrix $I - H + O(n^{-3/2})$ so

$$\begin{aligned} & E\left\{ \frac{\partial^2 \ell_C}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\gamma}} (\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma}) \frac{\partial \hat{\boldsymbol{\beta}}_\gamma}{\partial \boldsymbol{\gamma}} \right\} \\ = \;& -\dot{\boldsymbol{\phi}}^T \operatorname{diag}\left\{ \frac{1}{\phi_i} \right\} H \circ (I - H) \operatorname{diag}\left\{ \frac{1}{\phi_i} \right\} \dot{\boldsymbol{\phi}} + O(n^{-1/2}) \\ = \;& -2\dot{\boldsymbol{\phi}}^T \Sigma_\phi^{-1/2} H \circ (I - H) \Sigma_\phi^{-1/2} \dot{\boldsymbol{\phi}} + O(n^{-1/2}) \end{aligned}$$

where $H \circ (I - H)$ denotes the matrix with elements $h_{ij}(1 - h_{ij})$, i.e., the component-wise product of H and $I - H$. Putting the derivatives together gives the modified information matrix for $\boldsymbol{\gamma}$ as

$$E(-\ddot{\ell}_R) = \dot{\boldsymbol{\phi}}^T \Sigma_\phi^{-1/2} \left[\operatorname{diag}\left\{ \frac{v_d(\delta_i)}{\phi_i^2} - 1 \right\} + V \right] \Sigma_\phi^{-1/2} \dot{\boldsymbol{\phi}} + O(n^{-1/2})$$

where V is again the matrix $I - H$ with all elements squared.

This expression is the same as that obtained in Section 2.2 for the normal linear case with the addition of an extra term measuring the deviation of the variance of the unit deviances from χ_1^2-variance. It differs from the expression given for the modified information in Section 5 of Smyth & Verbyla (1999) in two respects. The earlier matrix did not include the extra term involved v_d just referred to. The third term of the earlier matrix also neglected a factor of $\{\phi_i \phi_j \dot{g}(\phi_i) \dot{g}(\phi_j)\}^{-1}$ associated with the elements of H^2, which means that matrix was correct to $O(n^{-1/2})$ only when $g()$ was the log-link.

7 Expectation of the Fitted Deviances

The results of Cox & Reid (1987) suggest that the derivatives of the approximate conditional log-likelihood are approximately unbiased. However the order of this approximation is unclear. Write $d_i = w_i d(y_i, \mu_i)$ and

$\hat{d}_i = w_i d(y_i, \hat{\mu}_i)$ where $\hat{\mu}_i = \mu_i(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma})$. In this section we show directly that

$$E(\hat{d}_i) = \delta_i - h_{ii}\phi + O(n^{-3/2}) \tag{11}$$

verifying that the modified score vector is unbiased to at least $O(n^{-1})$.

Consider the Taylor series expansion

$$\hat{d}_i = d_i + \frac{\partial d_i}{\partial \boldsymbol{\beta}}\boldsymbol{\epsilon} + \frac{1}{2}\boldsymbol{\epsilon}^T \frac{\partial^2 d_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\boldsymbol{\epsilon} + O(n^{-3/2})$$

where $\boldsymbol{\epsilon} = \hat{\boldsymbol{\beta}}_\gamma - \boldsymbol{\beta}$. This expansion is only useful if we have a corresponding expression for $\boldsymbol{\epsilon}$. Expanding the score vector $\dot{\ell}_\beta(\mathbf{y}; \hat{\boldsymbol{\beta}}_\gamma, \boldsymbol{\gamma})$ around the true value $\boldsymbol{\beta}$ gives

$$\boldsymbol{\epsilon} = \mathcal{I}_\beta^{-1}\dot{\ell}_\beta + \mathcal{I}_\beta^{-1}(\ddot{\ell}_\beta + \mathcal{I}_\beta)\boldsymbol{\epsilon} + \frac{1}{2}\mathcal{I}_\beta^{-1}\boldsymbol{\epsilon}^T \dddot{\ell}_\beta \boldsymbol{\epsilon} + O(n^{-3/2}) \tag{12}$$

where $\dddot{\ell}_\beta$ is the 3-way array of partial derivatives $\dddot{\ell}_{ijk} = \partial^3 \ell / \partial\beta_i \partial\beta_j \partial\beta_k$ and $\boldsymbol{\epsilon}^T \dddot{\ell}_\beta \boldsymbol{\epsilon}$ is the vector with ith element $\sum_{jk} \dddot{\ell}_{ijk}\epsilon_j \epsilon_k$. The first term $\boldsymbol{\epsilon}_1 = \mathcal{I}^{-1}\dot{\ell}_\beta$ in the expansion (12) is the one-step scoring approximation and is $O(n^{-1/2})$, while the second and third terms are $O(n^{-1})$. Substituting $\boldsymbol{\epsilon}_1$ into the second and third terms gives

$$\boldsymbol{\epsilon} = \boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2 + \boldsymbol{\epsilon}_3 + O(n^{-3/2})$$

where

$$\boldsymbol{\epsilon}_2 = \mathcal{I}_\beta^{-1}(\ddot{\ell}_\beta + \mathcal{I}_\beta)\boldsymbol{\epsilon}_1$$

and

$$\boldsymbol{\epsilon}_3 = \frac{1}{2}\mathcal{I}_\beta^{-1}\boldsymbol{\epsilon}_1^T \dddot{\ell}_\beta \boldsymbol{\epsilon}_1$$

and hence

$$\hat{d}_i = d_i + \frac{\partial d_i}{\partial \boldsymbol{\beta}}^T(\boldsymbol{\epsilon}_1 + \boldsymbol{\epsilon}_2 + \boldsymbol{\epsilon}_3) + \frac{1}{2}\boldsymbol{\epsilon}_1^T \frac{\partial^2 d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\boldsymbol{\epsilon}_1 + O(n^{-3/2}).$$

An analogous power series expansion has been used for normal nonlinear regression by Cook & Tsai (1985).

Now

$$\frac{\partial d_i}{\partial \boldsymbol{\beta}} = -\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\frac{2w_i(y_i - \mu_i)}{v(\mu_i)},$$

$$\frac{1}{2}\frac{\partial^2 d_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}^T \frac{w_i}{v(\mu_i)} - \frac{\partial}{\partial \boldsymbol{\beta}}\left\{\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\frac{1}{v(\mu_i)}\right\} w_i(y_i - \mu_i)$$

and

$$\boldsymbol{\epsilon}_1 = (\dot{\boldsymbol{\mu}}^T \Sigma_m^{-1} \dot{\boldsymbol{\mu}})^{-1} \dot{\boldsymbol{\mu}}^T \Sigma_m^{-1}(\mathbf{y} - \boldsymbol{\mu}).$$

It is straightforward to calculate

$$E\left(\frac{\partial d_i}{\partial \boldsymbol{\beta}}^T \boldsymbol{\epsilon}_1\right) = -2h_{ii}\phi_i$$

and

$$E\left\{\frac{w_i}{v(\mu_i)}\boldsymbol{\epsilon}_1^T\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}^T\boldsymbol{\epsilon}_1\right\}=h_{ii}\phi_i.$$

All other expectations except $E(d_i)=\delta_i$ are $O(n^{-3/2})$ or smaller, giving the required result (11). Note that this result is true for any generalized linear model or generalized nonlinear model. The proportion form for $b(y,\phi)$ as in (6) is not required if we define $\delta_i=E(d_i)$, and the dependence of $\boldsymbol{\mu}$ on $\boldsymbol{\beta}$ and of $\boldsymbol{\phi}$ on $\boldsymbol{\gamma}$ may be quite general.

8 The Single Sample Case

In this section we consider the case in which both the mean and the dispersion are constant. In this the simplest case it is possible to obtain exact distributions for the fitted deviances. Suppose that $y_i\sim\mathrm{ED}(\mu,\phi)$, $i=1,\ldots,n$. The maximum likelihood estimator $\hat\mu=\bar y$ is sufficient for μ and is distributed as $\mathrm{ED}(\mu,\phi/n)$. The conditional log-likelihood for $\mathbf{y}$ given $\bar y$ is, ignoring terms not depending on the parameters,

$$\begin{aligned}\ell(\mathbf{y}|\bar y;\phi) &= \sum_{i=1}^n\log f(y_i;\mu,\phi)-\log f(\bar y;\mu,\phi/n)\\ &= \frac{1}{2}\left\{\alpha\sum_{i=1}^n d(y_i,\mu)-n\kappa_d(\alpha)-n\alpha d(\bar y,\mu)+\kappa_d(n\alpha)\right\}\end{aligned}$$

where $\alpha=-1/\phi$. Sufficiency guarantees that this log-likelihood does not depend on μ, so we may substitute $\mu=\bar y$ to get

$$\ell(\mathbf{y}|\bar y;\phi)=\frac{1}{2}\left\{\alpha\sum_{i=1}^n d(y_i,\bar y)-n\kappa_d(\alpha)+\kappa_d(n\alpha)\right\}.$$

Differentiating gives

$$\frac{\partial}{\partial\alpha}\ell(\mathbf{y}|\bar y;\phi)=\frac{1}{2}\left\{\sum_{i=1}^n d(y_i,\bar y)-n\dot\kappa_d(\alpha)+n\dot\kappa_d(n\alpha)\right\}$$

which shows that

$$E\{d(y_i,\bar y)\}=-\dot\kappa_d(\alpha)+\dot\kappa_d(n\alpha)$$

For the normal and inverse-Gaussian distributions $\dot\kappa_d(\alpha)=-1/\alpha$ so

$$E(d(y_i,\bar y)=\frac{1}{\alpha}-\frac{1}{n\alpha}=\phi(1-1/n)$$

which agrees with sections 3 and 4. For the gamma distibution $\dot\kappa_d(\alpha)=2\{\log(-\alpha)-\psi(-\alpha)\}$ so

$$\begin{aligned}E\{d(y_i,\bar y)\} &= 2\{\log(-\alpha)-\psi(-\alpha)-\log(-n\alpha)+\psi(-n\alpha)\}\\ &= 2\{\psi(n/\phi)-\log n-\psi(1/\phi)\}\end{aligned}$$

As $n \to \infty$ we obtain

$$E\{d(y_i, \bar{y})\} = \delta - \frac{\phi}{n} - \frac{\phi^2}{6n^2} + O(n^{-4})$$

which agrees with sections 3 and 4.

Differentiating again gives

$$\frac{\partial^2}{\partial \alpha^2} \ell(\mathbf{y}|\bar{y}; \phi) = -\frac{n}{2} \{\ddot{\kappa}_d(\alpha) - n\ddot{\kappa}_d(n\alpha)\}$$

For the normal and inverse-Gaussian distributions, $\ddot{\kappa}_d(\alpha) = 1/\alpha^2$ so

$$\frac{\partial^2}{\partial \alpha^2} \ell(\mathbf{y}|\bar{y}; \phi) = -\frac{n}{2\alpha^2} \left\{1 - \frac{1}{n}\right\} = -\frac{\phi^2}{2}(n-1)$$

This agrees with Section 6.2 because $v_d(\delta) = \phi^2$ and $\mathbf{1}^T V \mathbf{1} = n - 1$. For the gamma distribution, $\ddot{\kappa}_d(\alpha) = 2\{1/\alpha + \dot{\psi}(-\alpha)\}$ so

$$\frac{\partial^2}{\partial \alpha^2} \ell(\mathbf{y}|\bar{y}; \phi) = -n \left\{\dot{\psi}(1/\phi) - n\dot{\psi}(n/\phi)\right\}$$

As $n \to \infty$ this can be written

$$-\frac{\phi^2}{2} \left\{ n \frac{v_d(\delta) - \phi^2}{\phi^2} + (n-1) \right\} + O(n^{-1})$$

which again agrees with Section 6.2.

9 Large Leverage Adjustment

The modified score vector computed in Section 6.1 is valid for small leverages but is evidently inconsistent for large leverages. The adjustment is not appropriate for large leverages in the gamma case because $d_i \to 0$ as $h_{ii} \to 1$ while $\delta_i^* = \delta_i - h_{ii}\phi_i \to \delta_i - \phi_i$. Motivated by the exact result in Section 8 we instead use

$$\delta_i^* = 2\{\psi[1/(h_{ii}\phi_i)] + \log h_{ii} - \psi(1/\phi_i)\}$$

for the gamma family. This agrees with the results of Sections 6.1 and 7 as $h_{ii} \to 0$ but also has the property that $\delta_i^* = 0$ for $h_{ii} = 1$. It also ensures that the score vector is exactly unbiased for single sample problems.

10 Computation of the Information Matrix

The modified information matrix $\mathcal{I}_R = E(-\ddot{\ell}_R)$ given in Section 6.2 involves the dense $n \times n$ matrix V which suggests that the computation of $\mathcal{I}_R$ is an $O(n^2)$ operation. One strategy is to approximate V with a diagonal matrix. If such as approximation is to be used, we suggest replacing V with $\mathrm{diag}(1-h_{ii})$. This produces the correct standard errors when the dispersion is constant, i.e., when the model is an ordinary generalized linear or nonlinear model. See Smyth et al. (2001) for a discussion of the accuracy of this approximation in the normal linear case.

In fact it is possible to compute the full matrix $\mathcal{I}_R$ in $O(n)$ operations using the strategy of Smyth (2002). Let Q be the orthogonal matrix from the QR-decomposition of $\Sigma_m^{-1/2}\dot{\boldsymbol{\mu}}$. Let S be the $n \times p$ matrix whose elements are the squares q_{ij}^2 of the elements of Q. Let T be the $n \times p(p-1)/2$ matrix whose elements t_{ij} are the pairwise products $q_{ik}q_{im}$ of the elements of Q where $k \neq m$. Then

$$\dot{\boldsymbol{\phi}}\Sigma_\phi^{-1/2} V \Sigma_\phi^{-1/2}\dot{\boldsymbol{\phi}} = \dot{\boldsymbol{\phi}}\Sigma_\phi^{-1/2}\mathrm{diag}(1-2h_{ii})\Sigma_\phi^{-1/2}\dot{\boldsymbol{\phi}} + B^T B$$

where

$$B = (S,\, 2T)^T\, \Sigma_\phi^{-1/2}\dot{\boldsymbol{\phi}}$$

and this is an $O(n)$ computation.

References

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

Barndorff-Nielsen, O. (1986). Inference on full or partial parameters, based on the standardized log likelihood ratio. *Biometrika* **73**, 307–322.

Barndorff-Nielsen, O. E. & Cox, D. R. (1979). Edgeworth and saddlepoint approximations with statistical applications. with discussion. *J. R. Statist. Soc.* B **41**, 279–312.

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

Blaesild, P. & Jensen, J. L. (1985). Saddlepoint formulas for reproductive exponential models. *Scandinavian Journal of Statistics* **12**, 193–202.

Cook, R. D. & Tsai, C.-L. (1985). Residuals in nonlinear regression. *Biometrika* **72**, 23–29.

Cordeiro, G. M. & McCullagh, P. (1991). Bias correction in generalized linear models. *J. R. Statist. Soc.* B **53**, 629–643.

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.

Davison, A. C. & Stafford, J. E. (1998). The score function and a comparison of various adjustments of the profile likelihood. *Canadian J. Statist.* **26**, 139–148.

Diaconis, P. & Efron, B. (1985). Testing the independence of a twoway table: new interpretations of the chi-square statistic (with discussion and rejoinder). *Ann. Statist.* **13**, 845–913.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *J. Amer. Statist. Ass.* **81**, 709–721.

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

Hall, P. & Carroll, R. J. (1989). Variance function estimation in regression: the effect of estimating the mean. *J. R. Statist. Soc.* B **51**, 3–14.

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

Jørgensen, B. (1998). *The Theory of Dispersion Models.* Chapman and Hall, New York.

Kalbfleisch, J. D. & Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *J. R. Statist. Soc.* B **32**, 175–194.

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

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

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

Nair, V. N. & Pregibon, D. (1986). A data analysis strategy for quality engineering experiments. *AT&T Technical J.* **65**, 73–84.

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 pseudolikelihood: some comparisons. *J. R. Statist. Soc.* B **54**, 273–284.

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

Nelder, J. A. & Wedderburn, R. W. M. (1972). Generalized linear models. *J. R. Statist. Soc.* A **135**, 370–384.

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.

Rigby, R. A. & Stasinopolous, D. M. (1996). Mean and dispersion additive models. In Hardle, W. & Schimek, M. G., editors, *Statistical Theory and Computational Aspects of Smoothing*, pages 215–230. Physica-Verlag.

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

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. (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.

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.