Harvard University
Harvard University Biostatistics Working Paper Series

Year 2006 *Paper* 45

A Computationally Tractable Multivariate Random Effects Model for Clustered Binary Data

Brent A. Coull* E. Andres Houseman†
Rebecca A. Betensky‡

*Harvard University, bcoull@hsph.harvard.edu
†Harvard School of Public Health, ahousema@hsph.harvard.edu
‡Harvard School of Public Health, betensky@hsphmail.harvard.edu

A Computationally Tractable Multivariate Random Effects Model for Clustered Binary Data

Brent A. Coull, E. Andres Houseman and Rebecca A. Betensky

Department of Biostatistics
Harvard School of Public Health
655 Huntington Avenue
Boston, Massachusetts 02115, U. S. A.

bcoull@hsph.harvard.edu, ahousema@hsph.harvard.edu,
betensky@sdac.harvard.edu

SUMMARY

We consider a multivariate random effects model for clustered binary data that is useful when interest focuses on the association structure among clustered observations. Based on a vector of gamma random effects and a complementary log-log link function, the model yields a likelihood that has closed form, making a frequentist approach to model fitting straightforward. This closed form yields several advantages over existing methods, including easy inspection of model identifiability and straightforward adjustment for nonrandom ascertainment of subjects, such as that which occurs in family studies of disease aggregation. We use the proposed model to analyse two different binary datasets concerning disease outcome data from a familial aggregation study of breast and ovarian cancer in women and loss of heterozygosity outcomes from a brain tumour study.

Some key words: Binary time series; Complementary log-log link; Generalised linear mixed model; Multivariate gamma.

1 Introduction

Use of generalised linear mixed models (Breslow & Clayton, 1993) has become a popular approach to modelling correlated discrete data, with the help of commercial software packages such as SAS, Stata and S-Plus/R. The models account for correlation among clustered observations by including random effects in the linear predictor component of the model.

In some scientific settings interest focuses primarily on the association structure among clustered observations. Examples include studies focusing on serially correlated observations (Fitzmaurice & Lipsitz, 1995; Aitkin & Alfo, 1998), familial aggregation of disease (Betensky & Whittemore, 1996; Hudson et al., 2001), and loss of heterozygosity analysis of brain tumours (Cairncross et al., 1998). A disadvantage of standard generalised linear mixed models in these instances is their inability to handle relatively complex dependence structures among clustered responses. Several authors have proposed adding additional random effects to model flexibly more complicated association structures (Aitchison & Ho, 1989; Diggle et al., 2002, §11.4.2; Agresti, 1997; Coull & Agresti, 2000). However, these more complicated structures add a layer of complexity in model fitting. For instance, Aitchison & Ho (1989) and Coull & Agresti (2000) noted that Gaussian quadrature methods are only feasible when the dimension of the random effects is at most four. Diggle et al. (2002) resorted to Markov chain Monte Carlo sampling to fit a logistic regression model with serially correlated random effects.

We consider for clustered binary data a multivariate random effects extension of the model with complementary log-log link and log-gamma random intercepts proposed by Conaway (1990). Henderson & Shimakura (2003) and Henderson et al. (2003) proposed the use of multivariate gamma random effects in log-linear models for serially correlated counts and spatial models for survival data, respectively. The first set of authors noted that this random effects assumption yields closed-form expressions for joint distributions of bivariate sets of counts, but showed that the calculation of joint distributions for higher dimensions is computationally prohibitive. We highlight the fact that use of this random effects distribution in conjunction with the complementary log-log link leads to computationally simple expressions for

the joint distribution of a multivariate binary response. As a result, model fitting via maximum likelihood is computationally simple, allowing for the likelihood-based analysis of moderately large datasets. Humphreys (1998) applied a special case of the model based on the additive formulation of the multivariate gamma distribution to some marketing data, but did not consider theoretical identifiability or parameter interpretation for the general model.

The model is attractive when interest focuses on the full joint probability distribution for the multivariate response. For instance, in studies of familial aggregation, interest focuses on measures of risk that are conditional on other family members, and the relevant conditional likelihood is derived from the full joint distribution. Thus, the model affords straightforward adjustment for nonrandom subject ascertainment, which is common in family studies of disease. Another example is the setting in which interest focuses on the union probability related to having at least one event (Lipsitz et al. 1995, 1996). The models are also useful for prediction, since under this formulation the empirical Bayes predictions of the random effects also have closed form expressions. The fact that the proposed approach is likelihood-based allows for deviance-based hypothesis testing and goodness-of-fit. Finally, it can be difficult to establish identifiability of all model parameters in existing multivariate random effects models. A closed-form likelihood allows the user to diagnose model identifiability relatively easily by evaluating the properties of the Fisher information matrix for parameter regions of interest.

A useful special case of the complementary log-log – multivariate gamma model is an autoregressive version for binary time series analysis. Cox (1981) classified time-series models for serially-correlated data into two classes, namely observation-driven and parameter-driven models. Observation-driven models specify the conditional distribution of a response at time t as a function of past responses, and are typically straightforward to fit (Diggle et al., 2002). In contrast, parameter-driven models specify an underlying serially correlated latent process and are typically much more difficult to fit. Existing approaches to fitting this class of models include Monte Carlo EM (Chan & Ledolter, 1995) and a fully Bayesian Markov chain Monte Carlo analysis (Diggle et al., 2002). Such Monte Carlo methods introduce a new set of computational issues requiring careful attention, such as prior elicitation and convergence properties

of the Markov chains.

2 A Multivariate Random Effects Model for Binary Data

We formulate the model using a vector of multivariate gamma random effects, as defined by Henderson & Shimakura (2003). Let $W_1, \ldots, W_q$ be independent p-variate Gaussian with standard marginals and common $p \times p$ correlation matrix C. Write $W_j = (W_{j1}, \ldots, W_{jp})'$ and let $Z_k = \sum_{j=1}^{q} W_{jk}^2/q$, for $k = 1, \ldots, p$. Then the vector $Z = (Z_1, \ldots, Z_p)'$ is said to be multivariate gamma with marginal $\mathrm{Ga}(q/2, q/2)$ distributions and Laplace transform

$$\mathcal{L} = E\left\{\exp\left(-u'Z\right)\right\} = \left|I + 2C\mathrm{diag}(u)/q\right|^{-q/2}, \tag{2.1}$$

for $u \in \mathcal{R}^n$ and $C = (c_{jk})$.

A large literature exists on the properties of the distribution defined by (2.1). Bapat (1989) showed that, for suitable choices of C, (2.1) defines a proper probability distribution more generally for noninteger values of q. He showed that, if there exists some diagonal matrix M having elements equal to 1 or -1 on the diagonal such that $(MCM)^{-1}$ has nonpositive off-diagonal elements and MCM has positive entries, then (2.1) defines an infinitely divisible distribution for any $q > 0$. If we let $\zeta = 2/q$, the resulting multivariate distribution with Laplace transformation

$$\mathcal{L} = E\left\{\exp\left(-u'Z\right)\right\} = \left|I + \zeta C\mathrm{diag}(u)\right|^{-1/\zeta}$$

defines a proper multivariate distribution for all $\zeta > 0$. Marginally, $Z_j \sim \mathrm{Ga}(1/\zeta, 1/\zeta)$, $j = 1, \ldots, n$, with correlation matrix describing the association among gamma variables equal to R with elements $r_{jk} = c_{jk}^2$. We denote this multivariate distribution by $Z \sim MG(\zeta, C)$.

Let Y_{ij} denote binary response j, $j = 1, \ldots, n_i$, in cluster i, $i = 1, \ldots, N$. Let $\theta_{ij} = \log(Z_{ij})$ be a random effect corresponding to Y_{ij}, and consider the generalised linear mixed model

$$\log\left[-\log\left\{\mathrm{E}\left(Y_{ij}|Z_i\right)\right\}\right] = \theta_{ij} + x_{ij}'\beta, \tag{2.2}$$

where x_{ij} is a $k \times 1$ vector of covariates associated with response j in cluster i, β is a $k \times 1$ vector of fixed effects, and $Z_i \sim MG(\zeta, C_i)$, independently over i, with C_i an $n_i \times n_i$ association matrix for subject i. In this framework, ζ is an overdispersion parameter, the interpretation of which we address in detail in § 4. Interest typically focuses on both the fixed effects β and the correlation matrix C_i parameterised as a known function of an $r \times 1$ vector of variance components ρ.

Under the generalised linear mixed model (2.2), the marginal probability of a response is

$$\text{pr}(Y_{ij} = 1) = \int \text{pr}(Y_{ij} = 1|Z) f(Z) dZ.$$

Although there exists no closed-form for $f(Z)$, note that

$$\begin{aligned} \text{pr}(Y_{ij} = 1) &= \int \exp\left\{-\exp\left(\theta_{ij} + x'_{ij}\beta\right)\right\} f(Z) dZ \\ &= \int \exp\left(-u'_{i,j} Z\right) f(Z) dZ \\ &= \left|I + \zeta C_i \text{diag}(u_{i,j})\right|^{-1/\zeta}, \end{aligned}$$

for vector $u_{i,j}$ having $\exp(x'_{ij}\beta)$ in position j and 0 elsewhere. Thus, an expression for the marginal, averaged over the random effects, probability of an event for a single observation exists in closed form under this model.

In order to derive the joint probability $\pi_{i,y} \equiv \pi_{i,\left(y_1 \ldots y_{n_i}\right)} = \text{pr}(Y_{i1} = y_1, Y_{i2} = y_2, \ldots, Y_{in_i} = y_{n_i})$, we use the method of Conaway (1990) that first computes marginal probabilities in the 2^{n_i} table formed by cross-classifying the binary responses in a given cluster, and subsequently transforms these marginal probabilities back to the joint probabilities of interest. Let T be a subset of the indices $\{1, 2, \ldots, n_i\}$. We define

$$\pi^*_{i,T} = \int \prod_{j \in T} \text{pr}(Y_{ij} = 1|Z) f(Z) dZ.$$

For example, for $n = 3$, $\pi^*_{i,\{1,2,3\}} = \text{pr}(Y_{i1} = 1, Y_{i2} = 1, Y_{i3} = 1)$, $\pi^*_{i,\{1,2\}} = \text{pr}(Y_{i1} = 1, Y_{i2} = 1)$ and $\pi^*_{i,\{1\}} = \text{pr}(Y_{i1} = 1)$. By the same arguments as above, these proba-

bilities also have closed form:

$$\begin{aligned}\pi^*_{i,T} &= \int \exp\left\{-\sum_{j\in T} Z_{ij}\exp\left(x'_{ij}\beta\right)\right\} f(Z)dZ \\ &= \left|I+\zeta C_i \mathrm{diag}(u_{i,T})\right|^{-1/\zeta},\end{aligned}$$

where now the jth element of $u_{i,T}$ equals $\exp(x'_{ij}\beta)$ if $j \in T$ and is 0 otherwise. Thus, only changes in the elements of $u_{i,T}$ are necessary to reflect differences among specific $\pi^*_{i,T}$. If $\pi^*_i = \left(\pi^*_{i,\{1\ldots n\}}, \pi^*_{i,\{2\ldots n\}}, \ \pi^*_{i,\{1,3,\ldots,n\}}, \ldots, \pi^*_{i,\{\emptyset\}}\right)'$ is the collection of all such marginal probabilities $\pi^*_{i,T}$, then the vector of joint probabilities π_i is a known linear transformation of π^*_i. For instance, for clusters of size $n = 3$ with $\pi^* = \left(\pi^*_{i,\{1,2,3\}}, \pi^*_{i,\{2,3\}}, \pi^*_{i,\{1,3\}}, \pi^*_{i,\{3\}}, \pi^*_{i,\{1,2\}}, \pi^*_{i,\{2\}}, \ \pi^*_{i,\{1\}}, \pi^*_{i,\{\emptyset\}}\right)'$, the probabilities π^* satisfy $\pi^* = A\pi$, where

$$A = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

and $\pi_i = \left(\pi_{i,(111)}, \pi_{i,(011)}, \pi_{i,(101)}, \pi_{i,(001)}, \pi_{i,(110)}, \pi_{i,(010)}, \pi_{i,(100)}, \pi_{i,(000)}\right)'$. Thus, $\pi_i = A^{-1}\pi^*_i$. The maximum likelihood estimates $\left(\widehat{\beta}', \widehat{\rho}', \widehat{\zeta}\right)$ are those values of the parameters that maximise the loglikelihood $l = \sum_{i=1}^N l_i$, where l_i is the log of the element of π_i corresponding to the observed response pattern for cluster i. We maximise this loglikelihood using numerical optimisation methods as implemented in the `optim` function in the R software package (R Development Core Team, 2003), and base inference on the inverse Hessian matrix for (β', ρ', ζ), evaluated at the maximum likelihood estimates. R programs for implementing the models and associated documentation are available from the web at `http://www.biostat.harvard.edu/~ahousema/software/mvg.htm`.

3 Prediction

In some instances, interest focuses on prediction of the random effects (Robinson, 1991). Standard practice in generalised linear mixed modelling uses the empirical

Bayes predictions of Z_i for prediction. These quantities are estimators of the posterior mean, $E\left(Z_i|Y_i\right)$, of the random effects Z_i given the observed data Y_i. In addition to a closed-form for the likelihood, the proposed complementary log-log multivariate gamma formulation has the advantage that it yields closed-form expressions for these predictions.

For a fixed cluster i, let $\mathcal{Y}_{n_i} = \{y : y = (y_1, \ldots, y_{n_i}), y_j \in (0,1), j = 1, \ldots, n_i\}$, and let e_y be the $2^{n_i} \times 1$ vector such that $\pi_{i,(y_1,\ldots,y_{n_i})} = e_y'\pi_i = e_y'A^{-1}\pi_i^*$. Furthermore, let $\pi_{i,y|Z} = \mathrm{pr}\left(Y_i = y|Z_i\right)$ and let $\pi_{i|Z}$ be the vector of all such probabilities ranging over $\mathcal{Y}_{n_i}$, ordered as in π_i. For subset T, let $\pi^*_{i,T|Z} = \prod_{j\in T}\mathrm{pr}\left(Y_{ij} = 1|Z_i\right)$, and let $\pi^*_{i|Z}$ be the vector containing all such probabilities in the order analogous to π^*. As shown in the Appendix, the empirical Bayes prediction for Z_{ij} is

$$E\left(Z_{ij}|Y_i = y\right) \;=\; \pi_{i,y}^{-1}e_y'A^{-1}\dot{L}_{ij}^*,$$

where $\dot{L}_i^*$ is the $2^{n_i} \times 1$ vector with elements

$$\begin{aligned}\dot{L}^*_{ij,T} &= \left.\frac{\partial}{\partial t_j}\mathcal{L}(t)\right|_{t=u_{i,T}} \\ &= |I + \zeta C_i \mathrm{diag}(u_{i,T})|^{-1/\zeta}\mathrm{tr}\left[\left\{I + \zeta C_i\mathrm{diag}(u_{i,T})\right\}^{-1}C_iE_j\right],\end{aligned}$$

for $t \in \mathcal{R}^n$ and $E_j = \mathrm{diag}\left(\partial t/\partial t_j\right)$. We have incorporated these predictions into our software that implements the model.

4 Parameter Interpretation and Identifiability

4.1 Interpretation of model parameters

Individually, the variance components (ρ', ζ) do not have straightforward interpretations, as they jointly parameterise the association structure of Y_i. However, primary scientific interest typically focuses on the overall structure of the within-cluster associations, and not the individual components that parameterise this structure. Thus, this joint parameterisation does not hinder the utility of the model. Since the model yields closed forms for the estimated joint probability distribution for a given cluster, we obtain closed-form expressions for the association structure in a familiar parameterisation such as log odds ratios, with these log odds ratios values specific to a given

pattern of the covariates in the model. We employ this strategy to obtain fitted log odds ratios for complex association structures in §§ 5.1 and 5.2.

Compared to a standard logistic model having log odds ratios as regression coefficients, interpretation of β in model (2.2) is also nonstandard. In the complementary log-log formulation, a positive value of a regression coefficient indicates a negative association between the corresponding covariate and the probability of response. Again, this is not a problem for interpretative purposes, as one can investigate the effect of a particular covariate on the joint distribution of Y_i in this model formulation. This advantage of the model allows the user to report estimates of the effect expressed either conditionally on the random effects or marginally in terms of the joint probability distribution of Y_i.

4.2 Parameter identifiability

For concreteness, we focus on the first order-autoregressive correlation structure $c_{ik} = \rho^{|t_i - t_k|}$, although similar reasoning applies for other correlation structures such as the compound symmetric structure $c_{ik} = \rho$. We focus on the intercept-only model

$$\log\{-\log(\pi_{ij})\} = \beta_0 + \theta_{ij}. \tag{4.1}$$

As pointed out by a referee, it is instructive to consider the latent response formulation for models with the complementary log-log link (Agresti, 2002, §6.6.4). The model for an underlying continuous response Y_{ij}^* can be written as

$$Y_{ij}^* = \beta_0 + \theta_{ij} + \epsilon_{ij}, \tag{4.2}$$

where $-\epsilon_{ij}$ has a Gumbel distribution with scale parameter 1, which yields variance of $\pi^2/6$, and the observed response Y_{ij} is 1 if $Y_{ij}^* > 0$. Since β_0 parameterises the mean of the Y_{ij}^*, the variance components ρ and ζ can only be identified through the correlation structure for $Y_i^* = \left(Y_{i1}^*, \ldots, Y_{in_i}^*\right)'$, if we assume that higher-order moments provide neglible information. When $\rho = 0$, the variance of θ_{ij}, or equivalently ζ, is not well identified because this variance does not relate to the correlations of Y_{ij}^*. In contrast, the special case of the model with $\rho = 1.0$ corresponds to a univariate random intercept model. In this case, ζ represents the variance component for the random intercepts in the model, and is clearly identifiable. Thus, identifiability

of the model parameters depends on the strength of the serial association among clustered responses, with the model being weakly identifiable, in the sense of high correlations between some pairs of parameters, for a wide range of ρ values within the two extremes.

More rigorously, we investigate the asymptotic identifiability of all model parameters in model (4.1). We do this by examining the Fisher information contained in one cluster for this model. For known values of β_0, ρ and ζ, we can easily calculate the values of each element in the Fisher information matrix, the condition number of the matrix and the asymptotic correlations among parameter estimates obtained from data generated from the model. Figure 1 shows the condition number of the Fisher information matrix over a wide range of ρ values, for the fixed value of $\zeta = 2.0$. Analogous results exist for different values of ζ, as can be seen if one plots the surface formed by this condition number as a function of ρ and ζ, not shown, and different values of β_0. The plot shows that the model that results from leaving ζ free to be estimated is well conditioned as long as ρ is greater than approximately 0.75, but that the condition number grows without bound as $\rho \to 0$. Figure 1 also shows the condition number for the Fisher information matrix for model (4.1) as a function of ρ when ζ is not treated as an unknown parameter. The figure shows that this constrained formulation results in a well-conditioned model for all values of ρ. The results of this exact calculation confirm the heuristic arguments suggested by latent response model (4.2): all model parameters are identifiable for some regions of the parameter space, and, for regions for which they are not, fixing ζ to a prespecified value results in an identified model. Although we demonstrate this strategy in the context of a specific autoregressive model, one can use it to investigate the theoretical identifiability of a model with any such structure for C.

Of course, the asymptotic arguments above do not ensure that the multivariate random effects model will be identifiable for a given finite sample. To address cases of weak identifiability in a given application, we propose first fitting the unconstrained model to the data and performing a battery of identifiability diagnostics on the resulting model fit, including inspection of the correlations among the parameter estimates and the condition number of the associated variance covariance matrix. The theoretical arguments above and our practical experience suggest that, in instances of strong

clustering, the resulting model fit is well conditioned. In cases in which the model is weakly identified, we propose refitting the model fixing the overdispersion parameter ζ at some value larger than the maximum likelihood estimate $\widehat{\zeta}$ obtained from the unconstrained fit. This ensures that we do not artificially constrain the magnitude of the within-cluster associations from above. This approach of fixing some parameters to arrive at a fully identified model is a standard approach in other latent response settings, such as the probit model (Agresti, 2002, §6.6) and the multivariate logistic-normal model (Rabe-Hesketh & Skrondal, 2001). In general, the fixed effect estimate $\widehat{\beta}$ will depend on the chosen value of ζ. However, this is not really a drawback for two reasons. First, for larger estimates, the corresponding standard error is also larger, so that conclusions concerning the strength of association between a response and a covariate are relatively invariant to the choice of ζ. Secondly, because the fitted joint probability distribution is easily calculated, one can express these associations using marginal odds ratios calculated from the joint probability distribution of Y_i. Since the fitted values are insensitive to choice of ζ when it is empirically unidentified, so are the estimates of the marginal effects of interest.

We stress that the above identifiability considerations are not unique to the complementary log-log multivariate gamma model considered here, but also apply to other multivariate random effects models with analogous covariance structures for the random effects. Diggle et al. (2002) considered a fully Bayesian analysis of the analogous logistic-normal autoregressive model, but, presumably to produce identifiable model parameters, placed a relatively sharp prior distribution of IG$(2, 2)$ on the random effects standard deviation. This Bayesian strategy of specifying sharp priors for weakly identified parameters has been proposed in other settings (Aitkin & Stansopolis, 1989). We view the fact that the complementary log-log model yields straightforward evaluation of model identifiability as a strength of the model as compared to existing multivariate random effects formulations for clustered binary data.

5 Applications

5.1 Example 1: Familial aggregation

This example demonstrates the ease with which one can use the model to condition on the response of a proband in case-control family studies, and thus adjust for nonrandom ascertainment. In familial aggregation studies, interest focuses on the association structure among disease indicators from members within the same family. A popular existing approach is the quadratic exponential model of Zhao & Prentice (1990). However, interpretation of parameters from this model is difficult when the cluster sizes vary, which is invariably the case in family studies (Betensky & Whittemore, 1996). In contrast, random effect models work well when the cluster sizes vary.

A second common complication in familial aggregation studies is the use of nonrandom sampling schemes, such as in a case-control design. This design samples individuals, known as probands, based on their disease status and subsequently obtains data on the family members of each proband in the study. The proper likelihood contribution from each family is the conditional distribution of that family's responses, conditional on the disease status of the proband. As a result, for correct inference we require the marginal probability of the proband's response. If the proband is identified as subject 1 in each family, the required marginal probability for this conditional probability is $\pi^*_{\{1\}}$, which is easily obtained under model (2.2). The resulting likelihood contribution for family i is $L_i / \left\{ \left(\pi^*_{\{1\}}\right)^{y_{i1}} \left(1-\pi^*_{\{1\}}\right)^{(1-y_{i1})} \right\}$, where L_i is the likelihood based on the full joint distribution for cluster i.

Here, we analyze data on the familial aggregation of the combined disease outcome of breast or ovarian cancer in women (Betensky & Whittemore, 1996). We fit the model that adjusts for nonrandom ascertainment to data from 5756 families, with each family consisting of a proband, the proband's mother, and the proband's sisters. The families range in size from two, just proband and mother, to six, made up of proband, mother and four sisters, with 384 'case' families, with proband's disease status = 1, and 5372 'control' families, with proband's disease status = 0.

One question of interest is whether or not the association among disease indicators from different family members depends on the relationship between the subjects. For

instance, in simple genetic settings, both a parent and child as well as two siblings share 50% of their genes on average, suggesting a simple compound symmetric structure (Andersen, 2004). For more complex diseases, it may be that parent-child pairs exhibit stronger dependence than do siblings. We fit the proposed complementary log-log model to evaluate the association structure among disease statuses of different family members. We consider the model with the family-specific covariate 'race' as a fixed effect and a covariance matrix C_i that specifies a correlation of ρ_{SS} for sister-sister pairs and $\rho_{SS}^{1/2}\rho_{MS}$ for mother-daughter pairs. This multiplicative form for the mother-daughter association satisfies the conditions on C_i necessary to ensure that (2.1) yields a proper probability distribution for all $0 \leq \rho_{MS}, \rho_{SS} \leq 1$. We focus on the estimates of association from this model, and whether or not there is evidence against the special case with $\rho_{MS} = \rho_{SS}^{1/2}$, which corresponds to the simpler compound symmetric covariance structure. Preliminary fits show that the models with ζ left to be freely estimated are weakly identified, with condition number of the estimated variance-covariance matrix being equal to 11658.0 and the estimated correlation between $\widehat{\beta}_0$ and ζ equal to 0.99. Thus, we fit the full model constraining $\zeta = 1.0$, which yields a condition number of 16.9. The model fit yields $\widehat{\rho}_{MS} = 1.0$, with standard error 0.12, and $\widehat{\rho}_{SS} = 0.50$, with standard error 0.09, which for the estimated intercept corresponds to log odds ratios of 1.99 for mother-daughter associations and 1.32 for sibling associations. These estimates are almost identical to those from the unconstrained model, which are 2.01 and 1.29, respectively. The difference between the deviance of this two-correlation model with $\zeta = 1$ and that from the simpler compound symmetric model, also fitted under the constraint $\zeta = 1.0$, is 9.54, providing strong evidence that these two familial associations differ for breast/ovarian cancer. These results are qualitatively similar to those obtained by Betensky & Whittemore (1996), who showed that these familial associations differed when one considered breast and ovarian cancer individually.

To assess the impact of properly accounting for the study design in the analysis, we re-fit the model without conditioning on the proband's observed response in each family. This incorrect analysis, also fitted constraining $\zeta = 1.0$, estimates the familial aggregation log odds ratios to be 1.13 for sister-sister pairs and 1.52 for mother-daughter pairs. Thus, once we correctly condition on the proband's response to

account for nonrandom sampling, the analysis suggests stronger familial aggregation of breast/ovarian disease status for both types of familial relationship.

5.2 Example 2: Brain tumour genetics

This is a case in which interest focuses on complex correlation structures for a relatively high-dimensional multivariate outcome. Loss of heterozygosity of chromosomal regions of tumours, a binary outcome, is of interest as it is suggestive of the presence of a tumour suppressor gene. Allelic losses on chromosome 1p have been frequently found in oligodendrogliomas, a common variant of brain tumour. Furthermore, loss of heterozygosity on chromosome 1p is of prognostic interest, as it has been shown to be highly associated with response to chemotherapy and long survival in patients with certain malignant brain tumours (Cairncross et al., 1998; Ino et al., 2001). Previous analyses of loss of heterozygosity in oligodendroglioma used three CA-repeat polymorphism markers to assess loss of heterozygosity of the whole chromosome arm. An entire chromosome arm was assumed to be lost if loss of heterozygosity was observed at all informative markers on that arm. Recently, a 'medium throughput' quantitative method for assessing loss of heterozygosity at 19 non-distal, approximately equally-spaced markers on two chromosomes has been developed. The markers consist of 15 markers from chromosome 1p, five of which are from the 'tip' of chromosome 1p, and 4 from chromosome 19q. The measurements were recorded on $N = 85$ brain tumours. One question of interest is whether segments of these chromosome arms, and not the entire arms, may be lost in some cases; that is, is there heterogeneity in the binary loss of heterozygosity outcomes across the two chromosomes, and, in particular, does this association among loss of heterozygosity outcomes vary according to location on chromosome 1p, or according to chromosome?

Since interest focuses on the strength of association as a function of the locations of two loss of heterozygosity outcomes, we consider an intercept-only complementary log-log multivariate gamma model with a correlation structure that specifies unique correlation parameters for both the intra- and inter-chromosomal associations. We refer to the tip of chromosome 1p as chromosome 1A and the remaining markers as chromosome 1B. Not all markers are informative for all tumours; these missing data are missing completely at random. Thus, let Y_{ij} denote the loss of heterozygosity

outcome at location j, $j = 1, \ldots, n_i$, on tumour i, $i = 1, \ldots, 85$. The model is

$$\log\left[-\log\left\{E\left(Y_{ij}|Z_i\right)\right\}\right] = \beta_0 + \theta_{ij}, \tag{5.1}$$

where $\theta_i = (\theta_{i1}, \ldots, \theta_{in_i})' \sim MG(\zeta, C_i)$, independently for each i. Although one might presume that loss of heterozygosity in 1p and 19q are independent, it is well known that the outcome is highly associated across these two chromosomes. Thus, we assume correlation structure $C_{\text{full}} = (c_{jk})$, such that

$$\begin{array}{ll}
c_{jk} = \rho_{1A} & \text{for } j,k \in \text{chromosome 1A} \\
c_{jk} = \rho_{1B} & \text{for } j,k \in \text{chromosome 1B,} \\
c_{jk} = \rho_{19} & \text{for } j,k \in \text{chromosome 19} \\
c_{jk} = \left(\rho_{1A}\rho_{1B}\right)^{1/2}\rho_{1A,1B} & \text{for } j \in \text{chromosome 1A, } k \in \text{chromosome 1B} \\
c_{jk} = \left(\rho_{1A}\rho_{19}\right)^{1/2}\rho_{1A,19} & \text{for } j \in \text{chromosome 1A, } k \in \text{chromosome 19} \\
c_{jk} = \left(\rho_{1B}\rho_{19}\right)^{1/2}\rho_{1B,19} & \text{for } j \in \text{chromosome 1B, } k \in \text{chromosome 19,}
\end{array}$$

for each cluster.

As in the first two examples, diagnostics for preliminary fits indicate that ζ, estimated as $\widehat{\zeta} = 2.3$, is weakly identified in the presence of β_0, with the condition number of the corresponding variance matrix being 15682.9 and the estimated correlation between the two estimates being 0.60. Table 1 shows the results of fitting the model to the data from the 19 markers, with ζ fixed at 2.5. This constrained model has a condition number of 1262.5. The first two columns of the table report the parameter estimates and associated standard errors for the correlation parameters. The third column reports the odds ratios implied by the above multiplicative correlation structure for each type of association. These estimates also hold for the unconstrained model. We see that the odds ratios implied by the correlation parameters range from 3.84 for the 1A and 19 association up to 9.89 for two markers on chromosome 1B. The results indicate that the within- and between-chromosome associations in loss of heterozygosity are strong. Interest focuses on whether this full model is necessary, or whether we can model the association structure among the 19 markers with a compound symmetric structure. The simpler compound symmetry model is a special case of the full model, holding when $\rho_{1A} = \rho_{1B} = \rho_{19} \equiv \rho$ and $\rho_{1A,1B} = \rho_{1A,19} = \rho_{1B,19} = \rho^2$. Thus we can assess whether or not the more complicated model provides a significantly better fit via likelihood ratio testing. The likelihood ratio statistic is 14.44 on 5 degrees of freedom, yielding strong evidence

that the full model is necessary. Thus, the pairwise associations among loss of heterozygosity markers vary according to location on chromosomes 1p and 19q.

6 Discussion

The multivariate gamma formulation used here is related to those used to represent multivariate frailties in correlated lifetime models (Hougaard, 2000, Ch. 10). That approach is useful in that it can yield specific forms for the correlation matrix C, but is somewhat less flexible than the direct correlation specification outlined here since certain correlation structures are not possible using simple sums. Henderson & Shimakura (2003) noted that the joint distributions based on the direct and additive correlation structures have the same marginal and association properties. These authors also noted that the differences between the joint distributions represented by these two constructions are generally small except in the tails. Thus, we anticipate differences in inferences obtained from latent variable models using these distributions also to be small.

A potential disadvantage of the model is the fact that the multivariate gamma distribution does not accommodate negative correlations. This is not a severe limitation, however, since such correlation structures can often be handled with relatively low-dimensional factor-analytic models (Skrondal & Rabe-Hesketh, 2004, Ch. 9), whereby a single latent variable is multiplied by fixed effects. When some of these parameters, or 'factor loadings', are negative, the latent variable induces negative correlations among some of the responses within the same cluster. Since such models often contain one or two latent variables, they can often be fitted easily using numerical integration, for example by PROC NLMIXED in SAS or `gllamm` in STATA. In contrast, our approach is appropriate when computation and the establishment of identifiability is difficult because of the dimension of the random effects.

Although it is computationally feasible to fit the model to the large majority of longitudinal or otherwise clustered datasets, there are computational limits since the computations are linear in 2^{n_i}. Thus, in situations with very large 'clusters', such as long binary time series or intervention trials performed at the school or community level, these methods are less applicable. For long binary time series, we have used a

pseudolikelihood approach to estimation based on the complementary log-log – multivariate gamma formulation. This approach, also used by Henderson & Shimakura (2003) for fitting other multivariate gamma models, bases inference on a set of estimating equations, where subsets of clusters of more manageable size are treated as new pseudo-clusters. Our R software implements these pseudolikelihood routines as well. Our model may also be useful in spatial settings and mixed-model formulations of regression splines for binary responses.

ACKNOWLEDGEMENT

This research was supported in part by grants from the U. S. National Institute of Environmental Health Sciences and National Cancer Institute. The authors thank O. Bogler, J.G. Cairncross and D.N. Louis for use of the loss of heterozygosity data and for helpful feedback, and three referees for insightful comments that significantly improved the manuscript.

APPENDIX

Derivation of the Empirical Bayes predictions of the random effects

For fixed cluster i, let $\pi_{i,y|Z} = \text{pr}\,(Y_i = y|Z_i)$ and let $\pi_{i|Z}$ be the vector of all such probabilities, ordered as in π_i. Let $\pi^*_{i|Z}$ be the corresponding vector containing elements $\pi^*_{i,T|Z} = \prod_{j\in T} \text{pr}\,(Y_{ij} = 1|Z_i)$. Finally, following the notation in § 3, let e_y be the $2^{n_i} \times 1$ vector such that $\pi_{i,y} = e_y'\pi_i = e_y'A^{-1}\pi^*_i$. Note that

$$\begin{aligned} E\left(Z_{ij}\pi_{i,y|Z}\right) = E\left(Z_{ij}e_y'\pi_{i|Z}\right) &= E\left(Z_{ij}e_y'A^{-1}\pi^*_{i|Z}\right) \\ &= E\left(e_y'A^{-1}\pi^*_{i|Z}Z_{ij}\right) \\ &= e_y'A^{-1}E\left(\pi^*_{i|Z}Z_{ij}\right). \end{aligned}$$

Here, $E\left(\pi^*_{i|Z}Z_{ij}\right)$ can be obtained by differentiating the Laplace transform $\mathcal{L}(t)$. Since $\pi^*_{i,T|Z} = \exp(-Z_i'u_{i,T})$ and

$$Z_{ij}\exp(-Z_i'u_{i,T}) = \frac{\partial}{\partial t_j}\exp(-Z_i't)\bigg|_{t=u_{i,T}},$$

assuming interchangeability of the differential and integral operators, we have

$$E\left(\pi^*_{i|Z}Z_{ij}\right) = E\left\{Z_{ij}\exp(-Z_i'u_{i,T})\right\} = \frac{\partial}{\partial t_j}\mathcal{L}(t_j)\bigg|_{t=u_{i,T}}.$$

Note that

$$\frac{\partial}{\partial t_j}\mathcal{L}(t_j) = |I - \zeta C_i \text{diag}(t)|^{-1/\zeta} \text{tr}\left[\{I - \zeta C_i \text{diag}(t)\}^{-1} C_i E_j\right],$$

where $E_j = \text{diag}(\partial t/\partial t_j)$.

Thus, if $f(Z_i)$ is the joint distribution of Z_i, then the posterior distribution of Z_i given $Y_i = y$ is equal to $\pi_{i,y}^{-1}\left\{\pi_{i,y|Z} f(Z_i)\right\}$, and the posterior mean of Z_{ij} is equal to

$$\begin{aligned} E\left(Z_{ij}|Y_i = y\right) &= E\left\{\pi_{i,y}^{-1}\left(Z_{ij}\pi_{i,y|Z}\right)\right\} \\ &= \pi_{i,y}^{-1} E\left(Z_{ij}\pi_{i,y|Z}\right) \\ &= \pi_{i,y}^{-1} e_y' A^{-1} E\left(\pi_{i|Z}^* Z_{ij}\right) \\ &= \pi_{i,y}^{-1} e_y' A^{-1} \dot{L}_{ij}^*, \end{aligned}$$

where $\dot{L}_i^*$ is the $2^{n_i} \times 1$ vector with elements

$$\begin{aligned} \dot{L}_{ij,T}^* &= \left.\frac{\partial}{\partial t_j}\mathcal{L}(t)\right|_{t=u_{i,T}} \\ &= |I + \zeta C_i \text{diag}(u_{i,T})|^{-1/\zeta} \text{tr}\left[\{I + \zeta C_i \text{diag}(u_{i,T})\}^{-1} C_i E_j\right]. \end{aligned}$$

REFERENCES

Agresti, A. (1997). A model for repeated measurements of a multivariate binary response. *J. Am. Statist. Assoc.* **92**, 315–21.

Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. New York: Wiley.

Aitchison, J. & Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika* **76**, 643–53.

Aitkin, M. & Alfo, M. (1998). Regression models for binary longitudinal responses. *Statist. Comp.* **8**, 289–307.

Aitkin, M. & Stasinopoulos, M. (1989). Likelihood analysis of a binomial sample size problem. In *Contributions to Probability and Statistics. Essays in Honor of Ingram Olkin*, Ed. L. J. Gleser, M. D. Perlman, S. J. Press and A. R. Simpson, pp. 399–411. New York: Springer-Verlag.

Andersen, E. W. (2004). Composite likelihood and two-stage estimation in family studies. *Biostatistics* **5**, 15–30.

Bapat, R. B. (1989). Infinite divisibility of multivariate gamma distributions and M-matrices. *Sankyha* A **51**, 73–8.

Betensky, R. A. & Whittemore, A. S. (1996). An analysis of correlated multivariate binary data: Application to familial cancers of the ovary and breast. *Appl. Statist.* **45**, 411–29.

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Am. Statist. Assoc.* **88**, 9–25.

Cairncross, J. G., Ueki, K., Zlatescu, M. C., Lisle, D. K., Finkelstein, D. M., Hammond, R. R., Silver, J. S., Stark, P. C., Macdonald, D. R., Ino, Y., Ramsay, D. A., & Louis, D. N. (1998). Specific genetic predictors of chemotherapeutic response and survival in patients with anaplastic oligodendroglioma. *J. Nat. Cancer Inst.* **90**, 1473–9.

Chan, K. S. & Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *J. Am. Statist. Assoc.* **90**, 242–52.

Conaway, M. R. (1990). A random effects model for binary data. *Biometrics* **46**, 317–28.

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

Cox, D. R. (1981). Statistical analysis of time-series – Some recent developments. *Scand. J. Statist.* **8**, 93–115.

Diggle, P. J., Heagerty, P., Liang, K.-Y. & Zeger, S. L. (2002). *Analysis of Longitudinal Data*, 2nd ed. Oxford: Clarendon Press.

Ekholm A., McDonald J. W. & Smith, P. W. F. (2000). Association models for a multivariate binary response. *Biometrics* **56**, 712–8.

Fitzmaurice, G. M. & Lipsitz, S. R. (1995). A model for binary time-series data with serial odds ratio patterns. *Appl. Statist.* **44**, 51–61.

Henderson, R. & Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika* **90**, 355–66.

Henderson, R., Shimakura, S. & Gorst, D. (2003). Modeling spatial variation in leukemia survival data. *J. Am. Statist. Assoc.* **97**, 965–72.

Hougaard, P. (2000). *Analysis of Multivariate Failure Time Data.* New York: Springer.

Hudson, J. I., Laird, N. M. & Betensky, R. A. (2001). Multivariate logistic regression for familial aggregation of two disorders. I. Development of models and methods. *Am. J. Epidem.* **153**, 500–5.

Humphreys, K. (1998). The latent Markov chain with multivariate random effects: An evaluation of instruments measuring labor market status in the British Household Panel Study. *Soc. Meth. Res.* **26**, 269–99.

Ino, Y., Betensky, R. A., Zlatescu, M. C., Sasaki, H., Macdonald, D. R., Stemmer-Rachamimov, A. O., Ramsay, D. A., Cairncross, J. G. & Louis, D. N. (2001). Molecular subtypes of anaplastic oligodendroglioma: implications for patient management at diagnosis. *Clin. Cancer Res.* **7**, 839–45.

Lipsitz, S. R., Fitzmaurice, G. M., Sleeper, L. & Zhao, L. P. (1995). Estimation methods for the joint distribution of repeated binary observations. *Biometrics* **51**, 562–70.

Lipsitz, S. R., Fitzmaurice, G. M., Sleeper, L. & Zhao, L. P. (1996). Estimating the joint distribution of repeated binary responses: Some small sample results. *Comp. Statist. Data Anal.* **23**, 219–27.

R Development Core Team (2003). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org`.

Rabe-Hesketh, S. & Skrondal, A. (2001). Parameterization of multivariate random effects models for categorical data. *Biometrics* **57**, 1256–63.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with Discussion). *Statist. Sci.* **6**, 15–51.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*. Boca Raton, FL: Chapman and Hall/CRC.

Spiegelhalter, D., Thomas, A. & Best, N. (2000). *WinBUGS Version 1.3. User's Manual*, MRC Biostatistics Unit. Institute of Public Health, Cambridge. `http://www.mrc-bsu.cam.ac.uk/bugs`.

Zhao, L. P. & Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika* **77**, 642–8.

Table 1: Maximum likelihood estimates $\widehat{\rho}$ and associated standard errors from the model applied to the brain tumour data. The third column presents the corresponding odds ratios for each type of association based on the correlation model C_{full} for the data.

Correlation Parameter	Estimate	Std. Err.	Corresponding Pairwise Odds Ratio
ρ_{1A}	0.92	0.03	5.25
ρ_{1B}	0.98	0.01	9.89
ρ_{19}	0.94	0.04	6.36
$\rho_{1A,1B}$	0.99	0.01	6.83
$\rho_{1A,19}$	0.93	0.04	3.84
$\rho_{1B,19}$	0.97	0.02	5.99

Std. Err., standard error

Figure 1: Plot of the condition number of the Fisher Information matrix for one cluster of size $n = 5$ for autoregressive model (4.1) as a function of autoregressive parameter ρ. The dotted line represents the result for the model that leaves ζ free, and the solid line represents the result for the model that treats ζ fixed.