BIOMETRICS 57, 1256–1264
December 2001

Parameterization of Multivariate Random Effects Models for Categorical Data

S. Rabe-Hesketh

Department of Biostatistics and Computing, Institute of Psychiatry, London SE5 8AF, U.K.
email: spaksrh@iop.kcl.ac.uk

and

A. Skrondal

Department of Epidemiology, National Institute of Public Health,
P.O. Box 4404 Nydalen, N-0403 Oslo, Norway

SUMMARY. Alternative parameterizations and problems of identification and estimation of multivariate random effects models for categorical responses are investigated. The issues are illustrated in the context of the multivariate binomial logit-normal (BLN) model introduced by Coull and Agresti (2000, *Biometrics* **56,** 73–80). We demonstrate that the BLN model is poorly identified unless proper restrictions are imposed on the parameters. Moreover, estimation of BLN models is unduly computationally complex. In the first application considered by Coull and Agresti, an identification problem results in highly unstable, highly correlated parameter estimates and large standard errors. A probit-normal version of the specified BLN model is demonstrated to be underidentified, whereas the BLN model is empirically underidentified. Identification can be achieved by constraining one of the parameters. We show that a one-factor probit model is equivalent to the probit version of the specified BLN model and that a one-factor logit model is empirically equivalent to the BLN model. Estimation is greatly simplified by using a factor model.

KEY WORDS: Equivalence; Factor model; Generalized linear mixed model; Identification; Multivariate binomial logit-normal model; Numerical integration.

1. Introduction

We discuss the parameterization of multivariate random effects models for categorical responses. We investigate two fundamental statistical problems, identification and equivalence, and suggest the use of factor models.

Identification is essential for consistent estimation of model parameters. The identification problem has been given ample attention in econometrics, where complex structural models have been used for a long time. This stands in contrast with biometrics, where much simpler models have traditionally been used. However, identification issues can no longer be ignored in biometrics due to the popularity of more complex models such as random effects models. A main objective of random effects models is to properly parameterize the dependence among responses. There are many ways of doing this; random intercept and random coefficient modeling are the standard approaches in biometrics. Such models are special cases of factor models, which have been employed in psychometrics for nearly 100 years (cf., Spearman, 1904) but are rarely used in biometrics. We show how factor models can be used to structure the covariance matrix of the random effects and to reduce the computational complexity of multivariate random effects models. The concept of equivalence concerns the possibility of empirically distinguishing between different statistical models. We may sometimes take advantage of equivalence in simplifying estimation problems through the use of factor models.

The importance of these ideas for multivariate random effects modeling of categorical data is best conveyed by considering a specific model. We investigate the multivariate binomial logit-normal (BLN) model for multiple binary responses recently suggested by Coull and Agresti (2000). Specifically, we will consider the BLN model used in their first application, where a dataset originally presented in Haber (1986) was analyzed. Infection status during four influenza outbreaks in Michigan in 1977, 1978, 1980, and 1981 was recorded for 263 subjects, yielding four binary repeated responses per subject. Our investigation will be informal, but we will refer to the literature for extensive formal treatments when appropriate.

2. The Multivariate Binomial Logit-Normal (BLN) Model

2.1 *The BLN Model as a Random Effects Logit Model*

Let y_{sr} be the binary response for subject s, $s = 1, \ldots, N$, during the rth influenza outbreak, $r = 1, \ldots, 4$, and let $\mathbf{y}_s = (y_{s1}, \ldots, y_{s4})'$ be the response vector for subject s with corresponding probability vector $\boldsymbol{\pi}_s$. The following model is

proposed by Coull and Agresti (2000):

$$\text{logit}(\boldsymbol{\pi}_s) = \mathbf{X}_s\boldsymbol{\beta} + \boldsymbol{\alpha}_s, \tag{1}$$

where $\mathbf{X}_s$ is a covariate matrix, $\boldsymbol{\beta}$ is a vector of fixed coefficients, and $\boldsymbol{\alpha}_s$ is a vector of random intercepts with $\boldsymbol{\alpha}_s \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma})$. The logits therefore have a multivariate normal distribution and, conditional on the $\boldsymbol{\pi}_s$, the responses are independent Bernouilli. In the influenza example, the covariate matrix is simply

$$\mathbf{X}_s = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2}$$

so that the parameters $\boldsymbol{\beta}$ represent the intercepts for each of the four responses.

The advantage of specifying a vector of random effects, one for each response, is that the covariance matrix can be unconstrained, whereas the conventional (single) random intercept model implies that all correlations between the logits are one. Instead of estimating the 10 variances and covariances freely, Coull and Agresti (2000) impose a structure on the covariance matrix. In their retained model, the matrix is

$$\boldsymbol{\Sigma} = \text{var}(\boldsymbol{\alpha}_s) = \begin{bmatrix} \sigma^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 \\ \rho_1\sigma^2 & \sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 \\ \rho_1\sigma^2 & \rho_1\sigma^2 & \sigma^2 & \rho_2\sigma^2 \\ \rho_2\sigma^2 & \rho_2\sigma^2 & \rho_2\sigma^2 & \sigma^2 \end{bmatrix} \tag{3}$$

The first three logits are therefore equicorrelated with correlation ρ_1, whereas the correlation between the first three logits and the fourth is ρ_2. This structure may be given a biological interpretation. The correlation among the first three logits, which is expected to be positive, could be due to a common susceptibility or frailty. On the other hand, a negative correlation is expected between the first three and the last logit if the final outbreak shares viruses with the preceding. In such a case, a prior infection would protect against infection during the last outbreak.

Coull and Agresti (2000) estimate the model using four-dimensional Gaussian quadrature, where the correlated random effects $\boldsymbol{\alpha}_s$ are represented by a linear combination of independent standard normal random effects $\mathbf{z}$ using the Cholesky decomposition of the covariance matrix $\mathbf{Q}$, with $\mathbf{QQ}' = \boldsymbol{\Sigma}$, so that $\boldsymbol{\alpha}_s = \mathbf{Qz}$. Integration is achieved by summing over a four-dimensional grid of quadrature points, in this case, using 20 quadrature points per dimension.

The likelihood can be expressed as

$$\begin{aligned} &l(\mathbf{y} \mid \mathbf{X}; \mathbf{Q}, \boldsymbol{\beta}) \\ &\quad = \prod_s \int f(z_1) \int f(z_2) \int f(z_3) \int f(z_4) \\ &\qquad \times \prod_r \left(\frac{e^{\mathbf{x}'_{sr}\boldsymbol{\beta} + \mathbf{q}'_r\mathbf{z}}}{1 + e^{\mathbf{x}'_{sr}\boldsymbol{\beta} + \mathbf{q}'_r\mathbf{z}}} \right)^{y_{sr}} \left(\frac{1}{1 + e^{\mathbf{x}'_{sr}\boldsymbol{\beta} + \mathbf{q}'_r\mathbf{z}}} \right)^{y_{sr}-1} \\ &\qquad \times dz_1 dz_2 dz_3 dz_4, \end{aligned} \tag{4}$$

where $f(\cdot)$ denotes standard normal densities, $\mathbf{y} = (\mathbf{y}'_1, \ldots, \mathbf{y}'_N)'$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$, and $\mathbf{x}'_{sr}$ and $\mathbf{q}'_r$ are the rth rows of $\mathbf{X}_s$ and $\mathbf{Q}$, respectively.

The parameter estimates for the BLN model reported in Coull and Agresti (2000) were $\hat{\boldsymbol{\beta}} = (-4.0, -4.4, -4.7, -4.5)$, $\hat{\sigma} = 4.05$, and $\hat{\rho}_1 = 0.43$ and $\hat{\rho}_2 = -0.25$. The deviance was $G^2 = 6.3$ with 8 $(= (2^4 - 1) - 7)$ d.f. No standard errors were reported.

2.2 *The BLN Model as a Random Effects Latent Response Model*

The BLN model can alternatively be written as a latent response model,

$$\mathbf{y}^*_s = \mathbf{X}_s\boldsymbol{\beta} + \boldsymbol{\alpha}_s + \mathbf{u}_s. \tag{5}$$

The elements of $\mathbf{u}_s$ have i.i.d. standard logistic distributions, with zero expectation and variance $\pi^2/3$, and are independent of $\boldsymbol{\alpha}_s$. $\mathbf{y}^*_s$ are latent responses generating the observed responses

$$y_{sr} = \begin{cases} 1 & \text{if } y^*_{sr} > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

The covariance matrix of the latent responses, conditional on the covariates, is

$$\begin{aligned} \boldsymbol{\Omega} &= \text{cov}(\mathbf{y}^*_s \mid \mathbf{X}_s) \\ &= \begin{bmatrix} \sigma^2 + \nu^2 & \rho_1\sigma^2 & \rho_1\sigma^2 & \rho_2\sigma^2 \\ \rho_1\sigma^2 & \sigma^2 + \nu^2 & \rho_1\sigma^2 & \rho_2\sigma^2 \\ \rho_1\sigma^2 & \rho_1\sigma^2 & \sigma^2 + \nu^2 & \rho_2\sigma^2 \\ \rho_2\sigma^2 & \rho_2\sigma^2 & \rho_2\sigma^2 & \sigma^2 + \nu^2 \end{bmatrix}, \end{aligned} \tag{7}$$

where ν^2 is equal to $\pi^2/3$. Note that $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \text{cov}(\mathbf{u}^*_s)$.

2.3 *The Multivariate Binomial Probit-Normal Model*

If a probit link is used instead of a logit link for the model, the elements of $\mathbf{u}_s$ in (5) have i.i.d. standard normal distributions instead of logistic distributions. Since the sum of two normal random variables $(\boldsymbol{\alpha}_s + \mathbf{u}_s)$ is normal, it follows that $\mathbf{y}^*_s \mid \mathbf{X}_s$ is multivariate normal for the binomial-probit normal (BPN) model. The covariance matrix of the latent responses conditional on the covariates is given by (7), where ν^2 in this case equals one.

3. Identification

A parametric statistical model is said to be identified if there is one and only one set of parameters that produces a given probability distribution for the observed variables.

Under suitable assumptions, a necessary and sufficient condition for identification at a parameter point is that the information matrix is nonsingular at the point. We refer to Koopmans and Reiersøl (1950), Rothenberg (1971), and Bekker, Merckens, and Wansbeek (1994) for extensive treatment of identification, including formal definitions, assumptions, and theorems. Identification is crucial since it is closely related to the prospects of consistent estimation (cf., Gabrielsen, 1978).

We note that Coull and Agresti (2000) did not discuss the identification status of the BLN model.

3.1 *Identification in the BPN Model*

For the BPN model, all information on the latent responses is contained in the first- and second-order moments since $\mathbf{y}^*_s \mid \mathbf{X}_s$ is multivariate normal. Since the observed responses are dichotomized versions of the latent responses, the probability distribution of the observed responses contains no information on the scale of the latent responses. Therefore, the information in BPN models is contained in the first- and second-order

moments of the normalized latent responses (where the scale is set to unity, i.e., $y^*_{sr}/(\sigma^2+1)^{1/2}$).

The marginal probability of the rth response is

$$\text{prob}(y_{sr}=1 \mid \mathbf{x}_{sr}) = \text{prob}(y^*_{sr}>0 \mid \mathbf{x}_{sr}) = \Phi\left(\frac{\beta_r}{\sqrt{\sigma^2+1}}\right), \tag{8}$$

where Φ denotes the cumulative standard normal distribution. It follows that the thresholds (or integration limits)

$$\tau_r = \frac{\beta_r}{\sqrt{\sigma^2+1}}, \qquad r=1,2,3,4, \tag{9}$$

are identified. The thresholds are simply the means of the normalized latent responses.

From equation (7), the off-diagonal elements of the correlation matrix of the latent responses are

$$\rho_{12}=\rho_{13}=\rho_{23}=\frac{\rho_1\sigma^2}{\sigma^2+1} \tag{10}$$

and

$$\rho_{14}=\rho_{24}=\rho_{34}=\frac{\rho_2\sigma^2}{\sigma^2+1}. \tag{11}$$

These tetrachoric correlations ρ_{ij} (Pearson, 1900) are well known to be identified.

Since the probability distribution of $\mathbf{y}_s$ is completely determined by the thresholds and tetrachoric correlations, the parameters of the BPN model are not identified in this case. This can be seen by considering an arbitrary set of values for the model parameters β_r, σ, ρ_1, and ρ_2. We can change σ and use equations (9)–(11) to obtain new values of the other parameters that generate the original thresholds and tetrachoric correlations. Underidentification is also evident from the fact that there are seven unknown parameters but only six identifying equations. In order for the model to be identified, suitable identification restrictions therefore need to be imposed.

A convenient way to proceed is by fixing σ, but σ cannot be set to any positive value. Since $|\rho_1| \leq 1$, the inequality $\sigma^2 \geq |\rho_{12}|/(1-|\rho_{12}|)$ must be satisfied, and similarly for ρ_{14}. When σ is fixed, it follows from the tetrachoric correlations that ρ_1 and ρ_2 are identified. We also see that β_1, β_2, β_3, and β_4 are identified from the τ_r.

Note that including a continuous covariate x_{sr} in the design matrix does not alleviate the identification problem. The thresholds then become

$$\tau_{sr} = \frac{\beta_{0r}+\beta_{1r}x_{sr}}{\sqrt{\sigma^2+1}}, \tag{12}$$

and a change in σ can be counteracted by a suitable rescaling of the regression parameters.

For a general treatment of identification in probit models with latent variables, see Skrondal (1996).

3.2 *Empirical Identification in the BLN Model*

A model is said to be empirically underidentified for a sample if the estimated information matrix at the maximum likelihood estimates is nearly singular (cf., Wiley, 1973; McDonald and Krane, 1977).

In this case, there are more than one set of parameters that produce almost identical maximum likelihoods, and standard errors and intercorrelations of parameter estimates corresponding to flat directions will be high. Collinearity among predictor variables in linear regression is a special case. The condition number, defined as the square root of the ratio of the largest to the smallest eigenvalue, is often used as a measure of how close a matrix is to singularity.

We first consider the first- and second-order moments of the normalized latent responses for the BLN model. The marginal probability of the rth response is

$$\begin{aligned}\text{prob}(y_{sr}=1 \mid \mathbf{x}_s) &= \text{prob}(y^*_{sr}>0 \mid \mathbf{x}_s)\\ &= \text{prob}\left(\frac{\alpha_{sr}+u_{sr}}{\sqrt{(\sigma^2+\pi^2/3)}} < \frac{\beta_r}{\sqrt{\sigma^2+\pi^2/3}}\right).\end{aligned} \tag{13}$$

The shape of the cumulative probability distribution function of the standardized compound normal-logistic $(\alpha_{sr}+u_{sr})/(\sigma^2+\pi^2/3)^{1/2}$ is determined by σ, the standard deviation of the normally distributed α_{sr}, because the scale of the logistic random variable u_{sr} is fixed. The compound distribution becomes increasingly normal as σ increases. For a given value of σ, the marginal probability depends only on $\beta_r/(\sigma^2+\pi^2/3)^{1/2}$, the thresholds for the BLN model. These thresholds are analogous to the thresholds for the BPN model in (9) with 1 replaced by $\pi^2/3$. The correlations between the y^*_{sr} (which are no longer called tetrachoric) take on the same form as in the probit case with the 1 in equations (10) and (11) replaced by $\pi^2/3$. We can constrain σ to a constant and still obtain any set of thresholds and correlations as long as

$$\sigma^2 \geq |\rho_{12}|(\pi^2/3)/(1-|\rho_{12}|), \tag{14}$$

and similarly for ρ_{14}.

However, unlike the BPN model, the BLN model may be identified from third-order and higher order moments since $\mathbf{y}^*_s$ is not multivariate normal (conditional on $\mathbf{X}_s$) but compound normal-logistic since $\boldsymbol{\alpha}_s$ is normal and $\mathbf{u}_s$ logistic. Although σ is not identified from the first- and second-order moments, it is likely to be identified from higher order moments because σ determines the degree of normality of the compound distribution. However, because of the constraint in equation (14), there may not be a value of σ that can generate both appropriate second-order and higher order moments of the latent responses for a particular dataset.

We moreover expect that there is little information on σ from the probability distribution of the observed (dichotomized) responses for a number of reasons:

(1) The density of the sum y^*_{sr} of two random variables, composed of a normal α_{sr} and a logistic u_{sr}, is fairly normal even for small σ due to the similarity of the normal and logistic distributions.
(2) The third-order and higher order moments of the latent responses $\mathbf{y}^*_s$ have a high sampling variability.
(3) The observed, dichotomized responses $\mathbf{y}_s$ retain little of the information in the higher order moments of $\mathbf{y}^*_s$.

Hence, we conjecture that the BLN model is empirically underidentified.

The validity of our conjecture was investigated in the context of the influenza example. Using purpose-written code in `Stata` (StataCorp, 1999), we estimated the parameters

Table 1

Parameter estimates, standard errors, and deviance for constrained and unconstrained versions of the BLN model (20 quadrature points per dimension)

		Standard errors	
	Estimate	Unconstrained	Constrained
β_1	−4.04	6.85	0.39
β_2	−4.42	7.34	0.41
β_3	−4.69	7.77	0.42
β_4	−4.56	7.57	0.42
σ	4.06	7.99	—
ρ_1	0.43	0.28	0.10
ρ_2	−0.25	0.20	0.12
G^2	6.28	—	—

of the BLN model retained by Coull and Agresti (2000) by maximum likelihood (using **Stata**'s modified Newton–Raphson algorithm). The parameter estimates were nearly identical to those of Coull and Agresti and are shown in Table 1.

We estimated the information matrix of the parameter estimates for both the Coull and Agresti (2000) version of the BLN model and a version where σ was constrained to 4.06. This particular constraint, equal to the maximum likelihood estimate, was simply chosen to enable us to compare the magnitude of the estimated standard errors for the models on the same scale. The information matrices were estimated by the negative Hessian matrix of the log-likelihood function, which was obtained using numerical first and second derivatives of the log-likelihood computed by **Stata**'s maximum likelihood functions. For the constrained model, the condition number for the information matrix was merely 5.2. In contrast, the condition number for the unconstrained model was 179.5, which is extremely large (the smallest eigenvalue was less than 0.004). Thus, the estimated information matrix for the BLN model as specified by Coull and Agresti is almost singular.

Inverting the estimated information matrices, we obtained the estimated covariance matrices of the parameter estimates. As can be seen in Table 1, the estimated standard errors decreased substantially when σ was fixed. The correlations of the parameter estimates are shown in Table 2. For the unconstrained model, the parameter estimates were highly intercorrelated, most correlations approaching ± 1, the smallest correlation (in absolute value) being -0.79, whereas the highest correlation for the constrained model was 0.19.

These results all suggest that the BLN model is empirically underidentified. An identification restriction, similar to the BPN model, should be imposed to obtain stable estimates.

We have argued that, although the BLN model is not identified from the first- and second-order moments, it may be identified from higher order moments. Having demonstrated empirical underidentification, we now investigate whether this is due to the scarce information in the higher order moments. For a range of values of σ, we computed the other parameters to preserve the thresholds and correlations implied by the maximum likelihood solution using equations (9), (10), and (11). The models with these different sets of parameter values imply identical first- and second-order moments of the latent responses but different higher order moments (the greater σ, the more the moments will resemble those of the multivariate normal distribution). The deviance G^2 of these models is plotted against σ in Figure 1, where σ increases from 1.35, the lowest value consistent with the correlations of the latent responses, to 8. The deviance hardly changes at all although the higher order moments were deliberately ignored in determining the other parameters for each value of σ. This provides direct evidence for the scarcity of information in the higher order moments of the latent responses. Note that the curve in Figure 1 represents an upper bound for the deviance corresponding to the profile likelihood for σ since the other parameters are not estimated by maximum likelihood. Estimating the model with σ fixed at 8.2, e.g., gives a deviance of only 6.53.

Table 2

Correlation matrices of parameter estimates for the BLN model; above the diagonal, constrained model; below the diagonal, unconstrained model

	β_1	β_2	β_3	β_4	σ	ρ_1	ρ_2
β_1	1	0.190	0.187	−0.083	—	0.050	−0.025
β_2	0.997	1	0.185	−0.080	—	0.062	−0.033
β_3	0.997	0.998	1	−0.077	—	0.069	−0.037
β_4	0.997	0.997	0.997	1	—	0.014	−0.045
σ	−0.998	−0.998	−0.999	−0.998	1	—	—
ρ_1	0.941	0.942	0.942	0.941	−0.942	1	−0.106
ρ_2	−0.814	−0.815	−0.815	−0.815	0.815	−0.788	1

It should be noted that empirical underidentification does not imply lack of theoretical identification because singularity is assessed at the parameter estimates (not the true values) and fallible numerical methods are used (e.g., McDonald and Krane, 1979). Also note that the empirical identification problem is a general feature of BLN models whatever the number of responses unless proper parameter restrictions are imposed.

4. Factor Models

A factor structured latent response model has the following form:

$$y^*_{sr} = \gamma_r \mathbf{x}_s + \boldsymbol{\lambda}'_r \boldsymbol{\eta}_s + u_{sr}, \tag{15}$$

where $\boldsymbol{\eta}_s$ is a vector of m factors that have a multivariate normal distribution, $\boldsymbol{\lambda}_r$ is a vector of factor loadings of the rth latent response on the m factors, γ_r is a vector of fixed effects, and u_{sr} an error term. Seminal contributions to the identification of factor models for directly observed responses include Reiersøl (1950a) and Anderson and Rubin (1956), and a modern account is given by Bekker et al. (1994). However, there is a paucity of research regarding identification of factor structured latent response models. The only detailed treatment we are aware of is Skrondal (1996).

Here we consider two versions of the factor structured latent response model:

- The factor structured logit (FSL) model, where u_{sr} is logistic.
- The factor structured probit (FSP) model, where u_{sr} is standard normal.

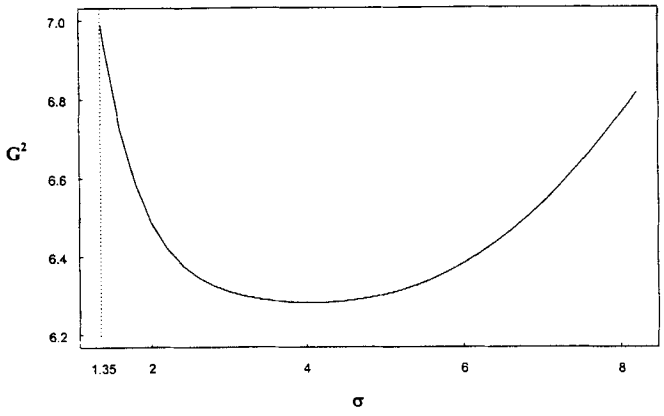

Figure 1. Deviance (G^2) for different values of σ. The other parameters have been computed to preserve the first- and second-order moments implied by the maximum likelihood solution.

The BLN model with an unstructured covariance matrix can be formulated as a special case of the general FSL model. Specifically, as many correlated factors are specified as there are responses and $\boldsymbol{\lambda}_r$ are zero vectors apart from ones in position r. However, typically the number of factors required will be much smaller than the number of responses. Consider the one-factor model,

$$y^*_{sr} = \gamma_r + \lambda_r \eta_s + u_{sr}. \tag{16}$$

The FSL version of this model is known as the (two-parameter) logistic test or item-response model in psychometrics (e.g., Birnbaum, 1968) and was presented in equation (11) of Coull and Agresti (2000). If there are at least three responses, the model is identified as long as the variance of the latent variable is constrained to an arbitrary positive constant (typically one) or, alternatively, if one of the factor loadings is constrained to an arbitrary nonzero constant (typically one). In this article, we use the latter approach. However, the choice of restriction is immaterial in the sense that the resulting models are equivalent as described in the subsequent section.

The advantage of using a one-factor model to structure the covariance matrix is that we now only have to integrate over one random effect, η_s. The likelihood is

$$\begin{aligned} l(\mathbf{y} \mid \mathbf{x}; \psi, \boldsymbol{\lambda}, \boldsymbol{\gamma}) &= \prod_s \int f(\eta_s; \psi) \\ &\quad \times \left\{ \prod_r F(\gamma_r + \lambda_r \eta_s)^{y_{sr}} (1 - F(\gamma_r + \lambda_r \eta_s))^{1 - y_{sr}} \right\} \\ &\quad \times \mathrm{d}\eta_s, \end{aligned} \tag{17}$$

where ψ is the factor variance, $f(\eta_s; \psi)$ is the normal density with variance ψ, and F represents the cumulative standard normal distribution for the FSP model and the cumulative logistic distribution for the FSL model.

5. Equivalence

Consider two statistical models, both identified and with the same number of unknown parameters. The models are said to be equivalent if they are reparameterizations, i.e., there are one-to-one functions relating the parameters across the models to produce identical probability distributions for the observed variables.

Equivalence is a fundamental statistical issue since it is impossible to distinguish between equivalent models empirically. This is, of course, detrimental when the models represent different substantive data-generating processes. On the other hand, we may sometimes take advantage of equivalence in simplifying estimation problems, as is shown below.

We refer to Luijben (1991) and Bekker et al. (1994) for extensive treatments of equivalence, including formal definitions, assumptions, and theorems.

5.1 *Equivalence Between the BPN and FSP Models*

Consider a one-factor FSP model with the constraints $\lambda_1 = \lambda_2 = \lambda_3 = 1$. If $\mathrm{var}(\eta_s) = \psi$, the covariance matrix of the latent responses becomes

$$\boldsymbol{\Omega} = \begin{bmatrix} \psi + 1 & \psi & \psi & \lambda_4 \psi \\ \psi & \psi + 1 & \psi & \lambda_4 \psi \\ \psi & \psi & \psi + 1 & \lambda_4 \psi \\ \lambda_4 \psi & \lambda_4 \psi & \lambda_4 \psi & \lambda_4^2 \psi + 1 \end{bmatrix}. \tag{18}$$

Therefore, the thresholds are

$$\tau_r = \frac{\gamma_r}{\sqrt{\psi + 1}}, \qquad r = 1, 2, 3, \tag{19}$$

and

$$\tau_4 = \frac{\gamma_4}{\sqrt{\lambda_4^2 \psi + 1}}, \tag{20}$$

and the tetrachoric correlations are

$$\rho_{12} = \rho_{13} = \rho_{23} = \frac{\psi}{\psi + 1} \tag{21}$$

and

$$\rho_{14} = \rho_{24} = \rho_{34} = \frac{\lambda_4 \psi}{\sqrt{(\psi + 1)(\lambda_4^2 \psi + 1)}}. \tag{22}$$

Note that these equations imply the same restrictions as the identified BPN model. There are no restrictions on the thresholds for either model since there is a free parameter for each threshold. The tetrachoric correlations are constrained in both models but in an identical way; the tetrachoric correlations involving the fourth occasion are constrained to be equal and the remaining correlations are specified as equal. Thus, the thresholds and the tetrachoric correlations are structured in the same way for the restricted one-factor FSP model and the BPN model. Since the probability distribution of probit models is completely characterized by these quantities, it follows that the restricted one-factor FSP model is equivalent to the BPN model.

Equating the structures put on the thresholds and tetrachoric correlations for the BPN and FSP models, functions relating the parameters across models can be expressed as follows:

$$\beta_r = \gamma_r \frac{\sqrt{\sigma^2 + 1}}{\sqrt{\psi + 1}}, \qquad r = 1, 2, 3, \tag{23}$$

$$\beta_4 = \gamma_4 \frac{\sqrt{\sigma^2 + 1}}{\sqrt{\lambda_4^2 \psi + 1}}, \tag{24}$$

$$\rho_1 = \frac{\psi(\sigma^2+1)}{\sigma^2(\psi+1)}, \tag{25}$$

and

$$\rho_2 = \frac{\lambda_4\psi(\sigma^2+1)}{\sigma^2\sqrt{(\psi+1)(\lambda_4^2\psi+1)}}, \tag{26}$$

where σ^2 is fixed to ensure identification.

5.2 *Empirical Equivalence Between the BLN and the Factor Models*

We say that two models are empirically equivalent for a sample if there are one-to-one functions relating the parameters across models such that almost identical likelihoods are produced.

Using the same arguments as in the previous section, we expect that one-factor FSL and FSP models can generate the same first- and second-order moments of the latent responses as the BLN model. Any difference in the likelihoods can only be due to higher order moments, which were shown in Section 3.2 to have minimal effect on the goodness of fit. We therefore expect the one-factor FSL (and FSP) model to have nearly identical fit to the (constrained or unconstrained) BLN model of Coull and Agresti (2000). Of course, the advantage of the FSL and FSP models is that the dimensionality of integration is substantially reduced.

The constrained single-factor models were fitted using the `Stata` program `gllamm` (Rabe-Hesketh, Pickles, and Skrondal, 2000; Rabe-Hesketh, Pickles, and Taylor, 2001) using 20 quadrature points. The parameter estimates and their standard errors are given in Table 3. The parameter estimates changed at most in the fifth decimal place when 40 quadrature points were used, and the condition numbers for the estimated information matrices were 3.5 and 3.9 for the FSL and FSP models, respectively.

Functions approximately relating the BLN and FSL parameters are obtained by substituting $\pi^2/3$ for 1 in the formulas derived in Section 5.1. Substituting the maximum likelihood estimates for the FSL parameters, we obtain values that are very close to the maximum likelihood estimates for the BLN model: $\hat{\beta} = (-3.87, -4.27, -4.55, -4.30)$, $\hat{\rho}_1 = 0.41$, and $\hat{\rho}_2 = -0.25$. The fit of the BLN, FSL, and FSP models was also almost identical, being $G^2 = 6.28$, 6.58, 6.26, respectively. This is due to the fact that, as expected, the implied first- and second-order moments of all three models were nearly identical: for the BLN, FSL, and FSP models, respectively, the thresholds are estimated as $\hat{\tau}_1 = -0.91, -0.87, -0.91$, $\hat{\tau}_2 = -0.99, -0.96, -1.00$, $\hat{\tau}_3 = -1.05, -1.03, -1.06$, $\hat{\tau}_4 = -1.03, -0.97, -1.03$. The correlations of the latent responses are estimated as $\rho_{12} = \rho_{13} = \rho_{22} = 0.36, 0.34, 0.36$ and $\rho_{14} = \rho_{24} = \rho_{34} = -0.21$, -0.21, -0.20. The greater similarity of the probit estimates with the BLN estimates could be due to the dominance of the normal component in the BLN model since the estimated variances of the normal and logistic components were 16.4 ($=4.05^2$) and 3.3 ($=\pi^2/3$), respectively.

We conclude that the BLN model is empirically equivalent to the factor structured alternatives FSL and FSP. There seems to be no biological argument in favor of the BLN specification as compared with the factor structured models. Coull and Agresti's (2000) motivation for the BLN model appears to be the desire to fit a realistic correlation structure

Table 3

Parameter estimates, standard errors, and deviance for the factor structured logit (FSL) and probit (FSP) models (20 quadrature points)

	Logit model		Probit model	
	Estimate	SE	Estimate	SE
γ_1	−1.95	0.24	−1.14	0.13
γ_2	−2.15	0.25	−1.24	0.14
γ_3	−2.29	0.26	−1.32	0.14
γ_4	−1.88	0.25	−1.09	0.12
ψ	1.71	0.60	0.56	0.19
λ_4	−0.54	0.32	−0.49	0.28
G^2	6.58	—	6.26	—

to the responses. However, the one-factor models are clearly preferred to the BLN model since they require integration in only one dimension instead of four. In practice, estimating the BLN model took about 8000 times as long as estimating the factor models (20 quadrature points used per dimension).

6. Discussion

The conclusion of our investigation is that (1) factor models should be more widely used in biometrics and (2) the fundamental statistical concepts of identification and equivalence should be given more attention in biometrics.

The BLN model is empirically underidentified unless proper restrictions are imposed on the parameters. This does not invalidate the BLN model per se but implies that identification must be taken seriously when BLN models are specified. Otherwise, identification problems such as that uncovered for the application of Coull and Agresti (2000) may be encountered.

Investigation of identification does not only serve to reveal underindentified specifications and suggest identifying restrictions. It may also lead to the discovery of identified models that are less restrictive than those conventionally used (e.g., Skrondal, 1996, Chapter 10). At first glance, the possibility of identification of all parameters in models with a logit link appears to be a definite advantage as compared with the probit models. The situation is somewhat similar to that discussed by Reiersøl (1950b), where an errors in variables model is identified if and only if at least one of the latent variables is not normally distributed. However, identification is based on scarce information and very large samples are likely to be required to ensure empirical identification of BLN models. Moreover, the fact that only first- and second-order moments can be used for identification in the probit models makes the analytical analysis of identification tractable.

We recommend the latent response formulation of random effects models for categorical responses to facilitate the analysis of identification and equivalence. Formulating such models as generalized linear mixed models with logit (or probit) links, $g(\cdot)$, would also give the misleading impression that the one-factor logit (or probit) model, $g(\pi_{sr}) = \gamma_r + \lambda_r\eta_s$, is very restrictive because, as pointed out in Coull and Agresti (2000), it specifies that all correlations among the logits (or probits) are equal to 1 or −1. However, these restrictions do

not apply to the latent responses underlying the observed responses.

Numerical integration is required over as many dimensions as there are responses in the approach of Coull and Agresti (2000). Presently, this is unfeasible if there are more than a few responses. To reduce the integration problem, we have suggested factor structured models. For the first example in Coull and Agresti, we have demonstrated that the probit version of their model is equivalent to a constrained one-factor model. Their BLN model is empirically equivalent to a constrained factor structured logit (FSL) model. For more complex situations, a higher dimensional factor model may be required. However, the dimensionality of the factor model will in practice be considerably lower than the number of responses. The factor approach is hence clearly superior from a computational point of view.

Models commonly used in biostatistics can be derived as special cases of factor models. The random intercept model is obtained when all factor loadings are constrained to be equal to one in the one-factor model. For balanced data like those considered here, random coefficient models are obtained as multidimensional confirmatory factor models where loadings are constrained to particular constants (Skrondal, 1996). Factor modeling can be performed both in exploratory and confirmatory mode. In the exploratory mode, we initially suggest approximating the unstructured covariance matrix with a well-fitting factor model with as low a dimensionality as possible. Subsequently, the estimates from such a model may suggest factor models with fewer parameters. In the confirmatory mode, theory, previous research, or the experimental design may prescribe a specific factor structure.

Factor models with a variety of link functions and error distributions can be estimated by maximum likelihood using the software `gllamm`. The program handles multiple correlated factors, multilevel models, mixed responses, linear parameter constraints, and semiparametric maximum likelihood where no distributional assumptions need be imposed on the factors. The class of models considered in this article can also be fitted in SAS PROC NLMIXED. Useful treatments of factor models include Lawley and Maxwell (1971) for continuous responses and Bartholomew and Knott (1999) for continuous and categorical responses.

Acknowledgements

The authors contributed equally to the research. The article was written while Anders Skrondal was visiting the Biostatistics Group at The University of Manchester. We wish to thank Alan Agresti, Brent Coull, and Andrew Pickles for helpful comments. `gllamm` may be downloaded from `http://www.iop.kcl.ac.uk/IoP/Departments/BioComp/programs/gllamm.html`.

Résumé

Nous étudions différents paramétrages et problème d'identification et d'estimation d'un model multidimensionnel à effets aléatoires pour des variables à expliquer qui sont catégorielles. Les résultats sont illustrés avec le model binomial logit-normale multidimensionnel (BLN) introduit par Coull et Agresti (2000). Nous démontrons que le model BLN est mal identifié a moins que des restrictions particulières ne soient imposées sur les paramètres. De plus ces estimations conduisent à des calculs trop complexes. Dans la première application étudiée par Coull et Agresti, l'identification du modèle est très instable avec des paramètres très corrélés et d'écart-type élevé. On démontre qu'une version probit normale du model BLN est sous identifiée alors que le BLN model est empiriquement sous identifié. L'identification peut être peut être complète en imposant une contrainte sur l'un des paramètres. Nous montrons que le model probit à un facteur est équivalent à la probit version du modèle BLN spécifié et que le model logit à un facteur est empiriquement équivalent au BLN modèle. L'estimation est grandement simplifiée par l'utilisation d'un modèle à facteur.

References

Anderson, T. W. and Rubin, H. (1956). Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability,* 111–150. Berkeley, California: University of California Press.

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis.* London: Arnold.

Bekker, P. A., Merckens, A., and Wansbeek, T. J. (1994). *Identification, Equivalent Models and Computer Algebra.* Boston: Academic.

Birnbaum, A. (1968). Some latent trait models. In *Statistical Theories of Mental Test Scores,* F. M. Lord and M. Novick (eds), 394–424. Reading, Massachusetts: Wesley.

Coull, B. A. and Agresti, A. (2000). Random effects modeling of multiple binomial responses using the multivariate binomial logit-normal distribution. *Biometrics* **56,** 73–80.

Gabrielsen, A. (1978). Consistency and identifiability. *Journal of Econometrics* **8,** 261–263.

Haber, M. (1986). Testing for pairwise independence. *Biometrics* **42,** 429–435.

Koopmans, T. C. and Reiersøl, O. (1950). The identification of structural characteristics. *Annals of Mathematical Statistics* **21,** 327–335.

Lawley, D. M. and Maxwell, A. E. (1971). *Factor Analysis as a Statistical Method.* London: Butterworths.

Luijben, T. W. (1991). Equivalent models in covariance structure analysis. *Psychometrika* **56,** 653–665.

McDonald, R. P. and Krane, W. R. (1977). A note on local identifiability and degrees of freedom in the asymptotic maximum likelihood test. *British Journal of Mathematical and Statistical Psychology* **30,** 198–203.

McDonald, R. P. and Krane, W. R. (1979). A Monte Carlo study of local identifiability and degrees of freedom in the asymptotic maximum likelihood test. *British Journal of Mathematical and Statistical Psychology* **32,** 121–132.

Pearson, K. (1900). Mathematical contributions to the theory of evolution vii. On the inheritance of characters not capable of exact quantitative measurement. *Philosophical Transactions of the Royal Society, Series A* **195,** 79–150.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2001). GLLAMM: A general class of multilevel models and a stata program. *Multilevel Modelling Newsletter* **13,** 17–23.

Rabe-Hesketh, S., Pickles, A., and Taylor, C. (2000). sg129:

Generalized linear latent and mixed models. *Stata Technical Bulletin* **53,** 47–57.

Reiersøl, O. (1950a). On the identifiability of parameters in Thurstone's multiple factor model. *Psychometrika* **15,** 121–149.

Reiersøl, O. (1950b). Identifiability of a linear relation between variables which are subject to error. *Econometrica* **42,** 731–736.

Rothenberg, T. (1971). Identification in parametric models. *Econometrica* **39,** 577–591.

Skrondal, A. (1996). *Latent Trait, Multilevel and Repeated Measurement Modelling with Incomplete Data of Mixed Measurement Levels.* Oslo: Section of Medical Statistics, University of Oslo.

Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology* **15,** 201–293.

StataCorp. (1999). *Stata Statistical Software,* Release 6. College Station, Texas: Stata Corporation.

Wiley, D. E. (1973). The identification problem for structural equation models with unmeasured variables. In *Structural Equation Models in the Social Sciences,* A. S. Goldberger and O. Duncan (eds), 69–83. New York: Academic.

Received July 2000. *Accepted May* 2001.

The authors replied as follows:

We thank Rabe-Hesketh and Skrondal for pointing out the empirical underidentification of the σ parameter in the first example we presented. Our investigations of the model in this example focused on the impact of different starting values and different numbers of quadrature points, and we did not compute standard errors. We should have done so since, as the authors show, this would have indicated the problem. We used only $\{\hat{\beta}_i\}$ and not $\hat{\mathbf{\Sigma}}$ in our interpretations of the model fit, so the substantive conclusions we made using this model are correct.

We agree that identifiability is an important aspect of model formulation and model checking and believe readers will benefit from reading the authors' impressive discussion of this. However, we do not agree with some general comments in the article, such as the statement in the abstract that "... the BLN model is poorly identified unless proper restrictions are imposed on the parameters" and the statement at the end of Section 3 that "... the empirical identification problem is a general feature of BLN models whatever the number of responses unless proper parameter restrictions are imposed." For instance, this problem does not occur in other applications we have made of the model. Thus, we worry that readers may get the impression that the model itself is problematic. We believe the model is fine as long as one does not make an identification error, and of course this is true for any form of model.

The developmental toxicity example in Section 3.2 of Coull and Agresti (2000) illustrates this. In this example, the condition numbers from SAS PROC NLMIXED for the BLN model with no constraints on the variance–covariance matrix (model (3) in Table 2 of our article) and the final model that we used (model (4) in Table 2) are both approximately seven. Thus, the parameters in the BLN model are identifiable.

Given that the parameters in a BLN model can be identified, the question becomes whether it is worthwhile to consider this class of models separately from the general factor structured logit models (FSL) described by Rabe-Hesketh and Skrondal. Indeed, the authors note that the BLN model is a special case of the FSL model since the latter specifies both factor loadings and a multivariate normal latent factor. We believe that this distinction is worthwhile for reasons of interpretability. Often, it is intuitively natural to envision correlation patterns among multivariate binomial responses as arising solely from correlated random effects in a logistic regression model. Focus on this smaller class of models is analogous to the separate study of the special class of generalized linear mixed models (McCulloch and Searle, 2001) within the larger class of generalized factor-analytic latent variable models.

As an illustration of this interpretability advantage, consider the leading crowd data analyzed with a random effects model by Agresti et al. (2000) and also in a technical report version of our article. A sample of schoolboys was interviewed twice, several months apart, and asked about their self-perceived membership in the "leading crowd" and about whether one must sometimes go against his principles to be part of that leading crowd. Thus, there are two binary variables, which we refer to as membership and attitude, measured at each of two interview times for each subject. Agresti et al. (2000) analyzed these data with a multivariate logit model that is a special case of the general BLN model with random effects for attitude and for membership. In this example, it is easy to imagine the existence of underlying correlated membership and attitude variables. As elements of a multivariate normal covariance matrix, the estimates in $\hat{\mathbf{\Sigma}}$ are easily understood; these suggest that there is more heterogeneity with respect to membership than attitude and that the random effects have a weak positive correlation. This approach is also attractive in that it is a continuous analog to discrete latent class models proposed by Goodman (1974) based on two associated binary latent variables. More general factor-analytic models that set some factor loadings equal to one and estimate others do not provide direct estimates and associated standard errors of this latent correlation.

We agree with Rabe-Hesketh and Skrondal that there are also advantages of the FSL formulation. Using the BLN formulation, one cannot specify a one-factor solution yet allow a general correlation structure among the multiple logits. As the authors mention, this flexibility may allow one to fit a simpler model, yielding a computational advantage in some settings. These models also represent a natural extension of the traditional factor analytic methods for normal data.

In summary, we believe that the BLN formulation for multivariate binomial responses should not be summarily dismissed. The parameters in the models can be made identifiable, and they often have appealing interpretations. In choosing between the correlated random effects and factor formulations, we believe that one should weigh the advantages and disadvantages of each approach in the context of a given application.

References

Agresti, A., Booth, J. G., Hobert, J. P., and Caffo, B. (2000). Random effects modeling of categorical response data. *Sociological Methodology* **30,** 27–80.

Goodman, L. A. (1974). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology* **79,** 1179–1259.

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

Brent A. Coull
Department of Biostatistics
Harvard School of Public Health
677 Huntington Avenue, Boston, MA 02115
email: bcoull@hsph.harvard.edu

and

Alan Agresti
University of Florida