REML in Generalized Linear Models: a Conditional Approach

Gordon K. Smyth
Department of Mathematics, University of Queensland, Brisbane, Q 4072, Australia

Abstract

Residual maximum likelihood estimation (REML) is often now the preferred method for estimating parameters in linear models with correlated or heteroscedastic errors. This note shows that the residual likelihood is a conditional likelihood where the conditioning is on an appropriate sufficient statistic to remove dependence on nuisance parameters. This interpretation allows a very concise derivation of the REML likelihood without the need for transformation and generalizes naturally and exactly to non-normal models in which there is a minimal sufficient statistic for the fitted values. The conditional interpretation of REML is applied to dispersion modelling in generalized linear models. It is also applied to estimate the index parameter in a power-variance family of generalized linear models.

1 Introduction

Consider the general linear model

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{e}$$

where $\mathbf{y}$ is an $n \times 1$ vector of responses, X is an $n \times p$ design matrix of full column rank and $\mathbf{e} \sim N(0, \Omega)$ is a random vector. The variance matrix Ω is a function of a q-dimensional parameter $\boldsymbol{\gamma}$, and is assumed positive definite for γ in a neighbourhood of the true value. For any given value of $\boldsymbol{\gamma}$, maximum likelihood or generalized least squares lead to the estimator

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

for $\boldsymbol{\beta}$. The problem considered in this paper is the estimation of $\boldsymbol{\gamma}$.

Patterson and Thompson (1971) introduced residual maximum likelihood estimation as a method of estimating variance components in the case of unbalanced incomplete block designs. The actual derivation of the likelihood function was somewhat involved, and this prompted Harville (1974), Cooper & Thompson (1977) and Verbyla (1990) to give alternative derivations. In all of these the residual likelihood is represented as the marginal likelihood of the error constrasts. This makes generalization of the residual likelihood principle to non-linear models or non-normal distributions difficult since zero mean error contrasts do not generally exist. The purpose of this note is to show that the residual likelihood can be viewed also as a conditional likelihood where the conditioning is on an appropriate sufficient statistic to remove dependence on the nuisance parameters. This interpretation may be of use in teaching because it clarifies the motivation for residual maximum likelihood estimation and because it allows a very concise derivation of the REML likelihood without the need for transformation of the data. It generalizes naturally and exactly to non-normal models in which there exists a minimal sufficient statistic for the fitted values.

The plan of this paper is as follows. Conditional likelihoods are discussed briefly in Section 2. The conditional derivation of REML is given in Section 3, and its generalization to generalized linear models in Section 4. Section 5 discusses dispersion estimation in generalized linear models, including the case where the dispersion is modelled using a link-linear model as in Smyth (1989). Section 6 discusses the estimation of parameters in the variance function, in a case where the exact likelihood can be specified. Emphasis in Sections 5 and 6 is given to the one-way experimental layout, since in this case the conditional likelihood can be written down in closed form. In other cases numerical evaluation or asymptotic approximation is necessary, and methods to do this are discussed also.

2 Conditional Likelihood

Consider an arbitrary likelihood function $L(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma})$ where $\boldsymbol{\beta}$ is a vector of nuisance parameters. If there exists a statistic $\mathbf{t}(\mathbf{y}; \boldsymbol{\gamma})$, possibly depending on $\boldsymbol{\gamma}$, that is sufficient for $\boldsymbol{\beta}$ then the nuisance parameters can be eliminated from the likelihood by conditioning on $\mathbf{t}$. If the maximum likelihood estimation of $\boldsymbol{\beta}$ is a one-to-one function of $\mathbf{t}$, then it can be argued that there is no available information in $\mathbf{t}$ about $\boldsymbol{\gamma}$ in the absence of knowledge of $\boldsymbol{\beta}$, i.e., the information in $\mathbf{t}$ is entirely consumed

in estimating $\boldsymbol{\beta}$. Therefore there should be no information loss in the conditional approach. The parameter of interest, $\boldsymbol{\gamma}$, can be estimated by maximizing the conditional log-likelihood $\ell_{y|t}(\mathbf{y};\boldsymbol{\gamma}) = \ell_y(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma}) - \ell_t(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma})$ which does not depend on $\boldsymbol{\beta}$.

The idea of conditioning to remove nuisance parameters is an old one (Bartlett, 1936, 1937). Kalbleisch and Sprott (1970) give an extensive discussion including the case in which $\mathbf{t}$ depends on $\boldsymbol{\gamma}$. General expressions for approximate conditional likelihoods based on saddle point approximations have been developed by Barndorff-Nielsen (1983) and Cox and Reid (1987). A long chain of related work is referenced in Cox and Reid (1987) and McCullagh and Nelder (1989, Chapter 7). Specific application to generalized linear models in made by Davison (1988).

3 A Conditional Derivation

Let $\mathbf{y}$ and $\hat{\boldsymbol{\beta}}$ be as in Section 1. For any Ω, $\hat{\boldsymbol{\beta}}$ is complete and minimal sufficient for $\boldsymbol{\beta}$, so we can eliminate $\boldsymbol{\beta}$ from the likelihood by conditioning on $\hat{\boldsymbol{\beta}}$. Since $\hat{\boldsymbol{\beta}} \sim N[\boldsymbol{\beta}, (X^T\Omega^{-1}X)^{-1}]$, the conditional log-likelihood is $\ell_{y|\hat{\beta}}(\mathbf{y};\boldsymbol{\gamma}) = \ell_y(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma}) - \ell_{\hat{\beta}}(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Omega| - \frac{1}{2}(\mathbf{y}-X\boldsymbol{\beta})^T\Omega^{-1}(\mathbf{y}-X\boldsymbol{\beta}) + \frac{p}{2}\log(2\pi) - \frac{1}{2}\log|X^T\Omega^{-1}X| + \frac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^TX^T\Omega^{-1}X(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) = \frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\Omega| - \frac{1}{2}\log|X^T\Omega^{-1}X| - \frac{1}{2}\mathbf{y}^TP\mathbf{y}$ where $P = \Omega^{-1} - \Omega^{-1}X(X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}$. This differs from the likelihood function given by Harville (1974) and Cooper and Thompson (1977) only in that it lacks the constant Jacobian term, $-\frac{1}{2}\log|X^TX|$, since no transformation of the data has been used.

That the conditional likelihood is equivalent to the marginal distribution of the error contrasts can be seen by transforming $\mathbf{y}$ to $\hat{\boldsymbol{\beta}}$ and $\mathbf{y}_2 = L^T\mathbf{y}$ where L is a $n\times(n-p)$ matrix of full column rank satisfying $L^TX = 0$. Conditionally, $\hat{\boldsymbol{\beta}}$ is constant, so maximizing the conditional likelihood of $\mathbf{y}$ is equivalent to maximizing the conditional likelihood of $\mathbf{y}_2$. Furthermore, $\mathbf{y}_2$ and $\hat{\boldsymbol{\beta}}$ are independent so the conditional distribution of $\mathbf{y}_2$ is the same as its marginal distribution.

In the above derivation, ℓ_y is decomposed as the sum of a marginal and a conditional likelihood. Estimation of $\boldsymbol{\gamma}$ proceeds by maximizing the conditional and then $\boldsymbol{\beta}$ is estimated by maximizing the marginal $\ell_{\hat{\beta}}$.

4 Generalized Linear Models

The generalization of REML to generalized linear models can now be stated. Consider the probability density function defined by

$$f(y;\theta,\phi) = \exp[\{y\theta - \kappa(\theta)\}/\phi + c(y,\phi)]$$

For given values of ϕ, this is a linear exponential family density function. Following Jørgensen (1987), the distribution defined by $f(y;\theta,\phi)$ is called an exponential dispersion model with dispersion parameter ϕ, and is denoted $\mathrm{ED}(\mu,\phi)$ where $\mu = E(y) = \dot{\kappa}(\theta)$. Let $y_i \sim \mathrm{ED}(\mu_i,\phi_i)$, $i = 1,\ldots,n$, be independent random variables. A generalized linear model arises if a link-linear model is assumed for the means, $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ where $\mathbf{x}_i$ is a vector of covariates, $\boldsymbol{\beta}$ is an unknown p-vector of regression parameters and $g()$ is a known link function. We assume also that the dispersions ϕ_i depend on an unknown parameter vector $\boldsymbol{\gamma}$, for example through a link-linear model $h(\phi_i) = \mathbf{z}_i^T\boldsymbol{\gamma}$ as in Smyth (1989), where $\mathbf{z}_i$ is a vector of covariates.

Let $\Phi = \mathrm{diag}(\phi_i)$ and X be the $n\times p$ matrix with $\mathbf{x}_i^T$ as ith row. We assume $g()$ to be the canonical link function such that $g(\mu_i) = \theta_i$, so that $\mathbf{t} = X^T\Phi^{-1}\mathbf{y}$ is a complete sufficient statistic for $\boldsymbol{\beta}$. We define the REML estimate of $\boldsymbol{\gamma}$ to be that which maximizes the conditional likelihood of $\mathbf{y}$ given $\mathbf{t}$.

REML can also be used to estimate parameters in the variance function of a generalized linear model if the probability density can be completely specified. Let $\boldsymbol{\psi}$ be a parameter vector which indexes a family of exponential dispersion models, $\mathrm{ED}_\psi(\mu,\phi)$, and assume $y_i \sim \mathrm{ED}_\psi(\mu_i,\phi_i)$ with μ_i and ϕ_i as given above. In general the functions $\kappa()$, $c()$ and $g()$ will depend on $\boldsymbol{\psi}$, and $\mathrm{var}(y) = \phi_i v(\mu_i,\boldsymbol{\psi})$ where $v(\mu,\boldsymbol{\psi}) = \ddot{\kappa}(\theta)$. We define the REML estimates of $\boldsymbol{\psi}$ and $\boldsymbol{\gamma}$ to be those which maximize the conditional likelihood of $\mathbf{y}$ given $\mathbf{t}$.

The next two sections of this paper work out REML estimates for certain generalized linear models in which the conditional likelihood can be obtained in closed form.

5 Dispersion Estimation

5.1 The one-way layout

Consider a generalized linear model with means described by a one-way classification, i.e., let y_{ij}, $i = 1,\ldots,b$, $j = 1,\ldots,n_i$, be independent random variables with $y_{ij} \sim \mathrm{ED}(\beta_i,\gamma)$. The group mean $\bar{y}_i$ is sufficient for β_i and is distributed as $\mathrm{ED}(\beta_i,\gamma/n_i)$. The conditional log-likelihood is

$$\ell_{y|\hat{\beta}} = \sum_{i=1}^{b}\left\{\sum_{j=1}^{n_i}\log f(y_{ij};\theta_i,\gamma) - \log f(\bar{y}_i;\theta_i,\gamma/n_i)\right\}$$

$$= \sum_{i=1}^{b} \left\{ \sum_{j=1}^{n_i} c(y_{ij}, \gamma) - c(\bar{y}_i, \gamma/n_i) \right\}$$

For example suppose the y_i are normally distributed. In that case $c(y, \gamma) = -\frac{1}{2}\log\gamma - \frac{1}{2}y^2 - \frac{1}{2}\log 2\pi$ (McCullagh and Nelder, 1989), so

$$\ell_{y|\hat{\beta}} = -\frac{1}{2\gamma}D(\mathbf{y}) - \frac{N-b}{2}\log 2\pi\gamma - \frac{1}{2}\sum_{i=1}^{b}\log n_i$$

where $N = \sum n_i$ and $D(\mathbf{y}) = \sum(y_{ij} - \bar{y}_i)^2$. The conditional maximum likelihood estimator is $\hat{\gamma} = D(\mathbf{y})/(N-b)$, which is the usual residual mean square estimator of the variance in one-way analysis of variance.

If the Y_i are inverse-Gaussian, then $c(y, \gamma) = 1/(2\gamma y) - \frac{1}{2}\log\gamma - \frac{3}{2}\log y - \frac{1}{2}\log 2\pi$. In that case

$$\begin{aligned} \ell_{y|\hat{\beta}} &= -\frac{1}{2\gamma}D(\mathbf{y}) - \frac{N-b}{2}\log 2\pi\gamma \\ &\quad -\frac{3}{2}\sum_{i=1}^{b}\left(\sum_{j=1}^{n_i}\log y_{ij} - \log\bar{y}_i\right) - \frac{1}{2}\sum_i \log n_i \end{aligned}$$

where

$$D(\mathbf{y}) = \sum_{i=1}^{b}\sum_{i=1}^{n_i}\left(\frac{1}{y_{ij}} - \frac{1}{\bar{y}_i}\right) = \sum_{i=1}^{b}\sum_{i=1}^{n_i}\frac{(y_{ij}-\bar{y}_i)^2}{\bar{y}_i^2 y_{ij}}$$

The REML estimator of γ is the residual mean square deviance, $\hat{\gamma} = D(\mathbf{y})/(N-b)$.

In both normal and inverse-Gaussian cases, the REML estimator $\hat{\gamma}$ is uniform minimum variance unbiased for γ, and $(N-b)\hat{\gamma}/\gamma \sim \chi^2_{N-b}$ independently of the $\bar{y}_i$.

For the gamma distribution we have $c(y, \gamma) = \log(y/\gamma)/\gamma - \log y - \log\Gamma(1/\gamma)$ so

$$\begin{aligned} \ell_{y|\hat{\beta}} &= \frac{1}{\gamma}\sum_{i=1}^{b}\sum_{j=1}^{n_i}\log(Y_{ij}/\bar{Y}_i) - N\log\Gamma(1/\gamma) \\ &\quad + \sum_{i=1}^{b}\log\Gamma(n_i/\gamma) - \sum_{i=1}^{b}\left(\sum_{j=1}^{n_i}\log Y_{ij} - \log\bar{Y}_i\right) \end{aligned}$$

This is an exponential family likelihood with canonical parameter $\nu = 1/\gamma$, sufficient statistic $D(\mathbf{y}) = \sum_{i=1}^{b}\sum_{j=1}^{n_i}\log(Y_{ij}/\bar{Y}_i)$ and cumulant function $\lambda(\nu) = N\log\Gamma(\nu) - \sum_{i=1}^{b}\log\Gamma(n_i\nu)$. The REML estimator of γ is obtained by equating $D(\mathbf{y})$ to its expectation,

$$D(\mathbf{y}) = \dot{\lambda}(\nu) = N\psi(\nu) - \sum_{i=1}^{b} n_i\psi(n_i\nu)$$

where $\psi()$ is the digamma function. This can be compared to maximum likelihood estimation of γ which would have $\log(\nu)$ in place of $\psi(n_i\nu)$ in the last term. Compare with Cox and Reid (1987, p. 12) and McCullagh and Nelder (1989, p. 295).

5.2 Dispersion Modelling

Now consider the one-way layout with a link-linear model for the dispersion, i.e., suppose that the $Y_{ij} \sim \mathrm{ED}(\beta_i, \phi_{ij})$ and the ϕ_{ij} are a function of a q-vector of parameters $\boldsymbol{\gamma}$. The log-likelihood is

$$\begin{aligned} \ell_y &= \sum_{i=1}^{b}\sum_{j=1}^{n}\left\{\frac{1}{\phi_{ij}}[y_{ij}\theta_i - \kappa(\theta_i)] + c(y_{ij}, \phi_{ij})\right\} \\ &= \sum_{i=1}^{b}\left\{\frac{1}{\alpha_i}[t_i\theta_i - \kappa(\theta_i)] + \sum_{j=1}^{n_i} c(y_{ij}, \phi_{ij})\right\} \end{aligned}$$

where $\alpha_i = (\sum_{i=1}^{n_i}\phi_{ij}^{-1})^{-1}$, $t_i = \alpha_i\sum_{j=1}^{n_i}\phi_{ij}^{-1}y_{ij}$ and $\beta_i = \dot{\kappa}(\theta_i)$. Each t_i is sufficient for β_i and is distributed as $\mathrm{ED}(\beta_i, \alpha_i)$. The conditional log-likelihood of $\mathbf{y}$ given the t_i is

$$\ell_{y|t} = \sum_{i=1}^{b}\left\{\sum_{j=1}^{n_i} c(y_{ij}, \phi_{ij}) - c(t_i, \alpha_i)\right\}$$

5.3 General Mean Models

We now leave the one-way layout and consider general link-linear models for the μ_i. Suppose that $y_i \sim \mathrm{ED}(\mu_i, \phi_i)$, $i = 1, \ldots, n$, with link-linear models for both μ_i and ϕ_i as described in Section 3. The sufficient statistic for $\boldsymbol{\beta}$ is $\mathbf{t} = X^T\Phi^{-1}\mathbf{y}$, and this has cumulant function

$$\kappa_t(\boldsymbol{\beta}) = \sum_{i=1}^{n}\phi_i^{-1}\kappa(\mathbf{x}_i^T\boldsymbol{\beta})$$

where $\kappa()$ is the cumulant function of the y_i. The cumulant generating function of $\mathbf{t}$ is $K(\mathbf{s}) = \kappa_t(\boldsymbol{\beta}+\mathbf{s}) - \kappa_t(\boldsymbol{\beta})$, so the probability density function of $\mathbf{t}$ is given by

$$f(\mathbf{t}) = \int \exp\left\{\sum_{i=1}^{n}\frac{\kappa(\mathbf{x}_i^T(\boldsymbol{\beta}+\mathbf{s})) - \kappa_t(\mathbf{x}_i^T\boldsymbol{\beta})}{\phi_i} - \mathbf{s}^T\mathbf{t}\right\} d\mathbf{s}$$

The required conditional log-likelihood is

$$\ell_{y|t} = \ell_y(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\gamma}) - \log f(\mathbf{t})$$

which doesn't depend on $\boldsymbol{\beta}$. Except in the normal case, the cumulant generating function of $\mathbf{t}$ is difficult to invert analytically, so either numerical evaluation or approximation will generally be necessary.

One possible approximation is to use, following a suggestion of A. T. James (James and Wiskich, 1993), the asymptotic normal approximation to the distribution of $\hat{\boldsymbol{\beta}}$. This leads to the approximate conditional log-likelihood

$$\begin{aligned}\ell_{y|\hat{\beta}} &= \ell_y(\mathbf{y};\boldsymbol{\beta},\boldsymbol{\gamma}) + \frac{p}{n}\log 2\pi - \frac{1}{2}\log|X^TWX| \\ &\quad + \frac{1}{2}(\hat{\boldsymbol{\beta}}-\beta)X^TWX(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\end{aligned}$$

where $W = \mathrm{diag}\{\phi_i^{-1}v(\mu_i)\}$ and $v()$ is the variance function defined by $v(\mu) = \dot{\kappa}(\theta)$. This expression depends on $\boldsymbol{\beta}$, but only slightly, so we can set $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, yielding the approximation

$$\ell_y(\mathbf{y},\hat{\boldsymbol{\beta}},\boldsymbol{\gamma}) + \frac{p}{n}\log 2\pi - \frac{1}{2}\log|X^TWX| \qquad (1)$$

i.e., the log-profile likelihood for $\boldsymbol{\gamma}$ adjusted by the log-determinant of the covariance matrix of $\hat{\boldsymbol{\beta}}$. This method is applicable even when the link function $g()$ is not canonical, although then $\mathbf{t}$ is not sufficient so it is impossible to entirely eliminate $\boldsymbol{\beta}$ from the estimation of $\boldsymbol{\gamma}$.

Another approach which leads to the same approximation in this case is to use the modified profile likelihood of Barndorff-Nielsen (1983) together with a suggestion of Cox and Reid (1987) for orthogonal parameters. The modified profile likelihood for $\boldsymbol{\gamma}$ is

$$\ell_y(\mathbf{y};\hat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\gamma}) - \frac{1}{2}\log|j_{\beta\beta}| + \log\left|\frac{\partial\hat{\boldsymbol{\beta}}}{\partial\hat{\boldsymbol{\beta}}_\gamma}\right|$$

where $\hat{\boldsymbol{\beta}}_\gamma$ is the maximum likelihood estimator for $\boldsymbol{\beta}$ for given $\boldsymbol{\gamma}$, $\hat{\boldsymbol{\beta}}$ is the unrestricted maximum likelihood estimator, $j_{\beta\beta}$ is the observed information matrix for $\boldsymbol{\beta}$ evaluated at $\hat{\boldsymbol{\beta}}_\gamma$, and $\ell_y(\mathbf{y};\hat{\boldsymbol{\beta}}_\gamma,\boldsymbol{\gamma})$ is the log-profile likelihood for $\boldsymbol{\gamma}$. Since $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are orthogonal, $\hat{\boldsymbol{\beta}}_\gamma$ varies only slowly with $\boldsymbol{\gamma}$ so the derivative term $\partial\hat{\boldsymbol{\beta}}/\partial\hat{\boldsymbol{\beta}}_\gamma$ can be neglected. For the current model we have

$$j_{\beta\beta} = X^TWX$$

and the modified profile likelihood is, apart from constants, the same as (1).

For normal linear models, the approximate conditional likelihood (1) is precisely the same as the standard residual likelihood given in Section 3. When the y_i are inverse-Gaussian and $\boldsymbol{\gamma}$ is scalar, modified profile likelihood leads to the residual mean deviance as the estimator of the dispersion. In other cases, the effectiveness of the approximation needs to be evaluated. This is not done here as our primary intention is to clarify the exact conditional approach.

Table 1: Simulation results for estimating γ and ϕ. One thousand data sets were generated. True values are $\gamma = 1.5$ and $\phi = 1.0$.

(a) Estimation of γ

	Mean	Std	MSE
Maximum likelihood	1.4731	0.0711	0.0058
REML	1.4873	0.0769	0.0061
Extended Quasi-Lik.	1.2345	0.0961	0.0798
Pseudo-Likelihood	1.5494	0.1894	0.0383

(b) Estimation of ϕ

	Mean	Std	MSE
Maximum likelihood	0.9010	0.1809	0.0425
REML	0.9915	0.2048	0.0420
Extended Quasi-Lik.	1.0008	0.2057	0.0423
Pseudo-Likelihood	0.9015	0.1904	0.0460

6 Variance Function Estimation

Suppose that $\boldsymbol{\gamma}$ is an unknown parameter than indexes a family of generalized linear models. That is, suppose that $y_i \sim \mathrm{ED}_\gamma(\mu_i,\phi)$, $i = 1,\ldots,n$ where $g(\mu_i) = \mathbf{x}_i^T\boldsymbol{\beta}$ and $\mathrm{var}(y_i) = \phi v(\mu_i,\gamma)$. The REML estimators of γ and ϕ are those which maximize the conditional likelihood of $\mathbf{y}$ given $X^T\mathbf{y}$. The purpose of this section is to consider a potentially important example, that of the compound Poisson exponential dispersion models introduced by Jørgensen (1987). The compound Poisson models have power variance functions $v(\mu,\gamma) = \mu^\gamma$ with γ between one and two. The compound Poisson distributions converge to Poisson as $\gamma \to 1$ and to gamma as $\gamma \to 2$, and so may be viewed as intermediate between the Poisson and gamma families. They are also positive and continuous except for mass at zero. Compound Poisson generalized linear models have potential applications in modelling continuous data with exact zeros, such as weather variables, insurance claims and waiting times, but the problem of estimating γ has not been satisfactorily solved (Burridge, 1987; Gilchrist, 1987).

The compound Poisson density function has been derived by Jørgensen (1992). See also Tweedie (1984). It has $\theta = \mu^{2-\gamma}/(2-\gamma)$, $\kappa(\theta) = \mu^{1-\gamma}/(1-\gamma)$ and

$$c(y,\phi) = \log\sum_{j=1}^{\infty}\frac{\{\alpha(\alpha+1)^{\alpha+1}\phi^{-\alpha-1}y^\alpha\}^j}{j!\Gamma(j\alpha)}$$

where $\alpha = (2-\gamma)/(\gamma-1)$. Tweedie (1984, p. 586) has identified $\exp c(y,\phi)$ as an instance of Wright's (1933) generalized Bessel function. It is not expressible however

in terms of the more common Bessel functions.

A simulation experiment was conducted to compare four estimators of ϕ and γ. These were maximum likelihood estimation, REML, extended quasi-likelihood (Nelder and Pregibon, 1987) and pseudo-likelihood (Davidian and Carroll, 1987). Data was simulated from a one-way classification with $n_1 = \ldots = n_5 = 10$, $\boldsymbol{\beta} = (0.1, 0.5, 1, 2, 5)^T$, $\phi = 1$ and $\gamma = 1.5$. One thousand such data sets were generated and, for each, γ and ϕ were estimated using the four methods. The results are tabulated in Table 1.

REML had the smallest bias for estimating γ. Maximum likelihood had the smallest standard deviation, and also the smallest mean square error, although this was not significantly different from that of REML. Pseudo-likelihood was also approximately unbiased, but with a largest standard deviation. Extended quasi-likelihood had a competitive standard deviation, but was biased down giving it the largest mean square error. Experimentation showed that the bias was due to the offset of 1/6 for zero observations. Positive and negative biases could be achieved by relatively small changes to this offset.

REML and extended quasi-likelihood were almost equally effective for estimating ϕ. The maximum likelihood estimator had again the smallest standard deviation and a mean square error not significantly greater than REML and extended quasi-likelihood, but was biased down by about 10%, as expected given the group size of ten. The pseudo-likelihood estimator was also biased down by about the same amount, despite incorporating a correction for degrees of freedom as recommended by Davidian and Carroll (1987).

We conclude that REML, in its conditional likelihood guise, is successful in reducing the bias of the maximum likelihood estimator while incurring minimal inflation to its standard deviation. Neither of its competitors, extended-quasi and pseudo likelihood, were as successful in doing this.

References

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

Bartlett, M. S. (1936). The information available in small samples. *Proc. Camb. Phil. Soc.*, **32**, 560–566.

Bartlett, M. S. (1937) Properties of sufficiency and statistical tests. *Proc. R. Soc.* A, **160**, 268–282.

Burridge, J. (1987). Discussion of Dr Jørgensen's paper. *J. R. Statist. Soc. B*, **49**, 150–151.

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

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

Davison, A. C. (1988) Approximate conditional inference in generalized linear models. *J. R. Statist. Soc.* B, **50**, 445–461.

Gilchrist, R. (1987). Discussion of Dr Jørgensen's paper. *J. R. Statist. Soc. B*, **49**, 145–147.

James, A. T. and Wiskich, J. T. (1993). *t*-REML for robust heteroscedastic regression analysis of mitochondrial power. *Biometrics* **49**, 339–356.

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

Jørgensen, B. (1992). *The theory of exponential dispersion models and analysis of deviance.* Monografias de Matemática No. 51, Instituto de Mathemática pura e Aplicada, Rio de Janeiro.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving a large number of nuisance parameters. *J. R. Statist. Soc.* B, **32**, 175–208.

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221–231.

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

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. (1990). A conditional derivation of residual maximum likelihood. *Aust. J. Statist.*, **32**, 221–224.

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

Wright, E. M. (1933). On the coefficients of power series having essential singularities. *J. London Math. Soc.*, **8**, 71–9.