GLOBAL JOURNAL OF MATHEMATICAL SCIENNCES VOL. 7, NO. 2, 2008: 115 – 124

PERFORMANCE ASSESSMENT OF ESTIMATION METHODS IN GENERALIZED LINEAR MODELS

M. E. NJA

(Received 27 February 2008. Revision Accepted 28 November 2008)

ABSTRACT

The Newton-Raphson and the Fisher's Scoring methods are two principal estimation methods in generalized linear models. The Fisher's Scoring method is derived from the Newton-Raphson method by replacing the Hessian matrix with its expected value. Most of the work done on performance of methods have focused on the convergence criterion only. Based on a proposed unified regression model and the logit link function, the relative performance of these methods are assessed with respect to convergence, efficiency and goodness of fit. Iteration history, variances of parameter estimates, the deviance and the Pearson chi-square are used as criteria of assessment. Using appropriate data ser, the claim that Newton-Raphson method converges faster than the Fisher's Scoring method is established. It is also found that both methods enjoy the same level of goodness of fit.

KEY WORDS: Generalized Linear Models, Newton-Raphson and Fisher's Methods, Parameter Estimation.

1.0 INTRODUCTION

The generalized linear model is a nonlinear model (Fox, 1997) and an extension of the general linear model such that each component of the response variable y has a distribution in the exponential family and linearization is achieved through a monotonic link function (McCullagh and Nelder, 1992). The model has extensions like the generalized additive models (Hastie and Tibshirani, 1990) generalized additive partially linear models (Rigby and Stasinoponhes, 2003), generalized linear mixed models (Schall, 1991; Larsen, et al, 2000). The maximum likelihood is the principal method of estimation in all generalized linear models (McCullagh and Nelder, 1992). This is the principle upon which the Newton-Raphson and the Fisher's scoring methods are based (McCullagh and Nelder, 1992).

Attempts have been made by several authors (McCullagh & Nelder, 1992; Knight, 2000; Schworer and Hovey, 2004) at comparing optimization methods in generalized linear models but till date the relative performance of these methods has not been adequately addressed. This is the motivation for the study.

McCullagh and Nelder (1992) in comparing the Newton-Raphson method and the Iterative weighted least squares method observed that the Fisher's Scoring method is a variant of the Newton-Raphson method. This is because the Newton-Raphson method uses the Hessian matrix while the Fisher's Scoring method uses the expected value of the Hessian Matrix. For this reason, the Fisher's Scoring method is referred to as quasi-Newton method. Their comparison was based on computational ease.

Knight (2000) compared the Newton-Raphson and the Fisher's scoring methods using the convergence criterion (number of iterations). Schworer and Hovey (2004) compared the Newton-Raphson and the Fisher's scoring algorithms with respect to convergence using the cumulative standard normal distribution function as response probability. The inverse cumulative standard normal distribution function is a non-canonical link function and is used in probit analysis

In this work assessment of the performance of optimization methods in generalized linear models is extended beyond the convergence criterion to include efficiency and goodness of fit of methods. The statistical methodology is based on a proposed unified regression model given as

$$z_i = \sum_{j}^{p} x_{ij} \beta_j^{\lambda} + \sum_{k=1}^{q} u_{ik} b_k + eh(\mu), \qquad i = 1, 2, \ldots, n \tag{1.1}$$

where h (μ) is a differentiable link function and λ a fixed constant.

This is a model that unifies the general linear model, general linear mixed model, generalized linear model, generalized linear mixed model and the non linear model.

The model reduces to five special case as follows:

1. general linear model:

M. E. Nja, Department of Mathematics/Statistics, Cross River University of Technology, Calabar, Nigeria

$$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i \qquad (1.1a)$$

when $\lambda = 1, \; u_{ik} = 0, \; h'(\mu) = 1, \; z = y$

2. generalized linear model:

$$z = \sum_{j}^{p} \beta_j x_{ij} + e h'(\mu) \qquad (1.1b)$$

when $\lambda = 1, \quad u_{ik} = 0$

3. general linear mixed model:

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{k=1}^{q} u_{ik} b_k + e_i \qquad (1.1c)$$

when $\lambda = 1, \; h'(\mu) = 1, \quad z_i = y_i$

4. generalized linear mixed model:

$$z_i = \sum_{j=1}^{p} x_{ij}\beta_j + \sum_{k=1}^{q} u_{ik} b_k + e_i h'(\mu_i) \qquad (1.1d)$$

when $\lambda = 1$

5. non-linear model

$$y_i = \sum_{j=1}^{p} x_{ij}\beta_j^{\lambda} + e_i \qquad (1.1e)$$

when $\lambda \neq 0,1, \; h'(\mu) = 1, \; u_{ij} = 0, \; z_i = y_i$

It is verified that (1.1) is a unified regression model as follows:
Proof. It is enough to show that (1.1a) and (1.1b) are equivalent.
In general linear modeling, the link function, $h(\mu)$ is an identify function (Mcullagh and Nelder, 1992)
Thus $h'(\mu) = 1$

Let $$y_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i \quad \text{(general linear model)} \qquad (1.1f)$$

$$z_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i h'(\mu_i) \quad \text{(generalized linear model)}$$

$$= \sum_{j=1}^{p} \beta_j x_{ij} + (y_i - \mu_i) h'(\mu_i)$$

$$= \sum_{j=1}^{p} \beta_j x_{ij} + \left(\sum_{j=1}^{p} \beta_j x_{ij} + e_i - \mu_i \right) h'(\mu_i) \quad \text{from (1.1f)}$$

$$= \sum_{j=1}^{p} \beta_j x_{ij} + e_i h'(\mu_i)$$

since $$\sum \beta_j x_{ij} = \mu_i = E(y_i)$$

$$\therefore z_i = \sum_{j=1}^{p} \beta_j x_{ij} + e_i$$

A numerical example on a study on coronary artery disease (Koch et al 1985) as given by Stokes et al (1995) is used to demonstrate the performance of these methods. Both the GENMOD and CATMOD procedures of the SAS software (Stokes et al, 1995) have been used to obtain the solution of the methods. These are then assessed based on convergence, efficiency and goodness of fit.

2.0 ESTIMATION METHODS

Two estimation methods in generalized linear models are considered as follows:

2.1 Newton-Raphson Method

The Newton Raphson optimization method (Silvey, 1970) arises from the application of Taylor's theorem on the system of equations:

$$\nabla_\beta \, l(y;\beta) = 0 \qquad (1.2)$$

where l (y;β), = log f (y; β), is the log likelihood function. ∇_β is the vector differential operator whose ith component is $\frac{\partial}{\partial \beta_i}$

The expansion by Taylor's theorem of (1.2) yields

$$O = \nabla_\beta l(y;\beta) \approx \nabla_\beta l(y, \beta^{(0)}) + \{\nabla^2_\beta l(y;\beta^{(0)})\} \; (\hat{\beta} + \beta^{(0)}) \qquad (1.3)$$

where ∇^2_β is the matrix operator whose (r, s) element is

$$\left(\frac{\partial^2}{\partial \beta_r \, \partial \beta_s}\right)$$

From (1.2)

$$\hat{\beta} \approx \beta^{(0)} - \{\nabla^2_\beta l(y;\beta^{(0)})\}^{-1} \nabla_\beta l(y;\beta^{(0)})$$

$$\hat{\beta}^2 \approx \beta^{(1)} - \{\nabla^2_\beta l(y;\beta^{(1)})\}^{-1} \nabla_\beta l(y;\beta^{(1)})$$

$\vdots$

$$\beta^{(k)} \cong \beta^{(k-1)} - \left(\nabla^2_\beta l(y;\beta^{(k-1)})\right)^{-1} \nabla_\beta l(y;\beta^{(k-1)}) \qquad (1.4)$$

(1.4) is referred to as Newton – Raphson iterative scheme (Silvey, 1970). $\nabla^2_\beta l(y;\beta^{(k-1)})$ is called Hessian (second derivative) matrix (McCullagh and Nelder, 1992). Convergence is established if $\beta^{(k)} \approx \beta^{(k-1)}$

If the initial approximation, $\beta^{(0)}$ is good, $\nabla^2_\beta l(y;\beta^{(0)})$ *and* $\nabla^2_\beta l(y;\beta^{(1)})$ will be very close, so that we can use the initial Hessian matrix for all the iterations. This is a modification of Newton-Raphson method (Silvey, 1970).

The gradient vector, $\nabla_\beta l(y;\beta^{(0)})$ is based on the design matrix and is given as (McCullagh and Nelder, 1992)

$$\nabla_\beta l(y;\beta^{(0)}) = \sum \frac{y_i - m_i\mu_i}{\mu_i(i-\mu_i)} \frac{d\mu_i}{d\eta_i} \qquad (1.5)$$

where the log likelihood function, $l(y;\beta)$ is given as

$$l(y_i:\beta) = l((y_i:\mu(\beta)))$$

$$= \sum \left[\log\binom{m_i}{y_i} + y_i \log \mu_i + (m_i - y_i)\log(1-\mu_i)\right]$$

$$= \sum \log\binom{m_i}{y_i} + \sum \left[y_i \log \mu_i + y_i \log(1-\mu_i) + m_i \log(1-\mu_i)\right] \quad (1.6)$$

2.2 Fisher's Scoring Method

This method makes use of the expected value of the Hessian matrix. The method uses (McCullagh and Nelder, 1992)

(i) the gradient vector $\partial l / \partial \beta = g_k U$ as in Newton-Raphson given in 2.1

(ii) minus the expected Hessian matrix

$$- E\left[\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\right] = A = -E(H) \quad (1.7)$$

Let β = current estimate of β

δb = an adjustment

$A\delta\hat{\beta} = U \Rightarrow \delta\hat{\beta} = [-E(H)]^{-1}g$

$= -[E(H)]^{-1}g$

The components of U (omitting the dispersion factor) are (McCullagh and Nelder, 1992)

$$g_r = \sum_{i=1}^{n} w_i (y - \mu) \frac{d\eta_i}{d\mu_i} x_r$$

$$\frac{\partial g}{\partial \beta} = \frac{\partial^2 l}{\partial \beta^2} = \left\{\frac{\partial^2 l}{\partial \beta r \partial \beta s}\right\}$$

$$\frac{\partial g_r}{\partial \beta_s} = \frac{\partial^2 l}{\partial \beta_r \partial \beta_s}$$

$$a_{rs} = -E\left(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\right)$$

$$= -\left[E\Sigma\left[\left((y_i - \mu_i)\frac{\partial}{\partial \beta_s}\left(w\frac{d\eta}{d\mu}xr\right)\right) + w\frac{d\eta}{d\mu}x_r\frac{\partial}{\partial \beta_s}(y_i - \mu_i)\right]\right]$$

$$= \sum^{n} w \frac{d\eta}{d\mu} x_r \frac{\partial \mu}{\partial \beta_s} = \Sigma w_i x_r x_s$$

Let $\hat{\beta}^{(k)}$ be estimate of parameter vector β at iteration m

$$\text{Then } \beta^{(k+1)} = \beta^{(k)} + \left(\delta\hat{\beta}\right)^{(k)}$$

$$= \hat{\beta}^{(k)} + \left(A^{(k)}\right)^{-1} g^{(k)}$$

$$= \beta^{(k)} + [-E(H)]^{-1} g^{(k)} \quad (1.8)$$

3.0 ASSESSMENT CRITERIA

Model – fit statistics were used to determine goodness of fit of the method. These are the deviance (Likelihood Ratio) and Pearson χ^2, of parameter estimates variances were used to assess the efficiency of each method. Convergence of methods was determined by iteration history (McCullagh and Nelder, 1992).

3.1 Efficiency

For a real parameter θ, Cramer-Rao inequality is stated (Silvey, 1970) as

$$\operatorname{var}\left(\beta\right) \geq \frac{1}{E\left[\left(\frac{\partial l}{\partial \beta}\right)^2\right]} \tag{1.9}$$

The quantity, $E\left[\left(\frac{\partial l}{\partial \beta}\right)^2\right]$ denoted by I_β, was called by Fisher, the amount of information about β contained in an observation of x (Silvey, 1970).

Thus, $\operatorname{var}\left(\hat{\beta}\right) \geq I_\beta^{-1}$

The efficiency of an estimator $\hat{\beta}$ is then defined (Silvey, 1970) as

$$eff\left(\hat{\beta}\right) = \frac{I_\beta^{-1}}{\operatorname{var}\left(\beta\right)} = \frac{1}{I_\beta \operatorname{var}\left(\hat{\beta}\right)} \leq 1$$

$$eff\left(\beta\right) = 1 \Rightarrow \text{we are using min } \operatorname{var}\left(\hat{\beta}\right).$$

Generally, the Cramer-Rao inequality can be generalized by considering a vector-valued parameter β $\left(i.e. \beta \in IR^s\right)$. In that case, the Fisher's information is replaced by an information **Matrix**, $B_\beta(sxs)$.

The $(i,j)th$ component of B_β is

$$E\left[\frac{\partial l}{\partial \beta_i}\frac{\partial l}{\partial \beta_j}\right] = -E\left[\frac{\partial^2 l}{\partial \beta_i \partial \beta_j}\right]$$

The diagonal elements of $B_\beta(i=j)$ are

$$-E\left[\frac{\partial^2 l}{\partial \beta_i^2}\right] = I_\beta$$

The variance matrix of β_β, denoted by $\operatorname{var}\left(\hat{\beta}\right)$ is

$$\operatorname{var}\left(\hat{\beta}\right) \geq \frac{1}{B_\beta} = B_\beta^{-1}$$

$\therefore$ B_β^{-1} is a "lower bound" for the variance matrix of an unbiased estimator of β.

$$\operatorname{var}\left(\hat{\beta}\right) \geq B_\beta^{-1} \Rightarrow \operatorname{var}\left(\hat{\beta}\right) - B_\beta^{-1} \geq 0$$

i.e. $\operatorname{var}\left(\beta\right) - B_\beta^{-1}$ is positive semi-definite.

3.2 The Deviance

The deviance can be obtained from the log-likelihood function. The **deviance assumes a minimum** value at the point b which minimizes the discrepancy of fit (McCullagh and Nelder, 1992).
The discrepancy of fit D is given as

$$D = (Y - Xb)'(Y - X\beta)$$
$$= (Y - Xb)'(Y - Xb) + (Y - Xb)'(Xb - X\beta)$$
$$= (Y - Xb)'(Y - Xb) + (Y - Xb)'X(b - \beta)$$

When b = β, the discrepancy D is denoted by D* and given as

$D^* = (Y - Xb)'(Y - Xb)$ D* is deviance of data from arbitrary point b.

The deviance is a measure of discrepancy between an observation and the solution locus (McCullagh and Nelder, 1992). It assumes a minimum value at the point which minimizes the discrepancy of fit. The scaled deviance is define as:

Let $\hat{\pi}$ = fitted probabilities

McCullagh and Nelder, (1992) obtained the loglikelihood as

$$l\left(\hat{\pi};y\right)=\sum\left\{y_i \log \mu_i+\left(m_i-y_i\right)\log\left(1-\pi_i\right)\right\}$$

= likelihood achievable for the model under investigation.

where y is the vector of observations, m_i is total number of observations in the sub group.

$l\left(y:y\right)$ = maximum likelihood achievable.

$l\left(y:y\right)$ is attained at the point $\overline{\pi}=\dfrac{y_i}{m_i}$ because $\dfrac{y_i}{m_i}$ is the observed proportion of success. It is estimated by the fitted probability $\hat{\pi}$.

$$2l(\hat{\pi};y)=2\sum\{y_i\log\hat{\pi}_i+(m_i-y_i)\log(1-\hat{\pi}_i)\}$$

$$2l(\overline{\pi};y)=2l\left(\frac{y}{m};y\right)$$

$$=2\sum\left\{y_i\log\frac{y_i}{m_i}+(m_i-y_i)\log\left(1-\frac{y_i}{m_i}\right)\right\}$$

$$D(y;\overline{\pi})=2l(\overline{\pi};y)-2l(\hat{\pi};y)$$

$$=2\sum\left\{y_i\log\frac{y_i}{\hat{\pi}_i}+(m-y_i)\left(\log\frac{m_i-y_i}{m_i-\hat{\pi}_i}\right)\right\}$$

4.0 P – VALUE

The p – value is the smallest level of significance at which the null hypothesis can be rejected. For the standard normal variate z, p – value is defined as follows:

P – value = prob (z < - z_{cal}) = prob (z > + z_{cal}) for a one-tailed test.

For a two-tailed test

P – value = prob (z < - z_{cal}) = 0.5 - z_{cal}^{-1} (1.10)

where z_{cal}^{-1} is the critical value of z_{cal} and z_{cal} is the calculated value of the standard normal variate corresponding to the null hypothesis being tested.

5.0 THE MODEL

The generalized linear model is used for estimation in a categorical data setting.

Let

$$y_{ij}=\beta_0+Sex(i)+ecg(j)+e_{ij} \qquad (1.11)$$

$$E\left(y_{ij}\right)=\beta_0+Sex(i)+ecg(j)$$

where $i=1,\ldots..n$, $j=1,\ldots..p$

$$\mu_{ij}=P_r\left\{y_{ij}\right\}=\frac{\exp[\beta_0+Sex(i)+ecg(j)]}{1+\exp[\beta_0+sex(i)+ecg(j)]}=\frac{\exp(\beta_0+\beta_1 sex+\beta_2 ecg)}{1+\exp(\beta_0+\beta_1 sex+\beta_2 ecg)}$$

y_{ij} = Number of people from the ith sex level and the jth ecg status who have coronary artery disease.

μ = overall mean

β_0, β_1, β_2 are model parameters

Sex *(i)* = effect of *ith* level of sex

ecg (j) = effect of *jth* ecg status

e_{ij} = random error associated with observation y_{ij}

μ, sex *(i)* and ecg *(j)* are fixed effects.

Equation (1.11) can be written as $y_i = \sum_{j=0}^{k} \beta_j x_{ij} + e_i$

where x_{ij} = value of the explanatory valuable corresponding to the ith sex level and jth ecg status. In terms of the proposed unified regression model (Nja, 2007) the generalized linear model is given as $z_i = \sum_{i=1}^{k} \beta_i x_{ij} + eh^1(\mu)$ where h(μ) is the link function.

x_{ij} is an element of the fixed effects design matrix corresponding to the ith row and jth column. U_{ik} is an element of the random effects design matrix, corresponding to the ith row and jth column. β_j and b_k are fixed effects and random effects parameter estimates respectively. μ is the mean of the observations. e is the random error term.

The probability that a person from the ith sex level and the jth ecg status has coronary artery disease is modeled, using the Newton-Raphson, equ (1.4) and the Fisher's Scoring method, equ. (1.8). Also modeled is the logit of this probability.

$$\text{Logit}\,(P\{y_{ij}\}) = \sum x_{ij} \beta_j$$

The logit link function is given as

$$\text{Logit}\,(P\{y_{ij}\}) = \log \frac{p(y_{ij})}{1 - p(y_{ij})}$$

6.0 NUMERICAL EXAMPLE

The following data is based on a study on coronary artery disease (Koch et al, 1985). It is required to model the probability of coronary artery disease, $P(y_{ij})$ by computing the parameter estimates β_0, β_1, and β_2 of the model using the methods under study. The methods of obtaining these estimates are then assessed based on the criteria under investigation.

Table 1: Data on coronary artery disease

I	Sex x_1	Ecg x_2	disease y_i	No disease	Total m_i
1	Female	< 0.1 ST segment depression	4	11	15
2	Female	≥ 0.1 ST segment depression	8	10	18
3	Male	< 0.1 ST segment depression	9	9	18
4	Male	≥ 0.1 ST segment depression	21	6	27

Ecg = electrocardiogram
ST segment depression is used to categorise Ecg.

6.1 RESULTS

These results have been obtained using the SAS statistical software. The results and performance assessment are shown below:

6.2 PARAMETER ESTIMATES

The parameter estimates using the logit link function (1.12) and the values for the methods are given in the following tables.

Newton-Raphson Method

Table 2: Estimates using Newton-Raphson Method

Parameter	Estimates	P-vales
β_0	1.1568	0.0042
β_1	-1.2770	0.0103
β_2	-1.0545	0.0342

Fisher's Method of Scoring

Table 3: Estimates using Fisher's Method

Parameter	Estimates	P-vales
β_0	-1.1747	0.0155
β_1	1.2770	0.0103
β_2	1.0545	0.0342

6.3 CONVERGENCE

Iteration history is used to determine the rate of convergence of the methods. The method that has a fewer number of iterations is said to converge faster. The Iterations are given tables 4 and 5.

Table 4: Iteration History for Newton-Raphson Method

Iteration	-2logL	Intercept (β_0)	Sex (β_1)	Ecg (β_2)
0	95.89973	1.1535088	-1.272435	-1.050579
1	95.89959	1.1567728	-1.276951	-1.054495
2	95.89959	1.1567765	-1.276955	-1.0545

Fisher's Method of Scoring

Table 5: Iteration History for Fisher's Method

Iteration	-2logL	Intercept (β_0)	Sex (β_1)	Ecg (β_2)
0	107.668965	0.154151	0	0
1	95.992676	-1.064377	1.167830	0.944285
2	95.899664	-1.171724	1.274025	1.051569
3	95.899598	-1.174676	1.276953	1.054497

6.4 GOODNESS OF FIT

Goodness of fit is judged by the values of the deviance and Pearson x^2 shown in tables 6 and 7 below.

Table 6: Model Fit Statistics for Newton-Raphson Method

Deviance	0.2141
Pearson X^2	0.2155

Model Fit Statistics (Fisher's Method)

Table 7: Model Fit Statistics for Fisher's Method

Deviance	0.2141
Pearson X^2	0.2155

Efficiency

We shall use variances of parameter estimates as provided in (1.9) to measure efficiency of methods. A smaller variance yields a higher efficiency. The data shows that there is no significant difference in efficiency between the two methods.

Table 8: Variances of Parameter Estimates

Optimization Method	Parameter	Parameter Estimates	Variance
	β_0	1.1568	0.1629
Newton - Raphson	β_1	-1.2770	0.2480
	β_2	-1.0545	0.2480
	β_0	-1.1747	0.2356
Fisher's	β_1	1.2770	0.2480
Method	β_2	1.0545	0.2480

7.0 DISCUSSION

Using the illustrative example, the values of the parameter estimates, β_0, β_1, β_2. Solutions of generalized linear models have been obtained using appropriate estimation methods. These methods are based on the maximum likelihood principles which employ the Newton-Raphson Scheme as described in section 1.0. Due to the limitations of the Newton-Raphson method which include the hectic computation of the Hessian matrix, an alternative quasi-Newton scheme was also used for purposes of assessing relative performance of the methods. This is the Fisher's scoring method. The logit link is applied to both methods and the solutions of the methods are given as:

Newton-Raphson: $\beta^{(k+1)} = \beta^{(k)} - \left\{ \frac{\partial^2 l}{\partial \beta_r \partial \beta_s} \right\}_g^{-1}$

$\beta_0 = 1.1568, \quad \beta_1 = -1.2770, \quad \beta_2 = -1.0545$

Fisher's Scoring Method: $\beta^{(k+1)} = \beta^{(k)} - \left[-(H^{(k)}) \right]_g^{-1}$

$\beta_0 = 1.1747, \quad \beta_1 = 1.2770, \quad \beta_2 = 1.0545$

Iteration history has been used to determine rates of convergence. Newton-Raphson method converged at the second iteration, Fisher's scoring method converged at the third. (Tables 4 and 5) because subsequent iterations do not yield significantly different results.

For efficiency, the variances of the fixed effects estimates (Table 8) show that there is no significant difference between the two methods. The deviance and Pearson χ^2, criteria were used as goodness of fit statistics. Both methods exhibited the same level of goodness of fit.

8.0 CONCLUSION

The Fisher's Scoring estimation method in generalized linear models is recommended against the Newton-Raphson method if efficiency and goodness of fit are the criteria under consideration. This is because the method uses the expected value of the Hessian matrix, a condition which renders it computationally less hectic than the Newton-Raphson method. However, where convergence is the criterion of interest, the Newton-Raphson method is preferred. This is because it converges faster than the Fisher's Scoring method as shown by the numerical illustration. This establishes the existing claim that Newton-Raphson method converges faster the Fisher's Scoring method.

REFERENCES

Fox, J., 1997. Nonparametric Regression Analysis: An Introduction, Memaster University.

Hastie, T. and Tibshirani, R., 1990. Generalized Additive" Models. Chapman and Hall, London.

Knight, K., 2000. Mathematical Statistics. Chapman & Hall/CRC, Boca Raton.

Koch, G. G., Imrey, P. B., Singer, J. M., Atkinson, S. S. and Stokes, M. E., 1985. Analysis of Categorical Data. Montreal, Canada: Les presses De L'Universite de Montreal.

Larsen, K; Peterson, J. H., Budtz-Jorgensen, E. and Endahl, L., 2000. Interpreting Parameters in the Logistic Regression Model with Random Effects. "Biometrics". 56: 909-914.

McCullagh, P. and Nelder, J. A., 1992. Generalized Linear Models. Chapman and Hall. Madras.

Nja, M. E., 2007. Relative Performance of Optimization methods in solutions of generalized linear models. (Unpublished Ph.D work, Dept. of Statistics, University of Ibadan, Nigeria.

Rigby, R. A., Stasinoponhes, D. M., 2003. Means and Dispersion Additive Models": University of North London.

Schall, R., 1991. Estimation in Generalized Linear Models with Random Effects Biometrika 78, 4, 719-27 Gr.Britain.

Schworer, A. and Hovey, P., 2004. Newton-Raphson versus Fisher Scoring Algorithms in Calculating Maximum Likelihood Estimates. Electronic Proceedings of Undergraduate Mathematics Day. No. 1 http://www.udayton.edu/~mathdept/EPUMD/EPMUD.htm University of Dayton Research Institute, Dayton, OH 45409.

Silvey, S. D., 1970. Statistical inference. Penguin Books Ltd, Harmondsworth. Middlesex, England.

Stokes, M. E., Davis. C. S. and Koch, G. G., 1995. Categorical Data Analysis using the SAS system. SAS Institute Inc. Cary, NC, USA.