Exponential Dispersion Models and the Gauss-Newton Algorithm

Gordon K. Smyth

Department of Mathematics, University of Queensland,
St Lucia 4067, Australia.

Abstract

It is well known that the Fisher scoring iteration for generalized linear models has the same form as the Gauss-Newton algorithm for normal regression. This note shows that exponential dispersion models are the most general families to preserve this form for the scoring iteration. Therefore exponential dispersion models are the most general extension of generalized linear models for which the analogy with normal regression is maintained. The multinomial distribution is used as an example.

Keywords: generalized linear models, scoring algorithm, multinomial distribution, quasi-likelihood.

1 Introduction

Recently, Jørgensen (1987) has shown how to construct a class of multivariate linear exponential families, called *exponential dispersion models*, which include as a special case the generalized linear model families of Nelder and Wedderburn (1972). These models were also discussed by McCullagh (1983) and others, including this author in an unpublished ANU PhD Thesis. Nelder and Wedderburn (1972) and Wedderburn (1974) showed that the Fisher scoring iteration for generalized linear models is a simple generalization of the Gauss-Newton algorithm for normal models, and much use is made of the analogy with normal regression in generalized linear model practice. The purpose of this note is to point out that exponential dispersion models are the most general families for which the Gauss-Newton structure of the scoring iteration is preserved. This result is implicit in the work of McCullagh and Jørgensen, but is worth emphasizing because it means that exponential dispersion models are the most general extension of generalized linear models for which the analogy with normal regression is maintained.

Exponential dispersion models increase the range of univariate variance functions for which generalized linear type models exits. For example, the variance may be proportional to any power of the mean μ^p with $p \geq 1$ or $p \leq 0$. However, finding a multivariate exponential dispersion model for a given mean and covariance function is more difficult. Each exponential dispersion family is generated from a cumulant function, from which the mean vector and covariance matrix are obtained as first and second derivative respectively. This imposes considerable structure on the mean and

covariance, so that multivariate exponential dispersion models exist only for special mean-covariance relationships.

Even if the data is not from an exponential dispersion model, use of the Gauss-Newton iteration can be justified by quasi-likelihood theory. Wedderburn (1974) and McCullagh (1983) show that the Gauss-Newton iteration produces consistent estimates provided only that the mean and covariance of the observations, rather their full distributional form, are correctly specified. Because of the relative paucity of multivariate dispersion models, this approach is likely to be even more important in the multivariate case than it has been in the univariate.

2 Exponential Dispersion Models and The Scoring Iteration

Much of the attractive unity of ordinary generalized linear models arises from the common form of the scoring iteration for the parameters in the linear model. The iteration can be expressed as a linear regression, with the current residuals as dependent variable and the current variance estimates as inverse weights. This largely explains the tendency for normal regression methods to carry over approximately to the more general class of models. The generalized linear model inherits local properties, such as standard errors, score tests and measures of influence, from the linear regression in the iteration. See for example Pregibon (1981), Cook and Weisberg (1982) and McCullagh and Nelder (1983). It is therefore natural to try to preserve the form of the scoring iteration in any extension of generalized linear models. We show below that exponential dispersion models are the most general families to do so.

Consider a normal observation y with mean vector $\mu(\beta)$ and covariance matrix $V\sigma^2$, with β and σ^2 unknown and V known. The scoring iteration consists of separate iterations for β and σ^2 since the two are orthogonal. Letting ℓ be the log-likelihood function, the score vector for β is

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma^2} \frac{d\mu}{d\beta}^T V^{-1}(y - \mu),$$

where $d\mu/d\beta = (\partial \mu_i / \partial \beta_j)$ is the matrix of partial derivatives, and the component of the Fisher information matrix corresponding to β is

$$\mathcal{I}_\beta = \frac{1}{\sigma^2} \frac{d\mu}{d\beta}^T V^{-1} \frac{d\mu}{d\beta}.$$

The scoring iteration for β is

$$\beta^{k+1} = \beta^k + \mathcal{I}_\beta^{-1} \frac{\partial \ell}{\partial \beta} = \beta^k + (\frac{d\mu}{d\beta}^T V^{-1} \frac{d\mu}{d\beta})^{-1} \frac{d\mu}{d\beta}^T V^{-1}(y - \mu) \Big|_{\beta = \beta^k}$$

which does not depend on σ^2. This is the Gauss-Newton algorithm for least squares estimation of β.

Note that it would not greatly complicate matters if V were to depend on β, provided the above formulae were preserved. Let ℓ now be an unknown log-likelihood function with scoring iteration

$$\beta^{k+1} = \beta^k + F(\beta^k) \tag{1}$$

where

$$F(\beta) = (\frac{d\mu}{d\beta}^T V^{-1} \frac{d\mu}{d\beta})^{-1} \frac{d\mu}{d\beta}^T V^{-1}(y - \mu) \tag{2}$$

for given functions $\mu(\beta)$ and $V(\beta)$. Let $\mathcal{I}$ be the corresponding information matrix. Assume that $d\mu/d\beta$ and V are of full rank, and that the support of the distribution does not depend on β. Now $F(\beta) = \mathcal{I}^{-1}\partial\ell/\partial\beta$, so $\mathrm{E}(F) = 0$ and $\mathrm{var}(F) = \mathcal{I}^{-1}$. The first identity shows that $\mathrm{E}(y) = \mu$, and the second that

$$\mathcal{I}^{-1} = (\frac{d\mu}{d\beta}^T V^{-1} \frac{d\mu}{d\beta})^{-1} \frac{d\mu}{d\beta}^T V^{-1}\mathrm{var}(y)V^{-1}\frac{d\mu}{d\beta}(\frac{d\mu}{d\beta}^T V^{-1}\frac{d\mu}{d\beta})^{-1},$$

from which it is apparent that $\mathrm{var}(y)$ is proportion to V. Therefore we must have

$$\frac{\partial\ell}{\partial\beta} = \frac{1}{\phi}\frac{d\mu}{d\beta}^T V^{-1}(y - \mu)$$

and

$$\mathcal{I} = \frac{1}{\phi}\frac{d\mu}{d\beta}^T V^{-1}\frac{d\mu}{d\beta},$$

where ϕ is some proportionality constant. Now $\partial\ell/\partial\beta$ is linear in y, i.e.

$$\phi\frac{d\ell}{d\beta} = \frac{d\mu}{d\beta}^T V^{-1}y - \frac{d\mu}{d\beta}^T V^{-1}\mu,$$

so ℓ must have the form

$$\ell = \{y^T\theta(\beta) - \kappa(\beta) + c(y, \phi)\}/\phi \tag{3}$$

where c is some function not depending on β, θ is such that

$$\frac{d\theta^T}{d\beta} = \frac{d\mu}{d\beta}^T V^{-1} \tag{4}$$

and κ is such that

$$\frac{d\kappa}{d\beta} = \frac{d\mu}{d\beta}^T V^{-1}\mu. \tag{5}$$

Now (4) shows that

$$\frac{d\theta}{d\mu} = V^{-1}$$

which, when substituted into (5), implies that

$$\frac{d\kappa}{d\theta} = \mu,$$

3

and hence that

$$\frac{d^2\kappa}{d\theta^T d\theta} = V.$$

The likelihood (3) represents the family of exponential dispersion distributions derived by Jørgensen (1987) starting with $\kappa(\theta)$ as a cumulant generating function.

Generalized linear models assume that the relationship between μ and the unknown parameter vector β takes the form

$$g(\mu) = X\beta, \tag{6}$$

where g is a known link function and X is a matrix of covariates, so that

$$\frac{d\mu}{d\beta} = \frac{dg}{d\mu}^{-1} X.$$

Also g acts component-wise on μ so that $dg/d\mu$ is diagonal. For ordinary generalized linear models then, the scoring iteration can be expressed as iteratively reweighted least squares,

$$\beta^{k+1} = \beta^k + (X^T W X)^{-1} X^T W z$$

with diagonal weight matrix $W = (dg/d\mu)^{-2} V$ and working vector $z = (dg/d\mu)(y - \mu)$, or even more compactly as

$$\beta^{k+1} = (X^T W X)^{-1} X^T W z \tag{7}$$

with $z = (dg/d\mu)(y - \mu) + X\beta$. The construction of the working vector z has no computational advantage, but highlights the analogy with normal regression. On the other hand the link linear model (6) seems to be important in limiting the amount of nonlinearity in the model and increasing the stability of the scoring iteration. See Smyth (1987) and Kass and Smyth (1990).

With a full covariance matrix V, the advantages of $dg/d\mu$ being diagonal are less obvious. Indeed g usually is composite in multivariate applications. This does not prevent the scoring iteration from being written in the weighted linear regression form (7) with a non-diagonal weight matrix W.

Green (1984) casts the scoring iteration into a weighted least squares mould for more general models than those considered here, in fact for any likelihood which is a function of its expected values. However this involves replacing the residuals $y - \mu$ with $d\ell/d\mu$ and so on, and the interpretation of the resulting iteration is much less straightforward than that of (1) and (2).

3   Multinomial Data

The multinomial distribution is undoubtedly the most commonly used exponential dispersion model with non-diagonal variance matrix. It is possible to fit the logistic multinomial model in GLIM, by fitting a Poisson model with log-link to all the cell counts and conditioning on a factor corresponding to the row totals (Aitkin *et al*, 1989). This however involves a large number of redundant parameters, making it suitable

4

only for small problems, and does not extend to other multinomial models such as the ordinal models of McCullagh (1980). It is therefore highly desirable to be able to fit the multinomial distribution directly.

Let $y = (y_1, \ldots, y_c)^T$ be an observation from a multinomial distribution satisfying $\sum_{i=1}^{c} y_i = n$. The log-likelihood function kernel is

$$
\begin{aligned}
\ell(y; p_1, \ldots, p_{c-1}) &= \sum_{i=1}^{c-1} y_i \log p_i + \left(n - \sum_{i=1}^{c-1} y_i \right) \log(p_c) \\
&= \sum_{i=1}^{c-1} y_i \log (p_i/p_c) + n \log(p_c)
\end{aligned}
$$

where the p_i are the cell probabilities. This is an exponential dispersion model with

$$
\theta_i = \log (p_i/p_c)
$$

for $i = 1, \ldots, c-1$ and

$$
\kappa(\theta) = -n \log(p_c) = n \log(1 + \sum_{i=1}^{c-1} e^{\theta_i}).
$$

This confirms that

$$
\mu_i = \frac{\partial \kappa}{\partial \theta_i} = n p_i
$$

and

$$
\mathrm{cov}(y_i, y_j) = \frac{\partial^2 \kappa}{\partial \theta_i \partial \theta_j} = n p_i \delta_{ij} - n p_i p_j
$$

where δ_{ij} is the kronecker delta function. In matrix terms

$$
V = \langle \mu \rangle - \mu \mu^T / n \tag{8}
$$

where $\langle \cdot \rangle$ represents the diagonal matrix with the components of the vector down the diagonal.

Now let y be a series of r independent multinomial observations in vector form, $y = (y_{ij})$ with $\sum_{j=1}^{c} y_{ij} = n_i$ say. Let θ be the vector (θ_{ij}) with $\theta_{ij} = \log(p_{ij}/p_{ic})$. Then the exponential dispersion model for y has cumulant generator

$$
\kappa(\theta) = \sum_{i=1}^{r} n_i \log(1 + \sum_{j=1}^{c-1} e^{\theta_{ij}}),
$$

and $\mathrm{var}(y) = V$ is block diagonal with r blocks of the form (8). The multinomial logistic model assumes that

$$
\theta = X\beta
$$

where X is a matrix of covariates. Hence

$$
\frac{d\mu}{d\beta} = \frac{d\mu}{d\theta} \frac{d\theta}{d\beta} = VX,
$$

5

so the scoring iteration for β is

$$\beta^{k+1} = \beta^k + (X^T V X)^{-1} X^T (y - \mu).\tag{9}$$

This can be evaluated efficiently using the fact that each block of V is a rank one displacement of a diagonal matrix. Writing $\mu_i = (\mu_{ij})_{j=1}^{c-1}$ and X_i for the corresponding $c-1$ rows of X,

$$X^T V X = \sum_{i=1}^{r} [X_i^T \langle \mu_i \rangle X_i - (X_i^T \mu_i)(X_i^T \mu_i)^T / n_i]$$

which can be calculated in $O(rc \dim \beta)$ operations. See also Green (1984).

The ordinal models of McCullagh (1980) assume that

$$g(\gamma) = X\beta$$

where γ is the vector (γ_{ij}) of cumulative expectations, and g is a link function acting component-wise on γ. The cumulative expectations are defined by $\gamma_{ij} = \sum_{a=1}^{j} \mu_{ia}$ for $j = 1, \ldots, c-1$ and $i = 1, \ldots, r$. That is, $\gamma = H\mu$ where H is block diagonal with blocks that are lower triangles of ones. Then

$$\frac{d\mu}{d\beta} = \frac{d\mu}{d\gamma}\frac{d\gamma}{d\beta} = H^{-1}\frac{dg}{d\gamma}^{-1} X,$$

and $dg/d\gamma$ is diagonal. Again efficient programming uses the special structures of V and H. The blocks of H^{-1} are the $c-1 \times c-1$ difference operators

$$\begin{pmatrix} 1 & & & 0 \\ -1 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & -1 & 1 \end{pmatrix}$$

so forming $d\mu/d\beta$ is a very inexpensive calculation. The blocks of V^{-1} are the $c-1 \times c-1$ matrices

$$\langle \mu_i \rangle^{-1} + uu^T / \mu_{ic},$$

where u is a $c-1$ vector of ones, and

$$x_i = \frac{d\mu_i}{d\beta}^T u$$

is simply the last column of the ith block of $X^T (dg/d\gamma)^{-1}$ since $H_i^{-T} u$ has only one nonzero component. So

$$\frac{d\mu^T}{d\beta} V^{-1} \frac{d\mu}{d\beta} = \sum_{i=1}^{r} \left[\frac{d\mu_i^T}{d\beta} \langle \mu_i \rangle^{-1} \frac{d\mu_i}{d\beta} + \frac{x_i x_i^T}{\mu_{ic}} \right]$$

and

$$\frac{d\mu^T}{d\beta} V^{-1} (y - \mu) = \sum_{i=1}^{r} \left[\frac{d\mu_i^T}{d\beta} \langle \mu_i \rangle^{-1} (y_i - \mu_i) + x_i \frac{\mu_{ic} - n_{ic}}{\mu_{ic}} \right],$$

where y_i and μ_i are the ith blocks of $c-1$ components of y and μ respectively. See also McCullagh (1980). The above formulae require $O(rc \dim \beta)$ operations. Perhaps the best known software to do this calculation is McCullagh's (1979) program PLUM, which is fairly widely available in the public domain.

References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989). *Statistical Modelling in GLIM.* Oxford Statistical Science Series, 4, Oxford: Clarendon Press.

Cook, R.D. and Weisberg, S. (1982). *Residuals and influence in regression.* London: Chapman and Hall.

Green, P.J. (1984). Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. R. Statist. Soc. B*, 46, 149–192.

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

Kass, R.E. and Smyth, G.K. (1990). The rate of convergence of the Fisher scoring: a geometricnterpretation. Technical Report, Department of Mathematics, University of Queensland, 18pp.

McCullagh, P. (1979). *PLUM: an interactive computer package for analysing ordinal data.* Unpublished manual.

McCullagh, P. (1980). Regression models for ordinal data. *J. R. Statist. Soc. B*, 42, 109–142.

McCullagh, P. (1983). Quasi-likelihood functions. *Ann. Statist.*, 11, 59–67.

McCullagh, P. and Nelder, J.A. (1983). *Generalized linear models.* Chapman and Hall: London.

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

Pregibon, D. (1981). Logistic regression diagnostics. *Ann. Statist.*, 9, 705–724.

Smyth, G. K. (1987). Curvature and convergence. *1987 Proceedings of the Statistical Computing Section.* American Statistical Association, Virginia, 278–283.

Wedderburn, R.W.M. (1974). Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. *Biometrika,* 61, 439–447.