Multivariate Logistic Models
Author(s): G. F. V. Glonek and P. McCullagh
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 57, No. 3 (1995), pp. 533-546
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2346155
Accessed: 16-03-2019 16:43 UTC

J. R. Statist. Soc. B (1995)
57, *No.* 3, *pp.* 533–546

Multivariate Logistic Models

By G. F. V. GLONEK† and P. McCULLAGH

Flinders University of South Australia, Adelaide, Australia

University of Chicago, USA

[Received November 1993. Final revision October 1994]

SUMMARY

When data composed of several categorical responses together with categorical or continuous predictors are observed, the multivariate logistic transform introduced by McCullagh and Nelder can be used to define a class of regression models that is, in many applications, particularly suitable for relating the joint distribution of the responses to predictors. In this paper we give a general definition of this class of models and study their properties. A computational scheme for performing maximum likelihood estimation for data sets of moderate size is described and a system of model formulae that succinctly define particular models is introduced. Applications of these models to longitudinal problems are illustrated by numerical examples.

Keywords: LOGISTIC REGRESSION; LONGITUDINAL DATA; MULTIVARIATE CATEGORICAL DATA; POLYTOMOUS RESPONSES

1. INTRODUCTION

Consider two binary responses A and B, and let $\pi_{ij} = \text{pr}(A = i,\ B = j)$. The bivariate logistic transform was defined by McCullagh and Nelder (1989) by the mapping $\boldsymbol{\pi} \mapsto \boldsymbol{\eta} = (\eta_a,\ \eta_b,\ \eta_{ab})$ where

$$\eta_a = \text{logit}(\pi_{1.}), \qquad \eta_b = \text{logit}(\pi_{.1}), \qquad \eta_{ab} = \log\left(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right),$$

the dot subscript denoting summation. When bivariate binary data are observed under different observational or experimental conditions as encoded in the predictor variable $\mathbf{x}$, we can consider a class of models that relate $\boldsymbol{\pi}$ to $\mathbf{x}$ by

$$\eta_a = \beta_a^{\mathrm{T}}\mathbf{x}_a, \qquad \eta_b = \beta_b^{\mathrm{T}}\mathbf{x}_b, \qquad \eta_{ab} = \beta_{ab}^{\mathrm{T}}\mathbf{x}_{ab},$$

where $\mathbf{x}_a$, $\mathbf{x}_b$ and $\mathbf{x}_{ab}$ are subsets of $\mathbf{x}$ and β_a, β_b and β_{ab} are parameters to be estimated. Such models are called bivariate logistic regression models.

The primary motivation for considering these models is best explained by contrasting them with the more familiar log-linear regression models. In particular, if we consider the transformation

$$\lambda_a = \log\left(\frac{\pi_{11}}{\pi_{21}}\right), \qquad \lambda_b = \log\left(\frac{\pi_{11}}{\pi_{12}}\right), \qquad \lambda_{ab} = \log\left(\frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}\right),$$

then a model of the form

†*Address for correspondence*: Department of Mathematics and Statistics, Flinders University of South Australia, GPO Box 2100, Adelaide, SA 5001, Australia.
E-mail: gary@stats.flinders.edu.au

 0035–9246/95/57533

$$\lambda_a = \beta_a^{\mathrm{T}} \mathbf{x}_a, \qquad \lambda_b = \beta_b^{\mathrm{T}} \mathbf{x}_b, \qquad \lambda_{ab} = \beta_{ab}^{\mathrm{T}} \mathbf{x}_{ab} \tag{1}$$

is called a log-linear regression model.

Suppose initially that only the single binary response A is observed. A natural class of models to consider in this case is the linear logistic models,

$$\mathrm{logit}(\pi) = \beta_a^{\mathrm{T}} \mathbf{x}_a,$$

where $\pi = \mathrm{pr}(A = 1) = \pi_{1.}$. When the second response B is added to the problem, both the bivariate logistic and the log-linear regression models can be seen as generalizations of the univariate logistic regression model. A serious objection to the log-linear model (1) is that it is incompatible with the univariate model that it seeks to generalize, i.e., given model (1), the marginal relationship between A and $\mathbf{x}$ is in general not linear on the logistic scale. Furthermore, if a third response C were recorded, the log-linear model that generalizes model (1) would be similarly inconsistent with both the univariate logistic model and the bivariate log-linear model.

Thus within the family of log-linear regression models the way that the marginal relationship between A and $\mathbf{x}$ is modelled depends on the number of response variables considered. By contrast, the bivariate logistic regression model implies univariate logistic models for both A and B marginally. It is this property, termed 'upward compatibility' by McCullagh (1989) or 'reproducibility' by Liang *et al.* (1992), that makes the bivariate logistic model preferable to the log-linear model for many applications. In cases where both responses arise on an equal footing and the objective of the analysis is to study how the joint distribution of A and B varies with $\mathbf{x}$, as opposed to trying to predict A from B or vice versa, it is difficult to see why the parameters λ_a and λ_b, which represent conditional logits in the log-linear decomposition of $\boldsymbol{\pi}$, would be of interest. By contrast, the parameters η_a and η_b of the bivariate logistic model are the marginal logits that would be considered in the univariate analyses and therefore have a natural and useful interpretation that is independent of other observed response variables.

In this light, it would seem that the log-linear approach has gained wide acceptance largely because of its convenient theoretical properties and the relative ease with which models can be specified and maximum likelihood calculations performed. To restore the balance, we define the class of multivariate logistic regression models in a way that leads to a computational scheme that is feasible for problems of moderate size. Various theoretical aspects of these models are also addressed.

2. MULTIVARIATE LOGISTIC TRANSFORM

The multivariate logistic transform as described for two binary variables in Section 1 can be extended to an arbitrary number of discrete response variables of either nominal or ordinal type. The general form of the transformation is

$$\boldsymbol{\eta} = C^{\mathrm{T}} \log(L\boldsymbol{\pi}), \tag{2}$$

where L and C are tensor products of suitably chosen marginal indicator and contrast matrices respectively. An explicit description of their construction is given in Glonek and McCullagh (1994). McCullagh and Nelder (1989), p. 219, illustrated this formulation of the multivariate logistic transform for three binary variables. Other models based on equation (2) have previously been considered by Grizzle *et al.* (1969).

When the transformation is applied in the general case with response variables $A_1, A_2, \ldots, A_d$ having $r_1, r_2, \ldots, r_d$ levels respectively, the components of $C^T \log(L\pi)$ are response contrasts denoted symbolically by

$$\boldsymbol{\eta} = (\phi, A_1, A_2, A_1A_2, A_3, A_1A_3, A_2A_3, A_1A_2A_3, \ldots, A_1A_2 \ldots A_d),$$

where ϕ represents the null contrast $\log(\Sigma\, \pi) = 0$. Although ϕ is strictly superfluous, it is convenient to retain it as a means of ensuring that the mapping $\boldsymbol{\pi} \mapsto \boldsymbol{\eta}$ is of full rank and also expressing the requirement that $\Sigma\, \pi = 1$.

The derivative matrix $(\partial\eta_i/\partial\pi_j)$ can be seen, from the chain rule, to be

$$\left(\frac{\partial\eta}{\partial\pi}\right) = C^T D^{-1} L$$

where $D = \mathrm{diag}(L\pi)$, as was stated in Grizzle *et al.* (1969). Glonek and McCullagh (1994) prove, for the class of matrices C and L defined there, the following result.

Theorem 1. The matrix $C^T D^{-1} L$ is non-singular provided that $\pi > 0$.

3. INVERTING LOGISTIC TRANSFORM

One way to fit marginal models is to maximize a Poisson likelihood kernel subject to constraints implied by the model (Balagtas *et al.*, 1995). In the more common direct method described here, the key step is the calculation of the inverse mapping and its derivative matrix, i.e. for given $\boldsymbol{\eta}$ it is necessary to invert the equation $\boldsymbol{\eta} = C^T \log(L\boldsymbol{\pi})$ to obtain $\boldsymbol{\pi}$ in terms of $\boldsymbol{\eta}$. The derivative matrix is then given by $(\partial\pi/\partial\eta) = (C^T D^{-1} L)^{-1}$, where $D = \mathrm{diag}(L\boldsymbol{\pi})$.

In inverting equation (2), it is also necessary to ensure that $\boldsymbol{\pi} > 0$ and, to do so, we work with $\boldsymbol{\nu} = \log \boldsymbol{\pi}$, i.e. we seek to solve for $\boldsymbol{\nu}$ in the equation $\boldsymbol{\eta} = C^T \log(L \exp \boldsymbol{\nu})$. Excluding the special cases $d = 1$ or $d = 2$ with $r_1 = r_2 = 2$ as discussed in Palmgren (1989), no explicit solution is available, so an iterative method must be used. In particular, the Newton–Raphson iterations can be easily applied as described below.

(a) Begin with an initial approximation $\boldsymbol{\nu}_0$.
(b) Then take

$$\boldsymbol{\nu}_n = \boldsymbol{\nu}_{n-1} - \{C^T D_{n-1}^{-1} L\, \mathrm{diag}(\boldsymbol{\nu}_{n-1})\}^{-1} \{C^T \log(L \exp \boldsymbol{\nu}_{n-1}) - \boldsymbol{\eta}\},$$

where $D_{n-1} = \mathrm{diag}(L \exp \boldsymbol{\nu}_{n-1})$, and iterate until convergence.

Although theorem 1 shows that the mapping $\boldsymbol{\pi} \mapsto \boldsymbol{\eta}$ is invertible, it must be kept in mind that the range of the mapping is usually a proper subset of $\mathbf{R}^k$, where $k = \Pi_i r_i$. Thus, for certain values of $\boldsymbol{\eta}$, no positive solution $\boldsymbol{\pi}$ to equation (2) exists. This may occur for at least two reasons.

First, when ordinal variables are involved, certain monotonicity requirements are placed on $\boldsymbol{\eta}$. For example, with a single ordinal variable having three levels, we find

$$\boldsymbol{\eta} = (\log \pi_{.},\ \mathrm{logit}(\pi_1),\ \mathrm{logit}(\pi_1 + \pi_2))^T$$

and since $\pi_2 \geqslant 0$ we must have $\eta_3 \geqslant \eta_2$.

Second, if there are more than two variables it can also happen that no solution exists because of incompatibility of the lower dimensional marginals. For example, the requirement that the variance matrix of the d categorical variables be non-negative definite places certain restrictions on the one- and two-dimensional marginal probabilities which, in turn, imply restrictions on the first- and second-order logistic contrasts. In fact, the restrictions on the marginal probabilities do not arise solely from moment considerations and a detailed discussion may be found in Darroch (1962) or Glonek *et al.* (1988). Unfortunately, no readily computable criterion, for determining whether a particular $\boldsymbol{\eta}$ is valid, is available.

4. MULTIVARIATE LOGISTIC REGRESSION MODELS

Multivariate logistic regression models are now defined to be those of the form

$$\boldsymbol{\eta} = X\boldsymbol{\beta} \tag{3}$$

where X is a $k \times p$ matrix of constants and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters. With the preceding definition of the multivariate logistic transform, η_1 is always 0 and therefore we assume that the elements of the first row of X are also 0. In practice it is more convenient simply to omit η_1 and the corresponding rows of $(\partial\pi/\partial\eta)$ and X.

Preliminary to fitting a multivariate logistic regression model to a given set of data, it is necessary to compute $\boldsymbol{\pi}$ as a function of $\boldsymbol{\beta}$ and also to evaluate the derivative matrix $(\partial\pi/\partial\beta) = (C^{\mathrm{T}}D^{-1}L)^{-1}X$. Given the method of computing $\boldsymbol{\pi}$ from $\boldsymbol{\eta}$, the computation of $\boldsymbol{\pi}$ from $\boldsymbol{\beta}$ is straightforward.

Consider a table of multinomial frequencies $\mathbf{y}_i \sim M(n_i, \boldsymbol{\pi}_i)$, where $C^{\mathrm{T}} \log(L\boldsymbol{\pi}_i) = X_i\boldsymbol{\beta}$. The log-likelihood is

$$l(\boldsymbol{\beta};\, \mathbf{y}_i) = \mathbf{y}_i^{\mathrm{T}} \log \boldsymbol{\pi}_i,$$

the score vector is

$$\mathbf{s}(\boldsymbol{\beta};\, \mathbf{y}_i) = \left(\frac{\partial\pi}{\partial\beta}\right)^{\mathrm{T}} \operatorname{diag}(\boldsymbol{\pi}_i)^{-1}\mathbf{y}_i$$

and the information matrix is

$$\mathscr{I}_i(\beta) = n_i\left(\frac{\partial\pi}{\partial\beta}\right)^{\mathrm{T}} \operatorname{diag}(\boldsymbol{\pi}_i)^{-1}\left(\frac{\partial\pi}{\partial\beta}\right).$$

Given m independent observations $(\mathbf{y}_1, n_1, X_1)$, $(\mathbf{y}_2, n_2, X_2)$, $\ldots$, $(\mathbf{y}_m, n_m, X_m)$ the log-likelihood is $l(\boldsymbol{\beta}) = \Sigma\, l(\boldsymbol{\beta}; \mathbf{y}_i)$, the score vector is $\mathbf{s}(\boldsymbol{\beta}) = \Sigma\, \mathbf{s}(\boldsymbol{\beta}, \mathbf{y}_i)$ and the information matrix is $\mathscr{I}(\boldsymbol{\beta}) = \Sigma \mathscr{I}_i(\boldsymbol{\beta})$. Using these formulae, $\boldsymbol{\beta}$ can be estimated by using the Fisher scoring algorithm.

Observations with incomplete responses can readily be incorporated into the analysis. In particular, if some subset of the response variables $A_{i_1}, A_{i_2}, \ldots, A_{i_{d_0}}$ is recorded for a particular unit, then the probability distribution on that d_0-dimensional marginal table is multinomial and, by upward compatibility, a multivariate logistic regression model applies to the table of probabilities. Furthermore, the design matrix relating the marginal probabilities to $\boldsymbol{\beta}$ is constructed by selecting the appropriate rows of the full design matrix that would be used if complete data were available for that unit.

5. MODEL SPECIFICATION AND RESTRICTIONS

Although the formulation $\boldsymbol{\eta} = X\boldsymbol{\beta}$ is completely general, the explicit construction of X can be tedious. It is therefore useful to consider a symbolic specification of common subclasses of models.

When no structure is assumed on the responses $A_1, A_2, \ldots, A_d$, it is natural to consider models based on a notion of interaction analogous to that used in the analysis of factorial experiments. However, simple model formulae of the type introduced by Wilkinson and Rogers (1973) are not suitable. For this reason, we use the syntax of McCullagh and Nelder (1989), pages 222–223, where the linear predictor for each component of the multivariate logistic transform is specified separately using conventional notation.

To illustrate, consider two ordinal variables A and B with r and s levels respectively and let $\gamma_{ij} = \mathrm{pr}(A \leqslant i, B \leqslant j)$. In this situation we might consider the regression model

$$M\colon \operatorname{logit}(\gamma_{is}) = \theta_i + \boldsymbol{\beta}_i^{\mathrm{T}}\mathbf{x}, \qquad \operatorname{logit}(\gamma_{rj}) = \phi_j + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \qquad \log\left\{\frac{\gamma_{ij}(1-\gamma_{is}-\gamma_{rj}+\gamma_{ij})}{(\gamma_{is}-\gamma_{ij})(\gamma_{rj}-\gamma_{ij})}\right\} = \alpha. \tag{4}$$

This model is most naturally thought of as the conjunction of three separate models: one for the marginal distribution of A, one for the marginal distribution of B and one for the set of global odds ratios; Dale (1986). This is made explicit in the symbolic representation

$$A\colon A + A.\mathbf{x}; \qquad B\colon B + \mathbf{x}; \qquad A.B;. \tag{5}$$

The fact that the components of $\boldsymbol{\eta}$ are not variation independent introduces some restrictions on the interpretation of the models. In particular, it is pertinent to ask whether there are model formulae for which no table of positive probabilities $\boldsymbol{\pi}$ exists for *any* possible value of the parameters. For nominal variables, no such model exists because setting all parameters to 0 yields uniform probabilities across the cells of the contingency table. However, when ordinal variables are involved there are models for which this can occur. For example, with an ordinal variable A, the null model

$$\operatorname{logit}(\gamma_i) = \theta, \qquad i = 1, 2, \ldots, r-1,$$

implies that $\pi_2 = \pi_3 = \ldots = \pi_{r-1} = 0$, irrespective of θ. For this reason it is sensible to require that the minimal model for a single ordinal variable A be $A\colon A$. When $A_1, A_2, \ldots, A_{d_0}$ are ordinal the minimal model is

$$A_1\colon A_1; \qquad A_2\colon A_2; \;\ldots; \qquad A_{d_0}\colon A_{d_0};.$$

Additional constraints on the log-odds ratios and higher order interactions are not required because setting all those quantities to 0 yields the independence model.

Even with this restriction, it will be the case, for many useful models, that no table of probabilities $\boldsymbol{\pi}$ satisfying $\boldsymbol{\eta} = C^{\mathrm{T}} \log(L\boldsymbol{\pi})$ exists for certain values of X and β. This can complicate the interpretation of such models when used for prediction because it may then happen that certain values of the predictor variable do not produce valid values of $\boldsymbol{\eta}$. The first component of model (4) provides a simple example of this. On the numerical side, it may also happen that intermediate values of $\hat{\boldsymbol{\beta}}$ chosen by the Fisher scoring algorithm lead to inconsistent values for $\boldsymbol{\eta}$ thus preventing completion of the calculations. However, we have not observed this with the data sets that we have studied. We surmise that it is very unlikely when a well fitting model is applied to adequate data with good initial estimates.

An important application of multivariate logistic models is to longitudinal categorical data. Suppose now that $A_1, A_2, \ldots, A_d$ are repeated measurements of the same variable A taken at times $t_1 < t_2 < \ldots < t_d$. Assume for simplicity that A is dichotomous. For any l-tuple $\mathbf{i} = (i_1, i_2, \ldots, i_l)$ of $(1, 2, \ldots, d)$ let $\boldsymbol{\pi}_{\mathbf{i}}$ denote the l-dimensional marginal table of probabilities for $A_{i_1}, A_{i_2}, \ldots, A_{i_l}$ and then let

$$\eta_{i_1} = \operatorname{logit}\{\pi_{i_1}(1)\},$$

$$\eta_{i_1 i_2} = \log\left\{\frac{\pi_{i_1 i_2}(1, 1)\, \pi_{i_1 i_2}(0, 0)}{\pi_{i_1 i_2}(1, 0)\, \pi_{i_1 i_2}(0, 1)}\right\},$$

$$\eta_{i_1 i_2 i_3} = \log\left\{\frac{\pi_{i_1 i_2 i_3}(1, 1, 1)\, \pi_{i_1 i_2 i_3}(1, 0, 0)\, \pi_{i_1 i_2 i_3}(0, 1, 0)\, \pi_{i_1 i_2 i_3}(0, 0, 1)}{\pi_{i_1 i_2 i_3}(1, 1, 0)\, \pi_{i_1 i_2 i_3}(1, 0, 1)\, \pi_{i_1 i_2 i_3}(0, 1, 1)\, \pi_{i_1 i_2 i_3}(0, 0, 0)}\right\},$$

and so on. In this context, it is of interest to consider models such as

$$\left.\begin{aligned} \eta_{i_1} &= \alpha_{i_1} + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} \\ \eta_{i_1 i_2} &= \gamma(|t_{i_1} - t_{i_2}|), \\ \eta_{i_1 i_2 i_3} &= \delta, \end{aligned}\right\} \tag{6}$$

and so on. Such models fall into the general class of multivariate logistic regression models (3) but cannot be specified by using model formulae of the form (5). Often the models that are appropriate for longitudinal data have the following properties.

(a) The parametric form of the model for $\eta_{i_1 i_2 \ldots i_l}$ depends only on l.
(b) The parametric form of the model for $\eta_{i_1 i_2 \ldots i_l}$ is symmetric in $(i_1, i_2, \ldots, i_l)$.
(c) Considering all l-dimensional terms, the parameters can vary in one of the following ways:
   (i) $[c]$—the parameter values are constant over all terms $\eta_{\mathbf{i}}$, with $i_1 < i_2 < \ldots < i_l$;
   (ii) $[s]$—the parameter values are the same in $\eta_{\mathbf{i}}$ and $\eta_{\mathbf{i}'}$ if $(t_{i'_1}, t_{i'_2}, \ldots, t_{i'_l}) = (t_{i_1} + c, t_{i_2} + c, \ldots, t_{i_l} + c)$ for some c;
   (iii) $[u]$—the parameter values are unrestricted.

A class of models that satisfy these criteria can be specified by model formulae of the form

$$1{:}\, f_1; \qquad 2{:}\, f_2; \qquad \ldots; \qquad K{:}\, f_K;$$

where each f_l is the model formula for all of the l-dimensional terms, $\eta_{i_1 i_2 \ldots i_l}$. The f_l themselves can include external factors and covariates in the same way as a standard Wilkinson–Rogers-type formula together with the symbols [c], [s], [u] and numbers indicating the highest order 'interaction' terms between $A_{i_1}, A_{i_2}, \ldots, A_{i_l}$ that would appear if that part of the model were specified in the notation of expression (5).

For example, model (6) is expressed in this notation as

$$1\colon\ [u] + x; \qquad 2\colon\ [s]; \qquad 3\colon\ [c];.$$

To provide a fuller illustration, suppose that A is polytomous and that we wish to consider a model

$$\begin{aligned} \eta_{i_1}(j_1) &= \alpha_{i_1}(j_1) + \beta x, \\ \eta_{i_1 i_2}(j_1, j_2) &= \gamma_{|t_{i_1} - t_{i_2}|}(j_1, j_2), \\ \eta_{i_1 i_2 i_3}(j_1, j_2, j_3) &= \delta_{12}(j_1, j_2) + \delta_{13}(j_1, j_3) + \delta_{23}(j_2, j_3) \qquad \text{for } i_1 < i_2 < i_3 \end{aligned}$$

and so on. This model can be expressed equivalently by the model formula

$$1\colon\ [u].1 + [c].x; \qquad 2\colon\ [s].2; \qquad 3\colon\ [c].2;.$$

It is worth making some further remarks on the interpretation of the symbols [c], [s] and [u]. These can be thought of as representing multilevel factors across the space of 'units' $\{\mathbf{i}\colon |\mathbf{i}| = l\}$. In this case, [$c$] represents the vector of 1s, [s] represents the levels defined by the equivalence relation $\mathbf{i} \sim \mathbf{i}'$ if $t_{\mathbf{i}} = t_{\mathbf{i}'} + c$ and [u] represents the case where each $\mathbf{i}$ is a distinct level. Finally, we note that the use of the three symbols [c], [s] and [u] introduces some redundancy in the sense that any two are sufficient if it is agreed that the third is to be the default. In what follows we adopt the convention that [c] is implicit in any expression where [s] and [u] are absent.

6. INFORMATION MATRIX

For two binary variables, McCullagh and Nelder (1989) gave the form of the information matrix as

$$\mathscr{I}(\eta_a, \eta_b, \eta_{ab}) = m \begin{pmatrix} V_a/\Delta & \Delta_\pi/\Delta & 0 \\ \Delta_\pi/\Delta & V_b/\Delta & 0 \\ 0 & 0 & V_{ab} \end{pmatrix}$$

where

$$V_a = \pi_{1.}\pi_{2.}, \qquad V_b = \pi_{.1}\pi_{.2}, \qquad V_{ab} = \left(\frac{1}{\pi_{11}} + \frac{1}{\pi_{12}} + \frac{1}{\pi_{21}} + \frac{1}{\pi_{22}}\right)^{-1},$$

$$\Delta_\pi = \pi_{12}\pi_{21} - \pi_{11}\pi_{22}, \qquad \Delta = \frac{V_a V_b - \Delta_\pi^2}{V_a V_b}.$$

If several units are observed and the multivariate logistic regression model

$$\boldsymbol{\eta}_a = X\boldsymbol{\beta}_a, \qquad \boldsymbol{\eta}_b = X\boldsymbol{\beta}_b, \qquad \boldsymbol{\eta}_{ab} = X\boldsymbol{\beta}_{ab}$$

is considered then the information matrix takes the form

$$i(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b, \boldsymbol{\beta}_{ab}) = \begin{pmatrix} X^{\mathrm{T}}D_1X & X^{\mathrm{T}}D_{12}X & 0 \\ X^{\mathrm{T}}D_{12}X & X^{\mathrm{T}}D_2X & 0 \\ 0 & 0 & X^{\mathrm{T}}D_3X \end{pmatrix}$$

where $D_1 = \mathrm{diag}(mV_a/\Delta)$, $D_2 = \mathrm{diag}(mV_b/\Delta)$, $D_{12} = \mathrm{diag}(m\Delta_\pi/\Delta)$ and $D_3 = \mathrm{diag}(mV_{ab})$.

One question raised by the multivariate logistic regression model concerns the potential gains in efficiency that they offer when compared with the marginal estimates of the same parameters. As was noted by McCullagh and Nelder (1989), the parameter $\boldsymbol{\beta}_{ab}$ is orthogonal to the marginal regression parameters $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$ and has the same variance as the log-linear estimate irrespective of $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$. Moreover, this is true irrespective of the form of the linear models that $\boldsymbol{\beta}_a$ and $\boldsymbol{\beta}_b$ are assumed to satisfy.

However, $\boldsymbol{\beta}_a$ and $\boldsymbol{\beta}_b$ are not orthogonal and therefore there is scope for gains in efficiency in estimating these parameters. The previously mentioned orthogonality shows that the estimation of $\boldsymbol{\beta}_{ab}$ has no bearing on the asymptotic variance matrix of $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$. Consider now the parameter $\boldsymbol{\beta}_a$ with estimates based on the full likelihood and the marginal likelihood. The asymptotic variance matrices of these two estimates are respectively

$$\{X^{\mathrm{T}}(D_1 - D_{12}X(X^{\mathrm{T}}D_2X)^{-1}X^{\mathrm{T}}D_{12})X\}^{-1} \quad \text{and} \quad (X^{\mathrm{T}}D_0X)^{-1},$$

where $D_0 = \mathrm{diag}(mV_a) = D_1 - D_{12}D_2^{-1}D_{12}$. Unfortunately these matrices cannot be inverted explicitly so an analytical comparison cannot be performed.

In the case of two binary responses, bounds on the gain in efficiency can be obtained by observing that the matrix $(X^{\mathrm{T}}D_1X)^{-1}$ is a lower bound for the asymptotic variance matrix of the estimate of $\boldsymbol{\beta}_a$ based on the multivariate logistic model. Thus the possible gain in efficiency is limited by Δ. Bounds for Δ in a single 2×2 contingency table $\boldsymbol{\pi} = \{\pi_{ij}\}$ can be found as follows. Taking $\pi_{\mathrm{l}} = \min(\pi_{1.}, \pi_{.1})$, $\pi_{\mathrm{u}} = \max(\pi_{1.}, \pi_{.1})$, $\bar{\pi}_{\mathrm{l}} = \min(\pi_{1.}, \pi_{.2})$ and $\bar{\pi}_{\mathrm{u}} = \max(\pi_{1.}, \pi_{.2})$ it can be shown that

$$\Delta \geqslant 1 - \max\left\{\left(\frac{\pi_{\mathrm{l}}}{1-\pi_{\mathrm{l}}}\right)\bigg/\left(\frac{\pi_{\mathrm{u}}}{1-\pi_{\mathrm{u}}}\right), \left(\frac{\bar{\pi}_{\mathrm{l}}}{1-\bar{\pi}_{\mathrm{l}}}\right)\bigg/\left(\frac{\bar{\pi}_{\mathrm{u}}}{1-\bar{\pi}_{\mathrm{u}}}\right)\right\}.$$

This shows, as might be expected, that large gains in efficiency cannot occur unless the marginals are such that most of the probability is concentrated either in the diagonal or in the counter-diagonal cells of the table. The potential for variance reduction also depends on the strength of the association between A and B and if $\psi = \pi_{11}\pi_{22}/\pi_{12}\pi_{21}$ then it can be shown that $\Delta \geqslant 4\sqrt{\psi}/(1+\sqrt{\psi})^2$. However, practical experience is that the bounds given above tend to be very generous and actual improvements are usually very small. This is particularly so for regression models of the form 'A: **x**; B: **x**; $A.B$: **x**;', which suggests that the potential gain in efficiency may be bounded in that situation. In fact, this is not the case. An example of such a regression model where the potential gain in efficiency is not theoretically bounded is constructed in Glonek and McCullagh (1994). However, in that example the gains in efficiency are negligible for values of the log-odds ratios that are typically found in practice.

The analysis for two binary responses is greatly simplified by the block diagonal form of the information matrix. Unfortunately, the generalization to higher dimensions and polytomous variables is less useful. The following result is proved in Glonek and McCullagh (1994).

Theorem 2. Suppose that $A_1, A_2, \ldots, A_d$ are nominal variables with $r_1, r_2, \ldots, r_d$ levels respectively. Then the information matrix, for a single observation, has the form

$$\mathscr{I} = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & B_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & B_2 \end{pmatrix},$$

where B_1 has dimension $(k - k_1 - 1) \times (k - k_1 - 1)$ and B_2 has dimension $k_1 \times k_1$ with $k = \Pi_i r_i$ and $k_1 = \Pi_i (r_i - 1)$.

In fact, the matrix $V = \mathscr{I}^{-1}$ contains other zero entries corresponding to 'orthogonality' between each other interaction term and all subordinate interactions and main effects. This can be seen by noting that V is just the large sample variance matrix of the multivariate logistic transform of a table of multinomial proportions and applying the previous result to all the lower dimensional marginal tables to verify this claim. However, those other entries do not produce a finer block diagonal structure in V than that prescribed by theorem 2 and do not produce any other zero entries in the information matrix. Consequently, they do not have an obvious effect on the structure of the large sample variance matrix for the parameter estimates in a multivariate regression problem. For example if A, B and C are binary variables we can consider, as previously, a bivariate logistic model for A and B,

$$A\colon \mathbf{x};\quad B\colon \mathbf{x};\quad A.B\colon \mathbf{x};, \tag{7}$$

and also a trivariate logistic model,

$$A\colon \mathbf{x};\quad B\colon \mathbf{x};\quad A.B\colon \mathbf{x};\quad C\colon \mathbf{x};\quad A.C\colon \mathbf{x};\quad B.C\colon \mathbf{x};\quad A.B.C\colon \mathbf{x};. \tag{8}$$

Now, model (8) implies model (7), and, as has been discussed previously, the estimates of $\boldsymbol{\beta}_a$ and $\boldsymbol{\beta}_b$ are asymptotically uncorrelated with the estimate $\boldsymbol{\beta}_{ab}$ when model (7) is fitted. However, if model (8) is fitted then $\hat{\boldsymbol{\beta}}_{abc}$ is asymptotically uncorrelated with all other parameter estimates, but, excluding special cases such as $\boldsymbol{X} = \mathbf{1}$, estimates of $(\boldsymbol{\beta}_a, \boldsymbol{\beta}_b)$ and $\boldsymbol{\beta}_{ab}$ are correlated.

When some of the variables are ordinal, theorem 2 does not hold although the information matrix can be seen to have the form

$$\mathscr{I} = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathrm{B} \end{pmatrix}.$$

Because the multivariate logistic transform for ordinal variables can be derived by applying the multivariate logistic transform for nominal variables to tables formed by dichotomizing the ordinal variables at various cutpoints, theorem 2 can be used to show that the inverse information matrix has other zero entries. However, these entries do not imply a block diagonal structure and thus appear to have little significance.

7. EXAMPLES

7.1. *Six Cities Data*

We consider the subset of the data from the six cities study on the health effects of air pollution by Ware *et al.* (1984) that has subsequently been analysed by Zeger *et al.* (1988), and more recently by Fitzmaurice and Laird (1993). These data consist of an annual binary response indicating the presence or absence of wheeze at ages 7, 8, 9 and 10 years for each of 537 children from Stuebenville, Ohio. The only explanatory variable is a binary indicator variable for the mother's smoking habits during the first year of the study. The data are given in Table 1 of Fitzmaurice and Laird (1993).

A sequence of multivariate logistic models was fitted to the data, beginning with the saturated model and subsequently eliminating the highest order non-significant interaction terms. In the model formulae in Table 1, the symbol 1 refers to the marginal log-odds for wheeze, the symbol 2 refers to bivariate marginal log-odds ratios, 3 refers to trivariate logistic contrasts, and so on. Thus, in the model '(1; 2): $[u].S$; 3: S;', the fourth-order logistic contrast is 0, the third-order contrasts depend on the mother's smoking habits and the lower order contrasts are unrestricted. In the final model, the third- and fourth-order contrasts are 0, the log-odds ratios are constant for each pair of time points and the log-odds for wheeze has a time-dependent intercept with a constant effect of mother's smoking habit.

The fitted parameters from the final model, '1: $[u] + S$; 2;', are given in Table 2. The positive coefficient of '1: $S(2)$' means that the estimated odds for wheeze is higher by a factor of $\exp 0.27 = 1.31$ for children of smoking mothers ($S = 2$) than for children of non-smoking mothers ($S = 1$). The deviance for this model is 17.27 on 24 degrees of freedom and the positive coefficient for maternal smoking, although not statistically significant, is in broad agreement with the previous analyses of Zeger *et al.* (1988) and Fitzmaurice and Laird (1993).

To make a closer comparison with those analyses we also fitted the model, '1: AGE $*$ S; 2;' where AGE is the age in years since the child's ninth birthday. The deviance for this model is 16.76 on 25 degrees of freedom and the parameter estimates with their standard errors are given in Table 3. The parameter estimates and standard errors of the one-dimensional marginal parameters agree, for all practical purposes, with those given by Fitzmaurice and Laird (1993) but the

TABLE 1
Deviances for a sequence of models

Model	*Incremental deviance*	*Degrees of freedom*
(1; 2; 3; 4): $[u].S$;	–	–
(1; 2; 3): $[u].S$; 4: $[u]$;	0.008	1
(1; 2; 3): $[u].S$;	0.41	1
(1; 2): $[u].S$; 3: S;	2.73	4
(1; 2): $[u].S$;	0.40	4
1: $[u].S$; 2: $[u]$;	0.66	6
1: $[u].S$; 2: $[s]$;	2.19	3
1: $[u].S$; 2;	3.96	2
1: $[u] + S$; 2;	1.77	3

TABLE 2
Parameter estimates in the model 1: [u] + S; 2;

Order	*Parameter*	*Estimate*	*Standard error*
1:	$[u](1)$	−1.76	0.136
	$[u](2)$	−1.62	0.133
	$[u](3)$	−1.76	0.136
	$[u](4)$	−2.12	0.150
	$S(2)$	0.271	0.178
2:	(Intercept)	2.05	0.173

TABLE 3
*Parameter estimates in the model 1: AGE*S; 2;*

Order	*Parameter*	*Estimate*	*Standard error*
1:	(Intercept)	−1.894	0.116
	AGE	−0.131	0.056
	S	0.306	0.186
	AGE.S	0.062	0.088
2:	(Intercept)	2.032	0.173

marginal log-odds ratio used here is not directly comparable with the conditional log-odds ratio given in that paper. The parameter estimates obtained by Zeger *et al.* (1988) are likewise similar and, although their model of exchangeable marginal correlations is not the same as the model of exchangeable log-odds ratios, the values obtained are consistent. In particular, the common marginal correlation was estimated to be 0.346 in Zeger *et al.* (1988) and the marginal correlation for each two-way marginal table can also be calculated from the parameters of the common log-odds ratio model used here. For example, in the 2×2 cross-tabulation of response at ages 7 and 8 years without exposure to maternal smoking, we calculate the correlation to be 0.357.

Ordinarily, in longitudinal data of this sort, we might expect that the marginal association or correlation between observations at times t_1 and t_2 would decrease with $|t_1 - t_2|$. However, the fitted log-odds ratios in the models 2: $[u]$ and 2: $[s]$ show no evidence of such a pattern. The second model shows that the incidence declines with age at an estimated annual rate of 12%. The high value for the log-odds ratio (2.05) indicates that wheeze is strongly persistent over this age range, i.e. that the variability between children is much greater than the variability between time points for the same child. To say the same thing in another way, individual susceptibilities are highly variable but they decline slowly over time in an essentially deterministic way.

7.2. *Clinical Trial for Skin Disorder*

Koch *et al.* (1991) presented data from a certain clinical trial which, for confidentiality, was fictitiously described as pertaining to the treatment of a skin disorder. The study comprised 72 subjects of whom 36 received the treatment and 36 received a placebo. An ordinal response variable ($0 \equiv$ excellent, $1 \equiv$ good, $2 \equiv$ fair, $3 \equiv$ poor) was recorded for each subject on four occasions, 3 days after treatment, 7 days after treatment, 10 days after treatment and 14 days after treatment. The data are given in Koch *et al.* (1991), Table 9-2, p. 228. Because response category 3 was observed only twice, the adjacent categories 2 and 3 were amalgamated to form a response variable with three ordered categories ($0, 1, \geqslant 2$).

When the data are represented as two 3^4 contingency tables the cell frequencies are very small. In fact, 128 of the 162 cells are empty. For this reason it is not possible to model the higher order interaction terms and so '1: 1.[*u*].*X*; 2: 2;', with *X* indicating treatment, is the most complicated model considered. Since the response categories are ordered, the marginal logits are cumulative. Consequently all bivariate contrasts are of the global type as in Dale (1986) or Plackett (1965). The term '2: 2;' in the model formula implies that the four global cross-ratios are the same for all the two-dimensional marginal tables, and the term '1: 1.[*u*].*X*;' implies that the one-dimensional marginal probabilities are unrestricted. This initial model was simplified by the elimination of the highest order non-significant interaction terms as summarized in Table 4. The first simplification is the reduction in the global cross-ratio term to '2;' which signifies that the four global cross-ratios are equal, essentially a Plackett distribution with constant cross-ratio.

Although the incremental deviance for the final model is marginally significant ($p = 0.041$) we adopted it on account of its simple interpretation. The parameter estimates for that model are given in Table 5. The two parameters, 1(1) and 1(2),

TABLE 4
Deviances for a sequence of models

Model	*Incremental deviance*	*Degrees of freedom*
1: 1.[*u*].*X*; 2: 2;	—	—
1: 1.[*u*].*X*; 2;	1.62	3
1: 1.[*u*] + 1.*X* + [*u*].*X*; 2;	2.64	3
1: 1.[*u*] + 1.*X*; 2;	1.36	3
1: 1.[*u*] + *X*; 2;	0.63	3
1: 1 + [*u*] + *X*; 2;	8.25	3

TABLE 5
Parameter estimates in the model 1: 1 + [*u*] *+ X; 2;*

Order	*Parameter*	*Estimate*	*Standard error*
1:	1(1)	−3.236	0.377
	1(2)	−0.419	0.287
	[*u*](2)	1.353	0.279
	[*u*](3)	1.376	0.280
	[*u*](4)	2.443	0.313
	X	1.242	0.343
2:	(Intercept)	1.485	0.286

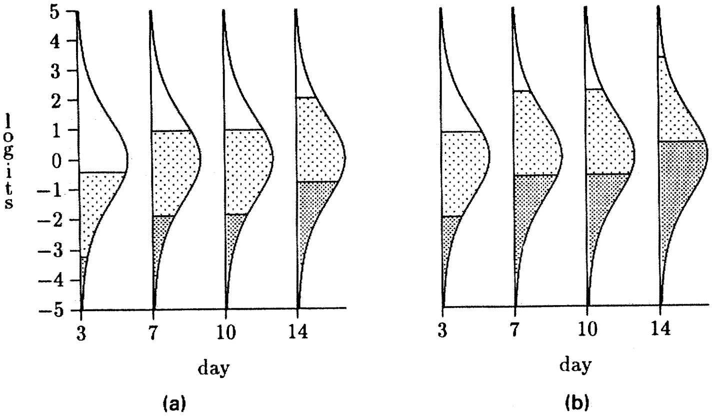

Fig. 1. Diagram of the final fitted model: (a) controls; (b) treatment (□, fair–poor; ⚄, good; ▩, excellent)

give the base-line logits, for day 3 with no treatment, of scoring 0 and 1 or less respectively. The parameters [*u*](2), [*u*](3) and [*u*](4) show the increase in both of these logits over time. The parameter X shows the change in these logits resulting from the treatment which in this case is positive and highly statistically significant. The additive nature of the model for the marginal probabilities is illustrated in Fig. 1. The common global cross-ratio parameter indicates a fairly strong positive association between the responses at different times.

7.3. *Computations*

All computations were programmed in C. At the time of writing the models described in Section 5 had not been fully implemented but it is envisaged that the computer programs will be available from the first author.

ACKNOWLEDGEMENTS

We are grateful to Dr A. J. Branford for useful discussions. The manuscript was prepared and the computations were performed on equipment partially funded by Flinders University Board of Research.

REFERENCES

Balagtas, C. C., Becker, M. P. and Lang, J. B. (1995) Marginal modelling of categorical data from crossover experiments. *Appl. Statist.*, **44**, 63–77.

Dale, J. R. (1986) Global cross-ratio models for bivariate discrete ordered responses. *Biometrics*, **42**, 909–917.

Darroch, J. N. (1962) Interaction in multi-factor contingency tables. *J. R. Statist. Soc.* B, **24**, 251–263.

Fitzmaurice, G. M. and Laird, N. M. (1993) A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141–151.

Glonek, G. F. V., Darroch, J. N. and Speed, T. P. (1988) On the existence of maximum likelihood estimators for hierarchical loglinear models. *Scand. J. Statist.*, **15**, 187–193.

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

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

Koch, G., Singer, J., Stokes, M., Carr, G., Cohen, S. and Forthofer, R. (1991) Some aspects of weighted least squares analysis for longitudinal categorical data. In *Statistical Models for Longitudinal Studies of Health* (eds J. H. Dwyer, M. Feinlib, P. Lippert and H. Hoffmeister), pp. 215–258. Oxford: Oxford University Press.

Liang, K.-Y., Zeger, S. L. and Qaqish, B. (1992) Multivariate regression analyses for categorical data (with discussion). *J. R. Statist. Soc.* B, **54**, 3–40.

McCullagh, P. (1989) Models for discrete multivariate responses. *Bull. Int. Statist. Inst.*, **53**, 407–418.

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

Palmgren, J. (1989) Regression models for bivariate binary responses. *Technical Report 101*. Department of Biostatistics, University of Washington, Seattle.

Plackett, R. L. (1965) A class of bivariate distributions. *J. Am. Statist. Ass.*, **60**, 516–522.

Ware, J. H., Dockery, D. W., Spiro III, A., Speizer, F. E. and Ferris, Jr, B. G. (1984) Passive smoking, gas cooking and respiratory health in children living in six cities. *Am. Rev. Resp. Dis.*, **129**, 366–374.

Wilkinson, G. N. and Rogers, C. E. (1973) Symbolic description of factorial models for analysis of variance. *Appl. Statist.*, **22**, 392–399.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.