Listy Biometryczne - Biometrical Letters
Vol. 39(2002), No. 1, 29-40

Maximum likelihood estimation for multivariate categorical data

Ewa Bakinowska and Radosław Kala

Department of Mathematical and Statistical Methods,
Agricultural University of Poznań,
Wojska Polskiego 28, 60-637 Poznań, Poland,
ewabak@owl.au.poznan.pl, kalar@owl.au.poznan.pl

SUMMARY

The generalized linear models were developed for a broad class of practical problems related with the sets of multivariate categorical data. The analysis of such models can be based on the maximum likelihood principle. This approach has been discussed in Lang (1996; *Ann. Statist.* 24, 726-752), where the likelihood stationary equations and the large-sample distributions, as well as an iterative fitting algorithm, were presented. In this paper we give a direct method of deriving the maximum likelihood equations. This approach is free from the superfluous assumptions and, in result, can be applied to the experiments with a priori empty cells. The theory is illustrated by four simple examples.

KEY WORDS: multivariate categorical data, logistic transform, linear model, multinomial distribution.

1. Introduction

In experimental research the scores of the multiple classification occur frequently. In such cases the multinomial distribution provides a helpful tool for modelling the response data. A broad class of such problems can be solved with the aid of the generalized linear models (GLM). They are widely discussed by McCullagh and Nelder in their well known monographs (McCullagh and Nelder, 1983, 1989). However, there are some areas which are still left for further active research, as recently was pointed out by McCulloch (2000) and Christensen (2000).

In construction of the GL models, the multivariate logistic transform and properly selected linear space play the key role. The analysis of such models can be conducted, among other methods, by the method of maximum likelihood (ML). This approach has been discussed in Lang (1996), where the ML estimators and their asymptotic distribution, as well as the relevant fitting algorithm, were presented. In this paper it is shown, how the ML stationary equations, which form the base for the whole theory, can be derived more directly, without superfluous assumptions.

In Section 2 we state the basic facts about the family of multinomial distributions. In the next section we describe the general multivariate logistic transform and characterize the ML equations for fitting the generalized linear model. The equations obtained are the extentions of those obtained by Lang (1996), as they allow modelling irregular experiments in which some cells of the multiple classification are empty by assumption. The last section contains some simple examples illustrating the diversity of problems covered by the GL models, and exhibiting the practical aspects of the ML estimation.

In what follows we will use the symbols $\mathbf{A}^T$, $\mathbf{A}^{-1}$, and $\mathcal{C}(\mathbf{A})$ for the transpose, the inverse, and the column space of $\mathbf{A}$. The orthogonal complement of $\mathcal{C}(\mathbf{A})$ under the standard inner product will be denoted by $\mathcal{C}^{\perp}(\mathbf{A})$. Moreover, we will use the symbol $\mathbf{u}^{\delta}$ to denote the diagonal matrix with elements of a vector $\mathbf{u}$ on its diagonal, and the symbol $\mathbf{u}^{-\delta}$ for the inverse of $\mathbf{u}^{\delta}$ if the latter matrix is non-singular. Observe that if $\mathbf{u}^{-\delta}$ exists, then $\mathbf{u}^{-\delta}\mathbf{u} = \mathbf{1}$, where $\mathbf{1}$ is a vector with all elements equal to one.

2. ML estimation of multinomial distribution parameters

Categorical response data are usually modelled with the use of the multinomial distribution. Let us assume that we have a system of s independent random vectors $\mathbf{y}_{(i)}$, $i = 1, 2, ..., s$, each $\mathbf{y}_{(i)}$ following the multinomial distribution specified by a number m_i and a probability vector $\boldsymbol{\pi}_{(i)}$ of order k_i. Formally the single vector $\mathbf{y}_{(i)}$ is composed of the random variables $y_{i1}, y_{i2}, ..., y_{ik_i}$ fulfilling a constraint $\sum_{j=1}^{k_i} y_{ij} = m_i$, where k_i is the number of categories, y_{ij} is the number of successes in the j-th category, and m_i is the number of objects being classified. The probability of the success of j-th category corresponding to the vector $\mathbf{y}_{(i)}$ is represented by the j-th element of the vector $\boldsymbol{\pi}_{(i)}$. These probabilities fulfill the multinomial sampling constraint,

$$\sum_{j=1}^{k_i} \pi_{ij} = 1, \tag{1}$$

as it is called by Lang (1996). Then the probability of observing $\mathbf{y}_{(i)}$ is

$$\frac{m_i!}{y_{i1}!y_{i2}!\cdots y_{ik_i}!}\pi_{i1}^{y_{i1}}\pi_{i2}^{y_{i2}}\cdots\pi_{ik_i}^{y_{ik_i}}$$

and the kernel of the log-likelihood function takes the form

$$l(\boldsymbol{\pi}_{(i)}\colon \mathbf{y}_{(i)}) = \mathbf{y}_{(i)}^T \log \boldsymbol{\pi}_{(i)}, \tag{2}$$

where $\log \boldsymbol{\pi}_{(i)}$ is the vector of logarithms, $\log \boldsymbol{\pi}_{(i)} = (\log \pi_{i1}, \log \pi_{i2}, ..., \log \pi_{ik_i})^T$.

Taking into account the sampling constraint (1), which has to be fulfilled by elements of each vector $\boldsymbol{\pi}_{(i)}$, the parameter space for the joint distribution of $\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, ..., \mathbf{y}_{(s)}$ takes the form

$$\Omega = \{\boldsymbol{\pi} : \boldsymbol{\pi} > \mathbf{0},\ \mathbf{B}^T\boldsymbol{\pi} = \mathbf{1}_s\},$$

where $\boldsymbol{\pi} = (\boldsymbol{\pi}_{(1)}^T, \boldsymbol{\pi}_{(2)}^T, ..., \boldsymbol{\pi}_{(s)}^T)^T$ and $\mathbf{B}$ is a block-diagonal matrix,

$$\mathbf{B} = \mathrm{diag}(\mathbf{1}_{k_1}, \mathbf{1}_{k_2}, ..., \mathbf{1}_{k_s}) = \oplus_{i=1}^{s}(\mathbf{1}_{k_i}). \tag{3}$$

In view of the constraint $\mathbf{B}^T\boldsymbol{\pi} = \mathbf{1}_s$ the set Ω is contained in the $k - s$ dimensional affine space, where $k = k_1 + k_2 + ... + k_s$. Since the vectors $\mathbf{y}_{(1)}, \mathbf{y}_{(2)}, ..., \mathbf{y}_{(s)}$ are considered as independent, the kernel of the joint log-likelihood function is the sum of the kernels of the form (2). In consequence, the likelihood equation for $\boldsymbol{\pi}$ can be written as

$$\mathbf{y}^T\boldsymbol{\pi}^{-\delta}\mathbf{H} = \mathbf{0}, \tag{4}$$

where $\mathbf{y} = (\mathbf{y}_{(1)}^T, \mathbf{y}_{(2)}^T, ..., \mathbf{y}_{(s)}^T)^T$ and $\mathbf{H}$ is any such matrix that $\mathcal{C}(\mathbf{H}) = \mathcal{C}^{\perp}(\mathbf{B})$. Observe that the natural estimates of the probabilities π_{ij}, provided by the frequencies $p_{ij} = y_{ij}/m_i$, $i = 1, 2, ..., s$, $j = 1, 2, ..., k_i$, form a solution of (4) and so, are the ML estimates.

3. Logistic transforms and GL models

The logistic transform links the expectation of the frequency vector $\mathbf{p}$, $E(\mathbf{p}) = \boldsymbol{\pi}$, with a new parameter vector $\boldsymbol{\eta}$. This mapping can be expressed as

$$\boldsymbol{\eta} = \mathbf{C}^T \log(\mathbf{L}\boldsymbol{\pi}), \tag{5}$$

where $\mathbf{L}$ is a fixed binary matrix, such that the product $\mathbf{L}\boldsymbol{\pi}$ compresses selected probability sums, and $\mathbf{C}^T$ is usually a matrix of contrasts. The constructions of matrices $\mathbf{L}$ and $\mathbf{C}$ are given by Glonek and McCullagh (1994). If $\mathbf{L}$ is the identity matrix, the individual elements of $\boldsymbol{\eta}$ are referred to as log-linear contrasts and, otherwise, they are referred to as multivariate logistic contrasts (Glonek, 1996). However, if it is desired that (5) represents the one-to-one transform, then not all elements of $\boldsymbol{\eta}$ can be considered as contrasts.

The GL model appears when the range of the logistic transform $\boldsymbol{\pi} \to \boldsymbol{\eta}$ is restricted to a priori given linear subspace. Such a relationship can be expressed as

$$\boldsymbol{\eta} \in \mathcal{C}(\mathbf{X}), \tag{6}$$

where $\mathbf{X}$ is a matrix of full column rank. The matrix $\mathbf{X}$ expresses some postulated properties of the model. It reflects, with the use of the mapping (5), some relationships among probabilities contained in the vector $\boldsymbol{\pi}$, or relates the probability vectors $\boldsymbol{\pi}_{(i)}$ with the experimental conditions. In the latter case, we assume that each $\mathbf{y}_{(i)}$ is observed under the conditions specified by the vector $\mathbf{x}_{(i)}$ of concomitant variables. Such models can be seen as the family of multinomial variables with the restricted parameter space

$$\Omega_x = \{\boldsymbol{\pi} : \boldsymbol{\pi} > \mathbf{0}, \mathbf{B}^T\boldsymbol{\pi} = \mathbf{1}_s, \mathbf{C}^T \log(\mathbf{L}\boldsymbol{\pi}) \in \mathcal{C}(\mathbf{X})\}.$$

The likelihood equation for estimating the probability vector $\boldsymbol{\pi}$ is stated in the following

THEOREM 1. *If $\mathbf{y}$ is a response vector from the product of s multinomial distributions, then the ML estimator of $\boldsymbol{\pi}$, under Ω_x, and a vector $\boldsymbol{\lambda}$ fulfill the equations:*

$$\mathbf{y} - (m_1\boldsymbol{\pi}_{(1)}^T, m_2\boldsymbol{\pi}_{(2)}^T, \ldots, m_s\boldsymbol{\pi}_{(s)}^T)^T = \oplus_{i=1}^s(\boldsymbol{\pi}_{(i)}^\delta - \boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^T)\mathbf{L}^T\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\boldsymbol{\lambda}, \tag{7}$$

and

$$\mathbf{M}^T\mathbf{C}^T \log(\mathbf{L}\boldsymbol{\pi}) = \mathbf{0}, \tag{8}$$

where $\mathbf{D} = (\mathbf{L}\boldsymbol{\pi})^\delta$ *and* $\mathcal{C}(\mathbf{M}) = \mathcal{C}^\perp(\mathbf{X})$.

Proof. In view of the definition of the logistic transform and of the property of the matrix $\mathbf{M}$, the condition (6) is equivalent to (8) or to the equality

$$\mathbf{M}^T\boldsymbol{\eta} = \mathbf{0}.$$

Taking into account the last restriction and the sampling constraint in the log-likelihood function, we have

$$l^+(\boldsymbol{\pi} : \mathbf{y}) = \mathbf{y}^T \log \boldsymbol{\pi} - \boldsymbol{\lambda}^T\mathbf{M}^T\boldsymbol{\eta} - \boldsymbol{\mu}^T(\mathbf{B}^T\boldsymbol{\pi} - \mathbf{1}_s), \tag{9}$$

where $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are the vectors of the Lagrange multipliers. If $\boldsymbol{\pi} > \mathbf{0}$ then, as was observed by Grizzle *at al.* (1969), the Jacobian of the transform $\boldsymbol{\pi} \to \boldsymbol{\eta}$ takes the form

$$\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\pi}^T} = \mathbf{C}^T\mathbf{D}^{-1}\mathbf{L}.$$

Differentiating (9) with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\lambda}$, we obtain

$$\frac{\partial l^{+}}{\partial \boldsymbol{\pi}} = \boldsymbol{\pi}^{-\delta}\mathbf{y} - \mathbf{L}^{T}\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\boldsymbol{\lambda} - \mathbf{B}\boldsymbol{\mu}, \tag{10}$$

and

$$\frac{\partial l^{+}}{\partial \boldsymbol{\lambda}} = \boldsymbol{\eta}^{T}\mathbf{M}. \tag{11}$$

Comparing (11) with zero vector, we obtain the condition (8). To show (7), first observe that the matrix $\mathbf{H} = \oplus_{i=1}^{s}(\boldsymbol{\pi}_{(i)}^{\delta} - \boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^{T}) = \boldsymbol{\pi}^{\delta} - \oplus_{i=1}^{s}(\boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^{T})$ spans the orthogonal complement of $\mathcal{C}(\mathbf{B})$, i.e. $\mathcal{C}(\mathbf{H}) = \mathcal{C}^{\perp}(\mathbf{B})$. Therefore, equating (10) with a zero vector and premultiplying the resulting equation by $\mathbf{H}$, leads to the equation

$$\mathbf{y} - \oplus_{i=1}^{s}(\boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^{T})\boldsymbol{\pi}^{-\delta}\mathbf{y} = \oplus_{i=1}^{s}(\boldsymbol{\pi}_{(i)}^{\delta} - \boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^{T})\mathbf{L}^{T}\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\boldsymbol{\lambda}. \tag{12}$$

But

$$\oplus_{i=1}^{s}(\boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^{T})\boldsymbol{\pi}^{-\delta}\mathbf{y} = \oplus_{i=1}^{s}(\boldsymbol{\pi}_{(i)}\mathbf{1}_{k_i}^{T})\mathbf{y} = (m_1\boldsymbol{\pi}_{(1)}^{T},\ m_2\boldsymbol{\pi}_{(2)}^{T},\ \dots,\ m_s\boldsymbol{\pi}_{(s)}^{T})^{T},$$

which together with (12) gives the equation (7). □

The following two corollaries exhibit some simplification of the equation (7).

COROLLARY 1. *If $s = 1$ and the matrix $\mathbf{C}$ in the logistic transform (5) is such that $\mathbf{1}^{T}\mathbf{C} = \mathbf{0}$, then the equation (7) takes the form*

$$\mathbf{y} - m_1\boldsymbol{\pi}_{(1)} = \boldsymbol{\pi}_{(1)}^{\delta}\mathbf{L}^{T}\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\boldsymbol{\lambda}.$$

If in addition $\mathbf{L} = \mathbf{I}$, it simplifies further to the form

$$\mathbf{y} - m_1\boldsymbol{\pi}_{(1)} = \mathbf{C}\mathbf{M}\boldsymbol{\lambda}.$$

Proof. The first simplification follows from the equalities

$$\boldsymbol{\pi}_{(1)}^{T}\mathbf{L}^{T}\mathbf{D}^{-1}\mathbf{C} = \boldsymbol{\pi}_{(1)}^{T}\mathbf{L}^{T}(\boldsymbol{\pi}_{(1)}^{T}\mathbf{L}^{T})^{-1}\mathbf{C} = \mathbf{1}^{T}\mathbf{C} = \mathbf{0},$$

while the second – from the observation that $\boldsymbol{\pi}_{(1)}^{\delta}\mathbf{L}^{T}\mathbf{D}^{-1} = \boldsymbol{\pi}_{(1)}^{\delta}(\boldsymbol{\pi}_{(1)}^{T})^{-1} = \mathbf{I}$. □

COROLLARY 2. *If $s = 1$, $\mathbf{1} \in \mathcal{C}(\mathbf{X})$, and $\mathbf{C} = \mathbf{I}$, then the equation (7) takes the form*

$$\mathbf{y} - m_1\boldsymbol{\pi}_{(1)} = \boldsymbol{\pi}_{(1)}^{\delta}\mathbf{L}^{T}\mathbf{D}^{-1}\mathbf{M}\boldsymbol{\lambda},$$

and, if in addition $\mathbf{L} = \mathbf{I}$, it simplifies further to the form

$$\mathbf{y} - m_1\boldsymbol{\pi}_{(1)} = \mathbf{M}\boldsymbol{\lambda}. \tag{13}$$

The next corollaries contain the direct generalization of the results established by Lang (1996), in his Theorem 3.1.

COROLLARY 3. *If in the logistic transform (5) the matrices* $\mathbf{C}$ *and* $\mathbf{L}$ *have the forms*

$$\mathbf{C}^T = \oplus_{i=1}^s(\mathbf{C}_i^T),\ \mathbf{L} = \oplus_{i=1}^s(\mathbf{L}_i), \tag{14}$$

where $\mathbf{C}_i^T$ *is of order* $c_i \times l_i$ *and such that* $\mathbf{1}_{l_i}^T\mathbf{C}_i = \mathbf{0}$*, and* $\mathbf{L}_i$ *is of order* $l_i \times k_i$*, then the equation (7) can be expressed as*

$$\mathbf{y} - (m_1\boldsymbol{\pi}_{(1)}^T,\ m_2\boldsymbol{\pi}_{(2)}^T,\ ...,\ m_s\boldsymbol{\pi}_{(s)}^T)^T = \boldsymbol{\pi}^\delta\mathbf{L}^T\mathbf{D}^{-1}\mathbf{C}\mathbf{M}\boldsymbol{\lambda}, \tag{15}$$

where $\mathbf{D} = \oplus_{i=1}^s(\mathbf{L}_i\boldsymbol{\pi}_i)^\delta$*. If in addition* $\mathbf{L} = \mathbf{I}$*, it simplifies further to the form*

$$\mathbf{y} - (m_1\boldsymbol{\pi}_{(1)}^T,\ m_2\boldsymbol{\pi}_{(2)}^T,\ ...,\ m_s\boldsymbol{\pi}_{(s)}^T)^T = \mathbf{C}\mathbf{M}\boldsymbol{\lambda}.$$

Proof. For the first simplification it suffices to observe that

$$\oplus_{i=1}^s(\boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^T)\mathbf{L}^T\mathbf{D}^{-1}\mathbf{C} = \oplus_{i=1}^s(\boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^T\mathbf{L}_i^T(\mathbf{L}_i\boldsymbol{\pi}_{(i)})^{-\delta}\mathbf{C}_i) = \oplus_{i=1}^s(\boldsymbol{\pi}_{(i)}\mathbf{1}_{l_i}^T\mathbf{C}_i),$$

while the second simplification follows from the argument used in the proof of Corollary 1. □

COROLLARY 4. *If in the logistic transform* $\boldsymbol{\eta} = \log(\mathbf{L}\boldsymbol{\pi})$ *the matrix* $\mathbf{L} = \oplus_{i=1}^s(\mathbf{L}_i)$*, where* $\mathbf{L}_i$ *is of order* $l_i \times k_i$*, and* $\oplus_{i=1}^s(\mathbf{1}_{l_i}^T)\mathbf{M} = \mathbf{0}$*, then the equation (7) reduces to the form*

$$\mathbf{y} - (m_1\boldsymbol{\pi}_{(1)}^T,\ m_2\boldsymbol{\pi}_{(2)}^T,\ ...,\ m_s\boldsymbol{\pi}_{(s)}^T)^T = \boldsymbol{\pi}^\delta\mathbf{L}^T\mathbf{D}^{-1}\mathbf{M}\boldsymbol{\lambda},$$

where $\mathbf{D} = \oplus_{i=1}^s(\mathbf{L}_i\boldsymbol{\pi}_i)^\delta$*. If in addition* $\mathbf{L} = \mathbf{I}$*, it simplifies further to the form*

$$\mathbf{y} - (m_1\boldsymbol{\pi}_{(1)}^T,\ m_2\boldsymbol{\pi}_{(2)}^T,\ ...,\ m_s\boldsymbol{\pi}_{(s)}^T)^T = \mathbf{M}\boldsymbol{\lambda}.$$

Proof. The first simplification results from the equalities:

$$\oplus_{i=1}^s(\boldsymbol{\pi}_{(i)}\boldsymbol{\pi}_{(i)}^T) = \oplus_{i=1}^s(\boldsymbol{\pi}_{(i)})\oplus_{i=1}^s(\boldsymbol{\pi}_{(i)}^T)$$

and

$$\oplus_{i=1}^s(\boldsymbol{\pi}_{(i)}^T)\mathbf{L}^T\mathbf{D}^{-1}\mathbf{M} = \oplus_{i=1}^s(\boldsymbol{\pi}_{(i)}^T\mathbf{L}_i^T(\mathbf{L}_i\boldsymbol{\pi}_{(i)})^{-\delta})\mathbf{M} = \oplus_{i=1}^s(\mathbf{1}_{l_i}^T)\mathbf{M} = \mathbf{0}.$$

The second simplification is obvious. □

Remark 1. Note that the equation (7) reduces also to the form (15), when $\mathbf{M}$ can be partitioned, $\mathbf{M} = (\mathbf{M}_1^T, \mathbf{M}_2^T, ..., \mathbf{M}_s^T)$, in such a way that if for some i $\mathbf{C}_i = \mathbf{I}_{k_i}$ then $\mathbf{1}_{k_i}^T\mathbf{M}_i = \mathbf{0}$, and $\mathbf{1}_{l_i}^T\mathbf{C}_i = \mathbf{0}$ for remaining i.

Remark 2. Contrary to the result of Lang (1996), the established equations enable one to consider the problem of the ML estimation of the probability vector $\boldsymbol{\pi}$ when some cells of multiple classification are empty by assumption and when the different logistic transfoms, defined by the submatrices $\mathbf{C}_i$ and $\mathbf{L}_i$, are applied to different vectors $\boldsymbol{\pi}_{(i)}$, i.e. when $\mathbf{C}_i \neq \mathbf{C}_j$ and $\mathbf{L}_i \neq \mathbf{L}_j$ for $i \neq j$.

4. Examples

In this section we present some examples which show a broad range of applications of the GL models and simultaneously exhibit the methods of solving the corresponding maximum likelihood equations. For simplicity of presentation we restrict our attention to the examples in which the logistic transform consists only of the log-linear contrasts, i.e., $\mathbf{L}$ is the identity matrix. This corresponds to the assumption that categories of the multinomial distribution are neither ordered nor form any hierarchical structure, in which cases the cumulative and conditional probabilities are of main interest (Agresti, 1990, Lang and Agresti, 1994).

For the first example let us assume that there are two discrete random variables A and B with two and three categories, respectively. Moreover, let the joint distribution of A and B be defined by the probabilities $\pi_{ij} = P(A = i,\ B = j)$,

$$\boxed{\begin{array}{ccc} \pi_{11} & \pi_{12} & \pi_{13} \\ \pi_{21} & \pi_{22} & \pi_{23} \end{array}},$$

which fulfill a single ($s = 1$) sampling constraint $\pi_{11}+\pi_{12}+\pi_{13}+\pi_{21}+\pi_{22}+\pi_{23} = 1$. The main problem which is usually posed in this context concerns the independence of the variables A and B . The appropriate condition takes then the well known form

$$\frac{\pi_{11}}{\pi_{21}} = \frac{\pi_{12}}{\pi_{22}} = \frac{\pi_{13}}{\pi_{23}}, \tag{16}$$

which can be expressed as the log-linear model $\boldsymbol{\eta} = \log(\boldsymbol{\pi}) \in \mathcal{C}(\mathbf{X})$, where

$$\boldsymbol{\pi} = \begin{pmatrix} \pi_{11} \\ \pi_{21} \\ \pi_{12} \\ \pi_{22} \\ \pi_{13} \\ \pi_{23} \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

Since $\mathbf{C} = \mathbf{L} = \mathbf{I}_6$ and $\mathbf{1}_6 \in \mathcal{C}(\mathbf{X})$, the assumptions of Corollary 2 are fulfilled. Thus the ML estimator of $\boldsymbol{\pi}$ is a solution of (8) together with (13). The first equation can be expressed as

$$\begin{pmatrix} 1 & -1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix} \log \boldsymbol{\pi} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{17}$$

while the second takes the form

$$\mathbf{y} - m\boldsymbol{\pi} = \mathbf{M}\boldsymbol{\lambda}, \tag{18}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)^T$ and $\mathbf{M}$ is specified in (17). Since the conditions (17) and (16) are equivalent, we are looking for $\boldsymbol{\pi}$ fulfilling (18) and (16). It can be checked that such a solution is provided by the estimates

$$\widehat{\pi}_{ij} = \frac{y_{i\cdot} y_{\cdot j}}{m^2},$$

where $y_{i\cdot} = y_{i1} + y_{i2} + y_{i3}$, $i = 1, 2$, $y_{\cdot j} = y_{1j} + y_{2j}$, $j = 1, 2, 3$ and $m = y_{1\cdot} + y_{2\cdot}$. These estimates are conventionally calculated for the two-dimensional contingency tables, under the independence assumption.

Note that the vector $\mathbf{y} - m\widehat{\boldsymbol{\pi}} = \mathbf{M}\widehat{\boldsymbol{\lambda}}$ describes the differences between the observed frequencies and their expected values when the assumption of independence is satisfied. Therefore, the norm of $\mathbf{M}\widehat{\boldsymbol{\lambda}}$ is a measure of fitting the model to the observed data.

As a second example let us consider the incomplete 3 × 3 contingency table of the form

π_{11}	π_{12}	π_{13}
π_{21}	π_{22}	$*$
π_{31}	$*$	$*$

.

In such case we can be interested in fitting the quasi-independence model, as it is called by Goodman (1968). This assumption implies the equality

$$\frac{\pi_{11}}{\pi_{21}} = \frac{\pi_{12}}{\pi_{22}}, \tag{19}$$

which again can be expressed in the form

$$\begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \end{pmatrix} \log \boldsymbol{\pi} = 0, \tag{20}$$

where $\boldsymbol{\pi}^T = (\pi_{11}, \pi_{12}, \pi_{13}, \pi_{21}, \pi_{22}, \pi_{31})$. Since the assumptions of Corollary 2 are again satisfied, the ML estimator of $\boldsymbol{\pi}$ is a solution of the equations (19) and (13), with $\mathbf{M}$ given in (20). In consequence we obtain the set of equations:

$$\begin{array}{lll} y_{11} - m\pi_{11} = \lambda, & y_{12} - m\pi_{12} = -\lambda, & y_{13} - m\pi_{13} = 0, \\ y_{21} - m\pi_{21} = -\lambda, & y_{22} - m\pi_{22} = \lambda, & \\ y_{31} - m\pi_{31} = 0, & & \end{array}$$

They lead to the solutions:

$$\begin{array}{lll}
\hat{\pi}_{11} = y_{1*}y_{*1}/(Mm), & \hat{\pi}_{12} = y_{1*}y_{*2}/(Mm), & \hat{\pi}_{13} = y_{13}/m, \\
\hat{\pi}_{21} = y_{2*}y_{*1}/(Mm), & \hat{\pi}_{22} = y_{2*}y_{*2}/(Mm), & \\
\hat{\pi}_{31} = y_{13}/m, & &
\end{array}$$

where $y_{i*} = y_{i1} + y_{i2}$, $y_{*j} = y_{1j} + y_{2j}$, i, $j = 1, 2$, and $M = y_{11} + y_{12} + y_{21} + y_{22}$, while $m = M + y_{13} + y_{31}$. Actually it means that the upper left 2×2 block of our triangular contingency table provides the estimates under the independence condition (19), which are then corrected with respect to the additional non-empty cells.

The general recursive procedure for fitting the expected cell frequencies for triangular tables under quasi-independence model was proposed by Bishop and Fienberg (1969). The alternative methods for establishing the maximum likelihood estimates can be found in Goodman (1968).

For the third example let us assume that there are four categories and two populations, but not all categories are possible for all objects of each population. To be more specific, let us assume that the two multinomial distributions, for the first and second group of m_1 and of m_2 objects, sampled from the first and the second population, respectively, are determined by the first and second row of the table

$$\boxed{\begin{array}{cccc}
\pi_{11} & \pi_{12} & \pi_{13} & * \\
* & \pi_{22} & \pi_{23} & \pi_{24}
\end{array}}\,.$$

Although the rows represent here independent distributions, we can postulate some relations between probabilities π_{ij}. Adopting the quasi-independence, we have the condition

$$\frac{\pi_{12}}{\pi_{22}} = \frac{\pi_{13}}{\pi_{23}}. \tag{21}$$

It can be expressed as

$$\begin{pmatrix} 0 & 1 & -1 & -1 & 1 & 0 \end{pmatrix} \log \boldsymbol{\pi} = 0, \tag{22}$$

where $\boldsymbol{\pi}^T = (\boldsymbol{\pi}_{(1)}^T, \boldsymbol{\pi}_{(2)}^T) = ((\pi_{11}, \pi_{12}, \pi_{13}), (\pi_{22}, \pi_{23}, \pi_{24}))$ is such that $\boldsymbol{\pi}_{(1)}^T \mathbf{1}_3 = 1$, $\boldsymbol{\pi}_{(2)}^T \mathbf{1}_3 = 1$. Using now Corollary 4 with $\mathbf{L} = \mathbf{I}$ the likelihood equation for $\boldsymbol{\pi}$ consists of the condition (21) and the equation

$$\mathbf{y} - (m_1 \boldsymbol{\pi}_{(1)}^T, m_2 \boldsymbol{\pi}_{(2)}^T)^T = \mathbf{M}\lambda,$$

where $\mathbf{M}$ is given in (22). They lead to the estimates

$$\begin{array}{llll}
\hat{\pi}_{11} = y_{11}/m_1, & \hat{\pi}_{12} = y_{1*}y_{*2}/(Mm_1), & \hat{\pi}_{13} = y_{1*}y_{*3}/(Mm_1), & * \\
* & \hat{\pi}_{22} = y_{2*}y_{*2}/(Mm_2), & \hat{\pi}_{23} = y_{2*}y_{*3}/(Mm_2), & \hat{\pi}_{24} = y_{24}/m_2,
\end{array}$$

where $y_{i*} = y_{i1} + y_{i2}$, $i = 1, 2$, $y_{*j} = y_{1j} + y_{2j}$, $j = 2, 3$, and $M = y_{12} + y_{13} + y_{22} + y_{23}$.

For the last example let us assume that A is a binary random variable, with the success probability π. Moreover, assume that A is observed together with a deterministic covariate variable x. In such case it is interesting to determine the probability π as a function of x, $\pi = \pi(x)$. Usually $\pi(x)$ is modelled in the form

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}},$$

which ensures the inequality: $0 < \pi(x) < 1$. In result

$$\log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta x. \tag{23}$$

Assume now that the variable A is observed only at three points: x_1, x_2, x_3. Moreover, assume that the number of classified objects for x_i is m_i and that

$$\pi_{1i} = \pi(x_i),\ \pi_{2i} = 1 - \pi(x_i),\ i = 1,\ 2,\ 3. \tag{24}$$

Then, in view of (23), we have the equality

$$\begin{pmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \tag{25}$$

where $\boldsymbol{\eta}$ results from mapping of the probability vector $\boldsymbol{\pi} = (\boldsymbol{\pi}_{(1)}^T, \boldsymbol{\pi}_{(2)}^T, \boldsymbol{\pi}_{(3)}^T)^T = (\pi_{11}, \pi_{21}, \pi_{12}, \pi_{22}, \pi_{13}, \pi_{23})^T$ by the logistic transform

$$\boldsymbol{\eta} = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \log \boldsymbol{\pi}. \tag{26}$$

Since the assumptions of Corollary 3 with $\mathbf{L} = \mathbf{I}$ are fulfilled, the ML estimator of $\boldsymbol{\pi}$ satisfies the equations

$$\mathbf{y} - (m_1\boldsymbol{\pi}_{(1)}^T, m_2\boldsymbol{\pi}_{(2)}^T, m_3\boldsymbol{\pi}_{(3)}^T)^T = \mathbf{CM}\lambda$$

and

$$\mathbf{M}^T\mathbf{C}^T \log \boldsymbol{\pi} = 0,$$

where $\mathbf{C}$ is given in (26) while

$$\mathbf{M}^T = (x_3 - x_2, x_1 - x_3, x_2 - x_1).$$

In view of (24), they lead to the system of linear equations

$$\begin{aligned} y_{11} - m_1\pi_{11} &= (x_3 - x_2)\lambda, \\ y_{12} - m_2\pi_{12} &= (x_1 - x_3)\lambda, \\ y_{13} - m_3\pi_{13} &= (x_2 - x_1)\lambda, \end{aligned} \tag{27}$$

which must be solved together with the nonlinear equation of the form

$$(x_3 - x_2)g(\pi_{11}) + (x_1 - x_3)g(\pi_{12}) + (x_2 - x_1)g(\pi_{13}) = 0, \tag{28}$$

where $g(\pi_{1i}) = \log(\pi_{1i}/(1 - \pi_{1i}))$, $i = 1, 2, 3$. This can be done with the use of the algorithm described by Lang (1996) or by Glonek (1996). Having the solution $\boldsymbol{\pi}$ one can find the ML estimator for the parameters α and β, since the equation (28) actually ensures the consistency of the system (25). Note also that the right hand side of (27) provides a measure of fitting the assumed model.

It can be surprising that so different experimental situations can be fitted into the same theoretical frames. But, it is not so, since the GL models enable one to combine the linear structures with nonlinear transformations and the ML approach is really very powerful inference tool.

Acknowledgments

The authors would like to thank the referee for his helpful comments.

References

Agresti, A. (1990). *Categorical Data Analysis.* Wiley, New York.

Bishop, Y. M. M. and Fienberg, S. E. (1969). Incomplete two-dimensional contingency tables, *Biometrics*, 25, 119-128.

Christensen, R. (2000). Linear and log-linear models. *Journal of the American Statistical Association*, 95, 1290-1293.

Glonek, G. F. V. (1996). A class of regression models for multivariate categorical responses. *Biometrika*, 83, 15-28.

Glonek, G. F. V. and McCullagh, P. (1994). Multivariate logistic models. *Technical Report* 94-31. School of Information Sciences and Technology, Flinders University of South Australia, Adelaide.

Goodman, L. A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency table with or without missing entries. R. A. Fisher Memorial Lecture, delivered at the annual meeting of the American Statistical Association and the Biometric Society, Pittsburgh, Pennsylvania.

Grizzle J. E., Starmer, C. F. and Koch, G. G. (1969). Analysis of categorical data by linear models, *Biometrics*, 25, 489-504.

Lang, J. B. (1996). Maximum likelihood methods for a generalized class of log-linear models. *Ann. Statist.*, 24, 726-752.

Lang, J. B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses. *Journal of the American Statistical Association*, 89, 625-632.

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

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

McCulloch, C. (2000). Generalized linear models. *Journal of the American Statistical Association*, 95, 1320-1324.

Received 10 February 2002; revised 20 September 2002

Estymacja metodą największej wiarogodności dla wielowymiarowych danych skategoryzowanych

Streszczenie

Szereg problemów dotyczących badań, w których odnotowuje się rezultaty wielokrotnej klasyfikacji ustalonego zespołu obiektów, może być rozwiązanych z użyciem uogólnionych modeli liniowych. Analizę takich modeli można wyprowadzić z zasady największej wiarogodności. Takie podejście zostało omówione przez Langa (1996; *Ann. Statist.* 24, 726-752), który przedstawił m.in. równania największej wiarogodności oraz podał iteracyjne algorytmy ich rozwiązywania. W obecnej pracy równania największej wiarogodności wyprowadzono bezpośrednio. Pokazane jest podejście wolne od zbędnych założeń, które może być stosowane do modelowania eksperymentów, w których ustalone podklasy wielokrotnej klasyfikacji są z założenia puste. Teorię zilustrowano czterema przykładami.

Słowa kluczowe: wielowymiarowe dane skategoryzowane, transformacja logistyczna, model liniowy, rozkład wielomianowy.