Journal of Modern Mathematics and Statistics 3 (1): 25-28, 2009
ISSN: 1994-5388
© Medwell Journals, 2009

On the Equivalence of Two Quasi-Newton Schemes in Generalized Linear Models

Mbe Egom Nja
Department of Mathematics/Statistics, Cross River University of Technology, Calabar, Nigeria

Abstract: The Iterative Weighted Least Squares and the Fisher's Scoring methods are two most commonly used iterative maximum likelihood optimization methods in generalized linear models. The Fisher's Scoring method is given in terms of the gradient vector. While, the Iterative Weighted Least Squares method is based on the adjusted dependent vector. Using the relation between the expected Hessian matrix and weighted sum of squares, established for quasi-likelihood function and the link between the expected Hessian and the weighted sum of cross product, a proof of the theorem on the equivalence of the two quasi-Newton schemes is presented.

Key words: Fisher's information, Gradient vector, Hessian matrix, likelihood function, quasi-likelihood function, weight function

INTRODUCTION

The maximum likelihood estimator is an alternative to the minimum variance unbiased estimator (Scott and Nowak, 2006). In generalized linear models parameter estimation is accomplished by iterative maximum likelihood procedure. Generalized linear models extend the idea of non linear regression to models with non-normal error distribution (Smyth, 2002). This is done (Allen, 1987) by replacing the objective function, $f_{(x)}$ with the log likelihood function l (θ, y).

Stokes *et al.* (1975), McCullagh and Nelder (1992) used the logit defined as the logarithm of the ratio between the probability of success and the probability of failure to demonstrate the concept of link function in generalized linear models. Based on this, the weight function of the Iterative weighted least squares method is defined.

Definitions: Let $\tilde{\beta}^{(k)}$ be estimate of parameter vector β at iteration k, then the Fisher's Scoring method is given as

$$\hat{\beta}^{(k-1)} = \hat{\beta}^{(k)} - \left[-E(H^{(k)})\right]^{-1} g^{(k)}$$

where,

H = The Hessian matrix and g is the gradient vector

$$H = \frac{\partial^2 l}{\partial\beta_r \partial\beta_s},\ g = \frac{\partial l}{\partial\beta}$$

The iterative weighted least squares method is a maximum likelihood estimation method for generalized linear models. The solution is given as follows:

$$\hat{\beta}^{(k+1)} = (X'WX)^{-1}X'WZ$$

where,

Z = Adjusted dependent vector

$$W = V^{-1}\left(\frac{d\mu}{d\eta}\right)^2,\ \eta = \beta_0 + \Sigma x_{ij}\beta_j$$

is the systematic component of the model. X is the design matrix. Wedderburn (1974) stated the theorem on the equivalence of the Fisher's Scoring method and the Iterative Weighted Least Squares method and showed that

$$E\left(\frac{\partial k \partial k}{\partial\beta_i \partial\beta_j}\right) = -E\left(\frac{\partial^2 k}{\partial\beta_i \partial\beta_j}\right) = \frac{1}{V(\mu)}\ \frac{\partial\mu}{\partial\beta_i}\ \frac{\partial\mu}{\partial\beta_j} =$$

$$X^1WX = -E(H),\ H = \frac{\partial^2 k}{\partial\beta_i \partial\beta_j}$$

where,

k = Quasi-likelihood function having properties similar to those of the log likelihood function

McCullagh and Nelder (1992) using the log likelihood function, l and an adjusted component β_r established that

$$\left(A\beta^*\right)_r = \Sigma W x_r z = X'WZ = -E(H)$$

Where:

$$A\beta^* = A\beta + g,\ H = \frac{\partial^2 l}{\partial\beta_i \partial\beta_j}$$

These facts are used to present a formal proof of the theorem. Wedderburn (1974) established in his theorem (1) and proof.

MATERIALS AND METHODS

Theorem 1: Let y_i (i = 1,...,n) be independent observations with expectations μ_i and variances V (μ_i). Let K (y_i, μ_i) be the quasi-likelihood function of the observation y_i and suppose that μ is expressed as a function of parameters $\beta_i,...,\beta_m$ then

$$E\left(\frac{\partial k \partial k}{\partial\beta_i\partial\beta_j}\right) = -E\left(\frac{\partial^2 k}{\partial\beta_i\partial\beta_j}\right) = \frac{1}{V(\mu)}\frac{\partial\mu}{\partial\beta_i}\frac{\partial\mu}{\partial\beta_j}$$

Proof : Note that

$$E\left(\frac{\partial k \partial k}{\partial\beta_i\partial\beta_j}\right) = E\left(\frac{\partial k}{\partial\mu}\right)^2 \frac{\partial\mu}{\partial\beta_i}\ \frac{\partial\mu}{\partial\beta_j},$$

$$= E\left(\frac{(y-\mu)^2}{\{V(\mu)\}^2}\right)\frac{\partial\mu}{\partial\beta_i}\frac{\partial\mu}{\partial\beta_j}, = \frac{1}{V(\mu)}\frac{\partial\mu}{\partial\beta_i}\frac{\partial\mu}{\partial\beta_j}$$

since, V (μ) = var (y).

Also, we have

$$-E\left(\frac{\partial^2 k}{\partial\beta_i\partial\beta_j}\right) = -E\left(\frac{\partial}{\partial\beta_j}\left\{\frac{z-\mu}{V(\mu)}\frac{\partial\mu}{\partial\beta_j}\right\}\right),$$

$$= -E\left((z-\mu)\frac{\partial}{\partial\beta_j}\left\{\frac{1}{V(\mu)}\frac{\partial\mu}{\partial\beta_j}\right\} - \frac{1}{V(\mu)}\frac{\partial\mu}{\partial\beta_i}\frac{\partial\mu}{\partial\beta_j}\right),$$

$$= \frac{1}{V(\mu)}\frac{\partial\mu}{\partial\beta_i}\frac{\partial\mu}{\partial\beta_j}$$

which completes the proof.

$$\frac{1}{V(\mu)}\frac{\partial\mu}{\partial\beta_i}\frac{\partial\mu}{\partial\beta_j} = Wx_ix_j = X^1WX$$

The quasi-likelihood function and the log likelihood function have similar properties. For this reason, we consider the expectation of the Hessian matrix defined on the log likelihood function.

The loglikelihood function and fisher's information . The loglikelihood for a binary response variable can be written as:

$$1(\mu;y) = \sum\left[y_i \log\left(\frac{\mu_i}{1-\mu_i}\right) + m_i\log(1-\mu_i)\right]$$

which becomes

$$1(\beta;y) = \sum_i\sum_j y_ix_{ij}\beta_j - \sum m_i \log\left(1+\exp\sum x_{ij}\beta_j\right)$$

From

$$\frac{\partial 1}{\partial\mu_i} = \frac{y_i - m_i\mu_i}{\mu_1(1-\mu_i)},\ \frac{\partial 1}{\partial\beta_r} = \sum\frac{y_i - m_i\mu_i}{\mu_1(1-\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}$$

$$= \sum\frac{y_i - m_i\mu_i}{\mu_1(1-\mu_i)}\frac{\partial\mu_i}{\partial\eta_i}x_{ir}$$

since

$$\frac{\partial\mu_i}{\partial\beta_r} = \frac{d\mu_i}{d\eta_i}\frac{\partial\eta_i}{\partial\beta_r} = \frac{d\mu_i}{d\eta_i}x_{ir}$$

so that

$$\frac{\partial 1}{\partial\beta_r} = \sum_i\frac{y_i - m_i\mu_i}{\mu_i(1-\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ir}$$

The fisher information for β is given (Silvey, 1970) as -E ($\partial^2 1/\partial\beta_r\partial\beta_s$):

$$-E\left(\frac{\partial^2 1}{\partial\beta_r\partial\beta_s}\right) = A$$

$$= \sum\frac{m_i}{\mu_i(1-\mu_i)}\frac{\partial\mu_i}{\partial\beta_r}\frac{\partial\mu_i}{\partial\beta_s},$$

$$= \sum\frac{m_i}{\mu_i(1-\mu_i)}\frac{d\mu_i}{d\eta_i}x_{ir}\frac{d\mu_i}{d\eta_i}x_{is}$$

$$= \sum m_i\frac{(d\mu_i/d\eta_i)^2}{\mu_i(1-\mu_i)}x_{ir}x_{is}, = \{X'WX\}_{rs}$$

Where:

$$W = \text{Diag}\left\{m_i\left(\frac{d\mu_i}{d\eta_i}\right)^2 / \mu_i(1-\mu_i)\right\}$$

Therefore, -E (H) =X'WX. = Fisher's information.

The iterative Weighted Least Squares method is derivable form the Fisher's scoring method as shown by the theorem and proof.

Theorem 2 (the main theorem): The iterative weighted least squares and the Fisher's Scoring methods are equivalent optimization schemes in generalized linear models.

Proof: Let the adjusted dependent variate

$$z = \eta + (y-\mu)\frac{d\eta}{d\mu}$$

where,

η = Systematic component of the model

Let the gradient vector, g = $\partial 1/\partial\beta$ and A = -E ($\partial^2 1/\partial\beta_r\partial\beta_s$)

Let

$$\delta\beta = \beta^{(k+1)} - \beta^{(k)}$$

The replacement of $\partial^2 1/\partial\beta_r\partial\beta_s$ with ($\partial^2 1/\partial\beta_r\partial\beta_s$) in the Newton-Raphson method yields the Fisher's Scoring method.

From the Newton-Raphson update

$$\delta\beta = A^{-1}g, \text{ as } \delta\beta = \beta^{(k+1)} - \beta^{(k)} \text{ and } \beta^{(k+1)} = \beta^{(k)} + A^{-1}g,$$

$$A^{-1}g = \beta^{(k+1)} - \beta^{(k)} = \delta\beta$$

$$g = A\delta\beta, \text{ but } g = \frac{\partial l}{\partial\beta} = \left\{\frac{\partial l}{\partial\theta}\frac{d\theta}{d\mu}\frac{d\mu}{d\eta}\frac{\partial\eta}{\partial\beta_j}\right\},$$

$$l = {\{y\theta - b(\theta)\}}\big/{a(\phi)} + c(y,\phi) \quad b^{'}(\theta) = \mu \text{ and } b^{''}(\theta) = \nu,$$

$$\therefore \frac{d\mu}{d\theta} = \nu \Rightarrow \frac{d\theta}{d\mu} = \frac{1}{\nu}, \ \eta = \sum\beta_j x_j \Rightarrow \frac{\partial\eta}{\partial\beta_j} = x_j$$

Hence,

$$\frac{\partial l}{\partial\beta_j} = \left[\left(\frac{(y-\mu)}{a(\phi)}\right)\frac{1}{v}\frac{d\mu}{d\eta}x_j\right] = \frac{W}{a(\phi)}(y-\mu)\frac{d\eta}{d\mu}x_j$$

For a single observation with constant dispersion, $\alpha(\phi)$ disappears,

$$g = W(y-\mu)\frac{d\eta}{d\mu}x_j$$

Taking all the observations together,

$$g = \sum W_i(y_i - \mu)\frac{d\eta}{d\mu}x_i = A\delta\beta$$

$x_i = x_{ij}$ (summation over all n individual observations). The components of g are

$$g_r = \sum W(y-\mu)\frac{d\eta}{d\mu}x_r, \ A_{rs} = -E\frac{\partial g_r}{\partial\beta_s}$$

$$= -E\sum\left[(y-\mu)\frac{\partial}{\partial\beta_s}\left\{W\frac{d\eta}{d\mu}x_r\right\} + W\frac{d\eta}{d\mu}x_r\frac{\partial}{\partial\beta_s}(y-\mu)\right],$$

$$= \sum W\frac{d\eta}{d\mu}x_r\frac{\partial\mu}{\partial\beta_s} = \sum Wx_r x_s$$

The new estimate $\beta^{(k+1)} = \beta^{(k)} + \delta\beta = \beta^{(k)} + A^{-1}$ g, $A\beta^{(k+1)} = A\beta^{(k)} + A\delta\beta = A\beta^{(k)} + g$, The component, β_r of β is given as:

$$(A\beta)_r = \sum_S A_{rs}\beta_s = \Sigma w_i x_r \eta_i$$

and the adjusted component based on *z* is given as

$$(A\beta^*)_r = \sum Wx_r z = \sum Wx_r\left\{\eta + (y-\mu)\frac{d\eta}{d\mu}\right\} = X'WZ$$

but

$$A\beta^* = A\beta + g, X^{'}WZ = A\beta + g, A\beta = X^{'}WZ - g$$

$$\beta = A^{-1}\left(X^{'}WZ - g\right) = A^{-1}\left(X^{'}WZ\right) - A^{-1}g = A^{-1}\left(X^{1}WZ\right) - \delta\beta$$

$$\beta + \delta\beta = A^{-1}\left(X^{1}WZ\right)$$

Hence,

$$\beta^{(k+1)} = A^{-1}\left(X^{1}WZ\right) = \left(X^{1}WX\right)^{-1}X^{1}WZ$$

which is the Iterative Weighted Least Squares update. Thus we have shown that the Fisher's scoring algorithm is the same as the iterative Weighted Least Squares algorithm.

RESULTS AND DISCUSSION

The exploration of alternative estimation schemes in generalized linear models arises from the complexities associated with the computation of the Hessian matrix in the Newton-Raphson method. Each member of the Hessian matrix involves a weight matrix, both partial and ordinary differential operators and the systematic component of the model. The quasi-Newton methods avoid the direct use of the Hessian matrix by considering its expected value. The prove that both methods are equivalent rests on the fact that the expected value of the Hessian matrix, E (H) = - (X^1WX) and that the gradient vector g is a product of the Hessian matrix and the discrepancy between current and previous quasi-Newton's updates. The loglikelihood function for a binary response variable has been used to first establish that the expected Hessian matrix used in Fisher's Scoring method is actually the Fisher's information matrix, X^1WX used in the Iterative Weighted Least Squares method. The gradient vector or score function is a weighted differential operator of the systematic component of the model. Computational ease remains the guiding factor in the choice of either of the method.

CONCLUSION

Parameter estimates of generalized linear models can be obtained using the Fisher's Scoring method or the Iterative Weighted Least Squares method. The Fisher's Scoring method uses the gradient vector while the Iterative Weighted Least Squares method uses the adjusted dependent variate. These differences not withstanding, bo th method yield the same solutions. The ease of computation of the gradient vector g and the adjusted dependent variate become the deciding factor as to which method to adopt in any given situation.

ACKNOWLEDGEMENTS

I acknowledge the contributions of Prof. T.A. Bamiduro, who introduced me to Generalized Linear Models. The research of R.W.M. Wedderburn on Quasi-likelihood functions and P. McCullagh and J.A. Nelder on Generalized Linear Models have been very useful in the development of this study.

REFERENCES

Allen, O.B., 1987. Marginal likelihood methods for estimating variance parameters. University of Guelphe, Department of Mathematics and Statistics, Statistical Series, 1986-183 (revised). DOI: 10-1080/ 03610918608812501.

McCullagh, P. and J.A. Nelder, 1992. Generalized Linear Models. Chapman and Hall Madras. 2nd Edn. pp: 40-43. ISBN: 0 412 31760 5.

Silvey, S.D., 1970. Statistical inference. Penguin Books Ltd, Harmondsworth, Middlesex, England. 1st Edn. ISBN: 13:97801-40800975. DOI: 10.1109/34.67641. PMID: 1461-0248200200374.

Smyth, G.K., 2002. Optimization, Encyclopedia of Environmetrics. John Wiley and Sons Ltd. Chichester. DOI: 101002/0470012234.

Scott, C. and R. Nowak, 2006. http://cnx.edu|content|in 11446|latest|. DOI: 10.1007/s10994-007-0717-6. http://ida.firstfraunhofer.de/blandchard/publi/index.html.

Stokes, M.E., C.S. Davis and G.G. Koch, 1975. Categorical Data analysis using the SAS system. SAS Institute Inc., Cary, NC, USA. 1st Edn., pp: 165-173. ISBN: 1-55544-2196.

Wedderburn, R.W.M., 1974. Quasi-likelihood functions, generalized linear models and the Gauss-Newton method. Biometrika, 61 (3): 439. DOI: 10.1214/00905-3605000000057. www.urban.org/url.cfm.