Biometrika (2001), **88**, 1, *pp*. 287–290
© 2001 Biometrika Trust
Printed in Great Britain

Modelling and generating correlated binary variables

By SAMUEL D. OMAN and DAVID M. ZUCKER
Department of Statistics, Hebrew University, Mount Scopus, Jerusalem, 91905 Israel
oman@mscc.huji.ac.il mszucker@mscc.huji.ac.il

Summary

Many applications use simple parametric models for the correlation structure of binary responses which are observed in clusters. The usual approach, to use correlation models appropriate for normally distributed responses, suffers from two drawbacks when the marginal probabilities within the clusters differ. First, as it does not explicitly take into account constraints on the second moments which must be satisfied for binary responses, it may fail to model realistically the range of correlations present in the data. Secondly, computer simulation of observations from these models is very difficult. We present an alternative class of correlation models which reflect the binary nature of the responses and allow for simple simulation.

Some key words: Binary variable; Computer simulation; Correlation structure; Generalised estimating equation.

1. Introduction

Many applications involve binary responses Y_i which are dependent, as they are observed in clusters. Often, instead of considering the full dependence structure, one uses a simple parametric model for the correlations between responses; these correlations are typically positive because of cluster effects or time dependence.

The usual approach uses models, such as intraclass, autoregressive or moving average models, which are commonly used with normally distributed responses; for convenience, we shall refer to these as 'normal models'. When the probabilities $p_i = E(Y_i)$ within a cluster differ, however, this gives rise to two problems. First, normal models do not explicitly take into account constraints involving the first and second moments which must be satisfied for binary responses (Bahadur, 1961; Prentice, 1988), and consequently they may not realistically model the range of correlations present in the data. The importance of accurately modelling correlations when using generalised estimating equations (Liang & Zeger, 1986; Zeger & Liang, 1986) is well known (Crowder, 1995; Sutradhar & Das, 1999). Secondly, it is extremely difficult to simulate observations from these models; see Lunn & Davies (1998) for a partial solution. Both problems arise because there is no natural, simple mechanism for generating binary variables with normal covariance structures when the p_i are unequal.

In this paper we propose a simple, constructive technique for defining binary variables with given marginals p_i and a variety of simple parametric correlation structures. As these structures correspond to rigorously defined joint distributions, they perforce take into account the binary nature of the responses. The correlation matrices are analogous to normal-model matrices, and reduce to them when the p_i within a cluster are equal. Also, the constructive definition of the variables makes their computer generation extremely simple. Finally, if the p_i are being modelled in terms of explanatory variables, the correlation matrices may be defined to correspond to the link function in a natural way, and their parameters may be easily estimated.

In the next section we define our models. We discuss data simulation in § 3 and parameter estimation in § 4.

2. A CLASS OF MODELS

Let $Y=(Y_1,\ldots,Y_n)$ denote the vector of binary responses for a cluster, and assume that $0<p_i<1$ for each i. We wish to define a matrix $R=R(\gamma)$, of simple parametric form, which is a valid correlation matrix in the sense that it corresponds to a rigorously defined joint distribution for Y which has the p_i as marginal probabilities.

To this end, denote $\mathrm{pr}(Y_i \times Y_j=1)$ by π_{ij}. Since $\pi_{ij}\leqslant \min(p_i,p_j)$, each correlation $r_{ij}=r_{ij}(\gamma)$ must satisfy

$$r_{ij}\leqslant \frac{[\min(p_i,p_j)/\{1-\min(p_i,p_j)\}]^{\frac{1}{2}}}{[\max(p_i,p_j)/\{1-\max(p_i,p_j)\}]^{\frac{1}{2}}}=\bar{r}_{ij}, \tag{2·1}$$

say. Observe that, if R is taken to be of intraclass form, the common correlation can be at most the minimum of the $\bar{r}_{ij}$, and thus very small, if the p_i in the cluster are very different from one another. Similar observations have been made by Heagerty & Zeger (1996), in the context of modelling correlations for clustered ordinal variables. Bahadur (1961) gives sufficient conditions for a candidate correlation matrix to be valid, but they are quite complicated and typically can be verified only if all the r_{ij} are sufficiently small.

Our models are of the form

$$r_{ij}(\gamma)=\bar{r}_{ij}c_{ij}(\gamma), \tag{2·2}$$

where, if we define $c_{ii}(\gamma)\equiv 1$, $C(\gamma)=(c_{ij}(\gamma))$ is a parametric normal-model correlation matrix. If the p_i in the cluster are all equal, then $\bar{r}_{ij}\equiv 1$ and (2·2) gives a normal-model matrix. If the p_i differ, then (2·2) extends these matrices in a natural way, essentially allowing each correlation to come as close as possible to the normal-model correlation as the pairwise constraints permit.

To define our models, first note that, if

$$\pi_{ij}=(1-v)p_ip_j+v\min(p_i,p_j) \tag{2·3}$$

for $0\leqslant v\leqslant 1$, then $r_{ij}=v\bar{r}_{ij}$. Observe that (2·3) defines a bivariate joint distribution which is a convex combination of that obtaining under independence and the distribution having maximum pairwise correlations, for the given marginals.

Now let F denote the cumulative distribution function of a continuous distribution, and define $\theta_i=F^{-1}(p_i)$. In what follows ε_i will denote independent variables distributed according to F, and U_i will denote independent Bernoulli variables with parameter γ $(0\leqslant\gamma\leqslant 1)$, which are independent of the ε_i.

To extend the intraclass model, we let

$$Y_i=1_{(Z_i\leqslant\theta_i)}, \tag{2·4}$$

where

$$Z_i=U_i\varepsilon_0+(1-U_i)\varepsilon_i. \tag{2·5}$$

This defines a joint distribution for all the Y_i in the cluster, and, since each $Z_i\sim F$, the definition of θ_i guarantees that the Y_i have the required marginal probabilities. Moreover, if $i\neq j$ then a simple calculation shows that (2·3) holds with $v=\gamma^2$, so that (2·2) obtains with $c_{ij}(\gamma)\equiv\gamma^2$. In particular, taking $\gamma=1$ shows that the matrix $(\bar{r}_{ij})$ of maximum pairwise correlations is a valid correlation matrix for the entire vector Y.

For the analogue of a MA(1) model, we keep (2·4) but replace (2·5) by

$$Z_i=U_i\varepsilon_{i-1}+(1-U_i)\varepsilon_i. \tag{2·6}$$

We now obtain

$$c_{ij}(\gamma)=\begin{cases}\gamma(1-\gamma), & \text{for } |i-j|=1,\\ 0, & \text{for } |i-j|>1.\end{cases}$$

Finally, for an AR(1) analogue we take $Z_1 \sim F$ independently of the U_i and ε_i, and set

$$Z_i = U_i Z_{i-1} + (1 - U_i)\varepsilon_i \quad (i \geqslant 2). \tag{2·7}$$

We have the following result.

LEMMA. *For any i, a and b, and* $l \geqslant 1$,

$$\mathrm{pr}(Z_i \leqslant a, Z_{i+l} \leqslant b) = \gamma^l \min\{F(a), F(b)\} + (1 - \gamma^l)F(a)F(b). \tag{2·8}$$

Proof. We proceed by induction. For $l = 1$, the left-hand side of (2·8) equals

$$\begin{aligned}\mathrm{pr}\{Z_i \leqslant a, U_{i+1}Z_i + (1 - U_{i+1})\varepsilon_{i+1} \leqslant b\} &= \mathrm{pr}(U_{i+1} = 0, Z_i \leqslant a, \varepsilon_{i+1} \leqslant b)\\ &\quad + \mathrm{pr}(U_{i+1} = 1, Z_i \leqslant a, Z_i \leqslant b)\\ &= (1 - \gamma)F(a)F(b) + \gamma \min\{F(a), F(b)\},\end{aligned}$$

as required. If we assume (2·8) for $l = k - 1$, a similar calculation gives

$$\begin{aligned}\mathrm{pr}(Z_i \leqslant a, Z_{i+k} \leqslant b) &= \mathrm{pr}(U_{i+k} = 0, Z_i \leqslant a, \varepsilon_{i+k} \leqslant b) + \mathrm{pr}(U_{i+k} = 1, Z_i \leqslant a, Z_{i+k-1} \leqslant b)\\ &= (1 - \gamma)F(a)F(b) + \gamma[\gamma^{k-1} \min\{F(a), F(b)\} + (1 - \gamma^{k-1})F(a)F(b)],\end{aligned}$$

proving the lemma. □

If we let $a = \theta_i$ and $b = \theta_{i+l}$ in (2·8) we obtain (2·3) with $j = i + l$ and $v = \gamma^l$, so that (2·2) holds with $c_{ij}(\gamma) = \gamma^{|i-j|}$.

We may obtain more complicated structures by using additional mixing Bernoulli variables with varying parameters, but we shall not pursue this here.

3. SIMULATION

Our Y_i are threshold variables for latent Z_i which are defined by a mixing procedure giving the desired covariance structure. This is very similar to the method of Lunn & Davies (1998) for efficiently generating variables with normal-family correlation structures when the probabilities $p_i \equiv p$ within a cluster. For example, to obtain an intraclass structure they define Y_i directly by

$$Y_i = U_i \varepsilon_0 + (1 - U_i)\varepsilon_i, \tag{3·1}$$

where the U_i are independent $\mathrm{Ber}(\gamma)$ variables and the ε_j are independent $\mathrm{Ber}(p)$ variables; compare with (2·5).

Unfortunately, this technique does not easily generalise when the p_i within a cluster differ. In this case, Lunn & Davies (1998) take $\varepsilon_j \sim \mathrm{Ber}(p)$, where $p = \max(p_i)$, generate $Y_i \sim \mathrm{Ber}(p)$ with the desired covariance structure, and then multiply the Y_i by independent $\mathrm{Ber}(p_i/p)$ variables. The resulting variables W_i have the desired marginal probabilities, but their correlation matrix is no longer of the desired normal form. For example, if the Y_i have an intraclass structure with correlation ρ, and $p_i \leqslant p_j$, Lunn & Davies' formula gives

$$\mathrm{corr}(W_i, W_j) = \frac{p_j/(1 - p_j)}{p/(1 - p)}\, \rho \bar{r}_{ij}. \tag{3·2}$$

The resulting correlation matrix is thus not of normal form, it does not allow for the maximum correlation possible when $p_j < p$, and it does not have an intuitive interpretation. In contrast, the method proposed here is equally straightforward when the p_i vary, and allows for maximal correlation.

4. USE WITH GENERALISED LINEAR MODELS

Our covariance structure may be incorporated into generalised linear modelling in a natural way. To obtain $g(p_i) = x_i^{\mathrm{T}}\beta$ for a link function g and column vector x_i of covariates, we simply take $F = g^{-1}$ and define $\theta_i = x_i^{\mathrm{T}}\beta$.

The covariance parameters may be estimated as part of an iterative procedure, as in Liang & Zeger (1986) or Carey et al. (1993). Specifically, define the binary variables $W_{ij} = (Y_j - Y_i)^2$, and note that, if Y_i and Y_j are in the same cluster and $p_i \leqslant p_j$, then (2·3) gives, when $v = c_{ij}(\gamma)$,

$$E(W_{ij}) = p_i(1-p_j) + p_j(1-p_i) + 2p_i(p_j - 1)c_{ij}(\gamma) = f_{ij} + h_{ij}c_{ij}(\gamma). \tag{4·1}$$

For a current estimate of β, we compute corresponding estimates of f_{ij} and h_{ij}, and then perform a weighted regression of the W_{ij} on these estimates, pooling over clusters and using those combinations of i and j implied by the form of c_{ij}. For computational simplicity we may use a diagonal weight matrix, approximating the variances of the W_{ij} by taking $E(W_{ij}) \simeq f_{ij}$ in (4·1).

The foregoing procedure also could be carried out using $W_{ij} = Y_i Y_j$, but numerical evidence suggests that this choice leads to more variable covariance parameter estimates.

REFERENCES

BAHADUR, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, Ed. H. Solomon, pp. 158–68. Stanford, CA: Stanford University Press.

CAREY, V., ZEGER, S. L. & DIGGLE, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **80**, 517–26.

CROWDER, M. C. (1995). On the use of a working correlation matrix in using generalised linear models for repeated measures. *Biometrika* **82**, 407–10.

HEAGERTY, P. J. & ZEGER, S. L. (1996). Marginal regression models for clustered ordinal measurements. *J. Am. Statist. Assoc.* **91**, 1024–36.

LIANG, K.-Y. & ZEGER, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

LUNN, A. D. & DAVIES, S. J. (1998). A note on generating correlated binary variables. *Biometrika* **85**, 487–90.

PRENTICE, R. L. (1988). Correlated binary regession with covariates specific to each binary observation. *Biometrics* **44**, 1033–48.

SUTRADHAR, B. C. & DAS, K. (1999). On the efficiency of regression estimators in generalised linear models for longitudinal data. *Biometrika* **86**, 459–65.

ZEGER, S. L. & LIANG, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* **42**, 121–30.

[*Received April* 2000. *Revised July* 2000]