Biometrical Letters
Vol. 48 (2011), No. 1, 23-32

A note on solving the likelihood equation in logistic model with the multinomial distribution

Ewa Bakinowska

Department of Mathematical and Statistical Methods, Poznań University of Life Sciences, Wojska Polskiego 28, 60-637 Poznań, e-mail: ewabak@up.poznan.pl

Summary

In this paper the maximum likelihood equation of a logistic model with multinomial distribution is presented. Two iterative methods, the Fisher scoring method and the Newton-Raphson method are described. The Hessian matrix of the likelihood function in a logistic model with multinomial distribution is derived.

Key words: logistic model, Fisher scoring method, Newton-Raphson method.

1. Introduction

Categorical response data are usually modelled using the multinomial distribution. If the expected value of the random variable of this distribution is mapped by the logistic transformation, we get a model which belongs to the class of generalized linear models. The unknown parameters of such models can be estimated by adopting various approaches. One of them is the maximum likelihood method, leading to a set of equations which are usually non-linear. The solution can be searched for using iterative methods such as the Fisher scoring method or the Newton-Raphson method. The last method uses the Hessian matrix of the likelihood function. In this paper an explicit derivation of the Hessian matrix of the likelihood function in a logistic model with multinomial distribution is presented.

2. Generalized linear model

Let us assume that $\mathbf{y}$ is the random vector observed in the experiment, and the expected value of $\mathbf{y}$ is μ,

$$E(\mathbf{y}) = \mu.$$

Moreover, let $\eta = \eta(\mu)$ be the fixed transformation of the expected value vector. If the image of this transformation belongs to the fixed linear subspace then we obtain the generalized linear model (see McCullagh and Nelder, 1983, 1989) as follows

$$\eta(\mu) = \mathbf{X}\beta. \tag{1}$$

If the transformation $\mu \to \eta$ is the identity then the model (1) is reduced to the standard linear model. The transformation $\eta(\mu)$ is called the link function. Matrix $\mathbf{X}$ is the known matrix of covariates and β is the vector of unknown parameters.

The main aim in such a model (1) is to find the expected value μ as the functions of covariates, which leads to estimation of the vector of unknown parameters β.

The estimation can be carried out using various methods: geometric (least squares method, weighted least squares method), or the maximum likelihood method.

3. Iterative methods

If the distribution of the random variable observed in the experiment is known (except for some parameters, which have to be estimated), the vector β can be estimated using the likelihood method provided that the transformation $\mu \to \eta$ is invertible. The last assumption makes it possible to determine the gradient $\mathbf{s}$ of the logarithm of the likelihood function $l = l(\mu, \mathbf{y})$ with respect to β,

$$\mathbf{s}^T(\beta) = \frac{\partial l}{\partial \beta^T} = \frac{\partial l}{\partial \mu^T} \left(\frac{\partial \eta}{\partial \mu^T} \right)^{-1} \frac{\partial \eta}{\partial \beta^T}, \tag{2}$$

(see Bakinowska, 2004). Putting the gradient (2) equal to the zero vector provides the maximum likelihood equation which takes the form

$$\mathbf{s} = \mathbf{0}. \tag{3}$$

Usually it is a non-linear equation. Its solution can be found using iterative methods: the Newton-Raphson method or the Fisher scoring method. In the Newton-Raphson method the consecutive approximations of the sought solution are defined by the following formula

$$\beta_{n+1} = \beta_n - \mathbf{H}^{-1}\mathbf{s}(\beta_n), \tag{4}$$

with $\mathbf{H} = \partial \mathbf{s}^T/\partial\beta$ being the matrix of second derivatives of the likelihood function. This matrix is called the Hessian matrix of the likelihood function and in (4) it is established for the approximation β_n (see Agresti, 1984, Appendix B).

The Fisher scoring method consists in replacing the Hessian matrix in (4) by its expectation,which leads to the Fisher information matrix $\mathbf{F}$

$$\mathbf{F} = -E\left(\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right) = -E\left(\frac{\partial \mathbf{s}^T}{\partial\beta}\right) = -E(\mathbf{H}), \tag{5}$$

(see Mardia et. al., 1979, p. 98 and Krzyśko 2000, p. 48). As a result the consecutive approximations of the solution are related by the equation

$$\beta_{n+1} = \beta_n + \mathbf{F}^{-1}\mathbf{s}.$$

This approach is known in the literature as the Fisher scoring method (see McCulloch and Searle 2001, p. 143)

4. Logistic transformation

In general the logistic transformation can be written in the following form

$$\eta(\mu) = \mathbf{C}^T \log\left(\mathbf{L}\mu\right), \tag{6}$$

where matrix $\mathbf{C}^T$ is usually the fixed matrix of contrasts or identity matrix, $\mathbf{L}$ is a fixed binary matrix such that in the product $\mathbf{L}\mu$ it forms the selected sums of elements of vector μ, and $\log(\cdot)$ is the vector of natural logarithms (compare Bakinowska and Kala, 2002a). The constructions of matrices $\mathbf{L}$ and $\mathbf{C}^T$ for discrete variables are given by Glonek and McCullagh (1995). The generalized linear model with link function (6) is called the logistic model.

The logistic transformation (6) is continuous and differentiable. The matrix of the partial derivatives $\mathbf{G} = \partial\eta/\partial\mu^T$ (compare Grizzle *et al.*, 1969) is determined by the formula

$$\mathbf{G} = \mathbf{G}\left(\mu\right) = \mathbf{C}^T\mathbf{D}^{-1}\mathbf{L},$$

with $\mathbf{D} = (\mathbf{L}\mu)^{\delta}$ being the diagonal matrix, having elements of the vector $\mathbf{L}\mu$ on its diagonal. If $\mathbf{C}^T$ in (6) is the matrix of contrasts, namely if this matrix fulfils the condition $\mathbf{C}^T\mathbf{1} = \mathbf{0}$, then

$$\mathbf{G}(\mu)\,\mu = \mathbf{0}. \tag{7}$$

5. Logistic model with multinomial distribution

Let us assume that the m units are characterized by the joint value $\mathbf{x}$ of some covariate. Now we assume that the units are classified in k separable categories. The results of counting the units in categories form a k-dimensional random vector $\mathbf{y} = (y_1, y_2, ..., y_k)^T$ which has multinomial distribution $\mathbf{y} \sim M(m, \pi)$, where m is the fixed number of units being classified, and $\pi = (\pi_1,\pi_2, ..., \pi_k)^T$ is the vector of fixed probabilities fulfilling

$$\sum_{j=1}^{k} \pi_j = 1. \tag{8}$$

The probability that within the sequence of m independent events one can observe y_1 events of the first category, y_2 events of the second category, etc. is expressed by

$$\frac{m!}{y_1!y_2!\cdots y_k!}\pi_1^{y_1}\pi_2^{y_2}\cdots\pi_k^{y_k}.$$

The m units are characterized by a common value $\mathbf{x}$. Hence we can expect that the vector π is the value of some transformation $\pi = \pi(\mathbf{x})$, i.e.

$$\pi = (\pi_1(\mathbf{x}), \pi_2(\mathbf{x}), ..., \pi_k(\mathbf{x}))^T.$$

In order to determine the system of functions $\pi_j = \pi_j(\mathbf{x})$, $j = 1, 2, ..., k$, one can apply the logistic model.

Let $\mathbf{p}$ be the vector of the experimentally observed frequencies, $\mathbf{p} = \mathbf{y}(1/m)$, then $E(\mathbf{p}) = \pi$ (see Fisz, 1958, p. 151 or Mardia *et al.*, 1979, p. 57).

Assuming that the link function is the logistic transformation (6), the model (1) now has the form

$$\eta = \mathbf{C}^T \log(\mathbf{L}\pi) = \mathbf{X}\beta. \tag{9}$$

Mapping the functions of probabilities $\pi_j = \pi_j(\mathbf{x})$, $j = 1, 2, ..., k$ as functions of covariates comes down to estimating the vector of parameters β.

6. Maximum likelihood method

According to the remarks in the previous paragraph, the "shape" of the joint distribution of the random variable observed in the experiment is known. Hence the estimator of the parameter vector β in (9) can be found using the maximum likelihood method. The maximum likelihood equation (3) now has the form

$$\mathbf{s}(\beta) = \frac{\partial l}{\partial \beta} = \frac{\partial \eta^T}{\partial \beta} \left(\frac{\partial \eta^T}{\partial \pi} \right)^{-1} \frac{\partial l}{\partial \pi} = \mathbf{0},$$

where l is the log-likelihood function of the multinomial distribution. We will solve this equation using the iterative method.

In the Newton-Raphson method the consecutive approximations of the searched solution are defined by (4). Now this formula contains the Hessian matrix of the likelihood function (the matrix of second derivatives of the likelihood function) of the logistic model with multinomial distribution. What kind of problems are involved with using this estimation method? We have to notice that in general the logistic transformation $\pi \to \eta$ is not invertible. Hence the matrix of derivatives $\partial \eta^T / \partial \pi$ is not nonsingular, usually the matrix of such derivatives is not a square matrix. The second problem is connected with calculation of the Hessian matrix of the likelihood function in the logistic model (9). The way to solve these problems will be shown in the next paragraph.

7. Hessian matrix of the likelihood function

According to the remarks contained in the paper by Bakinowska and Kala (2002b) the logistic transformation

$$\eta = \mathbf{C}_*^T \log (\mathbf{L}_* \pi) \tag{10}$$

is one-to-one if the matrices $\mathbf{C}_*^T$ and $\mathbf{L}_*$ are of the form

$$\mathbf{C}_*^T = \begin{pmatrix} 1 & \mathbf{0}^T \\ 0 & \mathbf{C}^T \end{pmatrix} \qquad \mathbf{L}_* = \begin{pmatrix} \mathbf{1}^T \\ \mathbf{L} \end{pmatrix}$$

and the matrix

$$\mathbf{G}_* = \mathbf{C}_*^T \mathbf{D}_*^{-1} \mathbf{L}_* = \begin{pmatrix} \mathbf{1}^T \\ \mathbf{G} \end{pmatrix} = \begin{pmatrix} \mathbf{1}^T \\ \mathbf{C}^T \mathbf{D}^{-1} \mathbf{L} \end{pmatrix}$$

with $\mathbf{D}_* = (\mathbf{L}_*\pi)^\delta$, $\mathbf{D} = (\mathbf{L}\pi)^\delta$ being nonsingular matrix (the matrices $\mathbf{C}^T$ and $\mathbf{L}$ above are defined in section 4). The second condition is fulfilled if $\mathbf{C}^T\mathbf{1} = \mathbf{0}$ and simultaneously matrix $\mathbf{G}$ is of full row rank $k-1$. Then $\mathbf{G}\pi = \mathbf{0}$ and we can derive that the inverse of matrix $\mathbf{G}_*$ has the form

$$\mathbf{G}_*^{-1} = \left(\pi, \quad \pi^\delta \mathbf{G}^T \left(\mathbf{G}\pi^\delta \mathbf{G}^T \right)^{-1} \right), \tag{11}$$

where the vector π is the first column of matrix $\mathbf{G}_*^{-1}$, and the columns of matrix $\pi^\delta \mathbf{G}^T \left(\mathbf{G}\pi^\delta \mathbf{G}^T \right)^{-1}$ form the next columns in $\mathbf{G}_*^{-1}$.

For the transformation (10), $\pi \to \eta$, the first element of vector η equals $\log(\mathbf{1}^T\pi)$. Hence putting this element equal to zero is equivalent to the equality $\mathbf{1}^T\pi = 1$, which is required in determining the multinomial distribution (see formula 8). This condition can be directly taken into account in determining the generalized linear model. It is sufficient in the equality $\eta = \mathbf{X}_*\beta$, to assume that the matrix $\mathbf{X}_*$ is of the form

$$\mathbf{X}_* = \begin{pmatrix} \mathbf{0}^T \\ \mathbf{X} \end{pmatrix}.$$

According to the assumptions given above the matrices of derivatives are of the forms

$$\frac{\partial \pi^T}{\partial \eta} = \left(\left(\mathbf{C}_*^T \mathbf{D}_*^{-1} \mathbf{L}_* \right)^{-1} \right)^T = \left(\mathbf{G}_*^{-1} \right)^T \tag{12}$$

and

$$\frac{\partial \eta^T}{\partial \beta} = \mathbf{X}_*^T.$$

Hence the gradient of the likelihood function l with respect to vector β is of the form

$$\mathbf{s}(\beta) = \frac{\partial \eta^T}{\partial \beta} \frac{\partial \pi^T}{\partial \eta} \frac{\partial l}{\partial \pi} = \mathbf{X}_*^T \left(\mathbf{G}_*^{-1} \right)^T \pi^{-\delta} \mathbf{y}. \tag{13}$$

In order to use the Newton-Raphson method it is necessary to calculate the Hessian matrix of l with respect to vector β,

$$\mathbf{H}(\beta) = \frac{\partial \mathbf{s}}{\partial \beta^T} = \frac{\partial \mathbf{s}}{\partial \pi^T} \frac{\partial \pi}{\partial \eta^T} \frac{\partial \eta}{\partial \beta^T}. \tag{14}$$

The essential problem comes down to finding the matrix of derivatives $\partial \mathbf{s}/\partial \pi^T$. Nevertheless, we have to notice that in gradient (13) the matrices

$\left(\mathbf{G}_*^{-1}\right)^T$ and $\pi^{-\delta}$ (see (11), (12)) are dependent on the vector of probabilities π. Hence directly deriving the formula of the Hessian matrix (14) is difficult. However, using the rules of matrix derivatives we can determine the j-th column of the sought matrix . We obtain

$$\frac{\partial \mathbf{s}}{\partial \pi_j} = \mathbf{X}_*^T \left(\frac{\partial \left(\mathbf{G}_*^{-1}\right)}{\partial \pi_j} \right)^T \pi^{-\delta}\mathbf{y} + \mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T \frac{\partial \pi^{-\delta}}{\partial \pi_j}\mathbf{y},$$

where π_j is the single probability. Since for any nonsingular matrix $\mathbf{F} = (f_{ij}(x))$ the equality

$$\frac{\partial \mathbf{F}^{-1}}{\partial x} = -\mathbf{F}^{-1}\frac{\partial \mathbf{F}}{\partial x}\mathbf{F}^{-1}$$

holds, (see Harville, 1997, p. 307), we have

$$\frac{\partial \mathbf{s}}{\partial \pi_j} = -\mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T \frac{\partial \mathbf{G}_*^T}{\partial \pi_j} \left(\mathbf{G}_*^{-1}\right)^T \pi^{-\delta}\mathbf{y} - \mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T \pi^{-\delta}\frac{\partial \pi^{\delta}}{\partial \pi_j}\pi^{-\delta}\mathbf{y}$$

Recalling that matrix $\mathbf{G}_*^T$ is the matrix corresponding to the model (10) we obtain

$$\frac{\partial \mathbf{G}_*^T}{\partial \pi_j} = \left(\mathbf{0}, \frac{\partial \mathbf{L}^T\mathbf{D}^{-1}\mathbf{C}}{\partial \pi_j}\right) = \left(\mathbf{0}, -\mathbf{L}^T\mathbf{D}^{-1}\frac{\partial \left(\mathbf{L}\pi\right)^{\delta}}{\partial \pi_j}\mathbf{D}^{-1}\mathbf{C}\right) =$$

$$= \left(\mathbf{0}, -\mathbf{L}^T\mathbf{D}^{-1}\mathbf{l}_j^{\delta}\mathbf{D}^{-1}\mathbf{C}\right),$$

where $\mathbf{l}_j$ is the j-th column of matrix $\mathbf{L}$. Similarly,

$$\pi^{-\delta}\frac{\partial \pi^{\delta}}{\partial \pi_j}\pi^{-\delta} = \pi^{-\delta}\mathbf{e}_j^{\delta}\pi^{-\delta} = \frac{1}{\pi_j^2}\mathbf{e}_j^{\delta},$$

where $\mathbf{e}_j$ is the j-th column of $\mathbf{I}_k$.

On the other hand using the condition $\mathbf{G}\pi = \mathbf{0}$ we have

$$\left(\mathbf{G}_*^{-1}\right)^T \pi^{-\delta} = \begin{pmatrix} \mathbf{1}^T \\ \left(\mathbf{G}\pi^{\delta}\mathbf{G}^T\right)^{-1}\mathbf{G} \end{pmatrix}.$$

In consequence we obtain the vector of derivatives

$$\frac{\partial \mathbf{s}}{\partial \pi_j} = \mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T \mathbf{L}^T\mathbf{D}^{-1}\mathbf{l}_j^{\delta}\mathbf{D}^{-1}\mathbf{C}\left(\mathbf{G}\pi^{\delta}\mathbf{G}^{\mathbf{T}}\right)^{-\mathbf{1}}\mathbf{G}\mathbf{y} - \tag{15}$$

$$-\mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T \mathbf{e}_j \left(\frac{y_j}{\pi_j^2}\right).$$

Putting together the above obtained vectors for $j = 1, 2, ..., k$, as the columns of the sought matrix we obtain the whole matrix $\partial \mathbf{s}/\partial \pi^T$, which multiplied by the product $\mathbf{G}_*^{-1}\mathbf{X}_*$ gives the Hessian matrix (14).

8. Fisher information matrix

Based on the Hessian matrix (14), according to the formula (5), we can calculate the Fisher information matrix $\mathbf{F}$. We can notice that $E(\mathbf{y}) = m\pi$, and according to the equality $\mathbf{G}\pi = \mathbf{0}$, taking the expected value of the vector (15) the first element equals the zero vector for $j = 1, 2, ..., k$. Also the expected values of the second element in (15) equal

$$-m\mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T \mathbf{e}_j \frac{1}{\pi_j}, \qquad j = 1, 2, ..., k.$$

Hence

$$E\left(\frac{\partial \mathbf{s}}{\partial \pi^T}\right) = -m\mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T \pi^{-\delta}$$

and as a result we obtain the formula of the Fisher information matrix in the logistic model with multinomial distribution:

$$\mathbf{F} = -E\left(\frac{\partial \mathbf{s}}{\partial \pi^T}\right)\mathbf{G}_*^{-1}\mathbf{X}_* = m\mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T \pi^{-\delta}\mathbf{G}_*^{-1}\mathbf{X}_*. \tag{16}$$

It is easy to derive that the product $\mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T$ in (16) can be written in the following form

$$\mathbf{X}_*^T \left(\mathbf{G}_*^{-1}\right)^T = \mathbf{X}^T \left(\mathbf{G}\pi^{\delta}\mathbf{G}^T\right)^{-1}\mathbf{G}\pi^{\delta}.$$

Hence the gradient of the likelihood function and the Fisher information matrix take the forms:

$$\mathbf{s}(\beta) = \mathbf{X}^T\left(\mathbf{G}\pi^{\delta}\mathbf{G}^{\mathbf{T}}\right)^{-\mathbf{1}}\mathbf{G}\mathbf{y},$$

$$\mathbf{F} = m\mathbf{X}^T \left(\mathbf{G}\pi^{\delta}\mathbf{G}^T\right)^{-1}\mathbf{X}.$$

It is easy to notice that the formula of the Fisher information matrix is simpler than the formula of the corresponding Hessian matrix.

9. Conclusions

Iterative methods are widely used in the estimation of parameters of logistic models by general statistical packages (SAS, Genstat). However, the method of iteration is very often imposed from above, when we choose some available (open) procedures. When we use the *logistic GLM* procedure in SAS, the Fisher scoring method is imposed from above. However there is the possibility of choosing the optimization technique (in SAS and in Genstat). In this paper the theory concerning the two iterative methods (their advantages and disadvantages) is presented. On this basis, the reader can decide which one of the methods is useful, or (based on the above theory) can write their own computational procedure in any statistical package.

In the paper by Bakinowska and Zawieja (2011) two iterative methods (Fisher scoring method and Newton-Raphson method) based on real and simulated data were empirically compared for various logistic models (the procedures were written in Turbo Pascal).

Summing up the results presented above we can state that the Hessian matrix of the likelihood function in a logistic model with multinomial distribution is expressed by a very complex formula. In practice it facilitates the estimation by using the Fisher scoring method. But it must be borne in mind that the number of iterations in the Newton-Raphson method with the complicated Hessian matrix is often smaller than in the Fisher scoring method, and the time needed to obtain the solution is shorter. This is very important in experiments in which a large quantity of data (observations) are used and in models with a large number of categories.

References

Agresti A. (1984): Analysis of Ordinal Categorical Data. Wiley, New York.

Bakinowska E. (2004): Analyzing experiments by generalized linear models. Listy Biometryczne - Biometrical Letters 41: 37–49.

Bakinowska E., Kala R. (2002a): Maximum Likelihood Estimation for Multivariate Categorical Data. Listy Biometryczne - Biometrical Letters 39: 29–40.

Bakinowska E., Kala R. (2002b): Estimation Methods in Generalized Linear Model (in Polish). Colloquium Biometrycze 32: 241–251.

Bakinowska E., Zawieja B. (2011): Empirical comparison of iterative methods in logistic model based on three real experiments and simulated data (in a review)

Fisz M. (1958): Probability and Mathematical Statistics (in Polish). PWN, Warszawa.

Glonek G.F.V., McCullagh P. (1995): Multivariate Logistic Models. J. R. Statist. Soc. B 57: 533–546

Grizzle J.E., Starmer C.F., Koch G.G. (1969): Analysis of Categorical Data by Linear Models. Biometrics 25: 489–504.

Harville D.A. (1997): Matrix Algebra From a Statistician's Perspective. Springer - Verlag, New York.

Krzyśko, M. (2000): Multivariate Statistical Analysis (in Polish). UAM, Poznań.

Mardia K.V., Kent I.T., Bibby I.M. (1979): Multivariate Analysis. Academic Press, London.

McCullagh P., Nelder, J.A. (1983): Generalized Linear Models. Chapman and Hall, London.

McCullagh P., Nelder, J.A. (1989): Generalized Linear Models. 2nd. ed. Chapman and Hall, London.

McCulloch Ch.E., Searle S.R. (2001): Generalized, Linear, and Mixed Models. Wiley, New York.