Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Jamsen et al., J Biomet Biostat 2012, S1
DOI: 10.4172/2155-6180.S1-003

Research Article Open Access

Specification of Generalized Linear Mixed Models for Family Data using Markov Chain Monte Carlo Methods

Kris M Jamsen[1]*, Sophie G Zaloumis[1,2], Katrina J Scurrah[1,2] and Lyle C Gurrin[1]

[1]*Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, Melbourne School of Population Health, The University of Melbourne, Victoria 3010, Australia*
[2]*Department of Physiology, The University of Melbourne, Victoria 3010, Australia*

Abstract

Statistical models imposed on family data can be used to partition phenotypic variation into components due to sharing of both genetic and environmental risk factors for disease. Generalized linear mixed models (GLMMs) are useful tools for the analysis of family data, but it is not always clear how to specify individual-level regression equations so that the resulting within-family variance-covariance matrix of the phenotype reflects the correlation implied by the relatedness of individuals within families. This is particularly challenging when families are of varying sizes and compositions. In this paper we propose a general approach to specifying GLMMs for family data that uses a decomposition of the within-family variance-covariance matrix of the phenotype to set up a series of regression equations with fixed and random effects that corresponds to an appropriate genetic model. This "mechanistic" specification is particularly suited to estimation and evaluation of models within a Markov chain Monte Carlo (MCMC) framework. The proposed approach was assessed with simulated data to demonstrate the accuracy of estimation of the variance components. We analyzed data from the Victorian Family Heart Study (families with two generations over-sampled for those with monozygotic and dizygotic twins) and for a binary phenotype (hypertension) that resulted in substantially reduced computation time in the MCMC framework (via WinBUGS) compared with a maximum likelihood approach (via Stata). The proposed approach to model specification in this paper is easily implemented using standard software and can accommodate prior information on the magnitude of fixed or random effects.

Keywords: Family data; Generalized linear mixed models; Model specification; Markov chain Monte Carlo

Introduction

High-throughput genotyping technology now provides epidemiologists with an unprecedented opportunity to explore the association between measured genetic variants and the risk of disease. For most common complex diseases of adult life, however, putative genetic risk factors explain only a small proportion of phenotypic variation. Statistical analysis of family data therefore retains an important role in the search for genetic determinants of disease since, with suitable assumptions, phenotypic variation can be partitioned into shared genetic and common environmental components. Efforts to elucidate the role of genetic variants through their functional effects can then be concentrated on phenotypes where there is strong evidence that observed variation in the phenotype is consistent with the influence of genetic factors.

Generalized Linear Mixed Models (GLMMs) offer a convenient vehicle to perform the variance components analysis described above. Such an analysis relies on a correct specification of the within-family variance-covariance matrix of the phenotype, which is implied by the correlation structure of shared random effects and individual or residual error terms that appear in the linear predictor. It is not always obvious how to generate a series of individual-specific regression equations to achieve this aim, especially when families are of varying sizes and compositions.

Current approaches to specifying GLMMs for family data are tied to particular family compositions and/or phenotypic outcomes. Burton et al. [1] and Scurrah et al. [2] proposed methods for binary phenotypes and censored survival data (respectively), but their specifications were derived in an *ad-hoc* fashion that was dependent on the family structures in the data. Rabe-Hesketh et al. [3] proposed some model specifications to suit continuously-valued and categorical phenotypes, however they require specific family compositions (e.g. monozygotic (MZ) and dizygotic (DZ) twins grouped together, nuclear families with no MZ twins, etc.). Lange et al. developed FISHER [4,5], which inputs the within-family variance-covariance matrix (calculated from a pedigree file indicating relationships between individuals within families) directly into a multivariate normal likelihood function, thus only continuous phenotypes can be analyzed. Atkinson and Therneau [6] developed an R package to analyze family data that makes use of the generalized Cholesky decompostion of the matrix representing the relatedness between individuals within a family, but again only for continuously-valued phenotypes. SOLAR [7] uses a liability threshold model to analyze family data with discrete phenotypes, but these discrete phenotypes can only consist of two categories (binary phenotypes only).

In this paper we propose a general specification for GLMMs to analyze family data where families are of varying sizes and compositions. The specification utilises a decomposition of the within-family variance-covariance matrices as the basis for generating a system of regression equations that imply the desired correlation structure between phenotypes. It can be easily implemented in standard

***Corresponding author:** Kris M Jamsen, Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, Level 3, 207 Bouverie St, The University of Melbourne, Victoria 3010, Australia, Tel: +61 (0)3 8344 0700; Fax: +61 (0)3 9349 5815; E-mail: kjamsen@student.unimelb.edu.au

Received December 21, 2011; **Accepted** February 15, 2012; **Published** February 21, 2012

Citation: Jamsen KM, Zaloumis SG, Scurrah KJ, Gurrin LC (2012) Specification of Generalized Linear Mixed Models for Family Data using Markov Chain Monte Carlo Methods. J Biomet Biostat S1:003. doi:10.4172/2155-6180.S1-003

statistical packages that support mixed effects models, such as Stata [8], but is particularly suited to scenarios where the model is expressed in a "mechanistic" way via an explicit statement of a regression equation with fixed and random effects for each datapoint, such as those encapsulated in the WinBUGS software and programming language [9].

Derivation of the Model Specification

Our models for data have as their basis Fisher's polygenic model [10], where the genetic contribution to the phenotype is the combined effect of possibly a large number of separate locations or *loci* on the human genome. At each of these loci an individual inherits one allele (of possibly many) from each parent. The effect of these alleles is assumed to be additive - the presence of each additional allele of the same type changes the phenotype by the same amount (so the effect of two identical alleles is twice the effect of one allele acting alone), and the effect of two different alleles is the sum of the effect of each allele separately. The result of this assumption is a phenotypic model where covariation between individuals depends only on their level of relatedness. If the underlying additive model holds at a sufficient proportion of the active loci, then the polygenic model will be a reasonable approximation to the true but unknown effect of the genetic factors on the phenotype. Common environmental factors might also contribute to similarity of outcomes within families, with the additional assumption that the correlation between phenotypes does not depend of the degree of relatedness of the individuals.

We can formalise the above by specifying a model for individual participant data:

$$Y_{ij} = \mu_{ij} + a_{ij} + c_{ij} + \varepsilon_{ij,} \tag{1}$$

where Y_{ij} is the observed phenotype (continuously-valued) for the j^{th} member of the i^{th} family, μ_{ij} is the fixed-effect mean that can be expressed in a general linear predictor to capture the effect of measured environmental and/or genetic factors, including possible interactions between the two, $a_{ij} \sim \mathrm{N}(0,\sigma_a^2)$ is the additive genetic random effect, $c_{ij} \sim \mathrm{N}(0,\sigma_c^2)$ is a common environment random effect (typically we take $c_{ij} = c_i$ for all i) and $\varepsilon_{ij} \sim \mathrm{N}(0,\sigma_\varepsilon^2)$ is an individual-specific residual term. We assume that these three sources of variation or random effects are independent of each other, so that the total variation is the sum of the variance components, i.e. $\mathrm{Var}(Y_{ij}) = \sigma_a^2 + \sigma_c^2 + \sigma_\varepsilon^2$. This preliminary model specifies the marginal, univariate distribution of the random effects a,c and ε, and hence the distribution of the phenotype Y, but we need further structure to express the assumed correlation between phenotypes within a family and thus specify the joint distribution of phenotypes.This can be achieved through within-family sharing of the random effects a_{ij} and c_{ij}. For a family i with $j = 1,2,\ldots,n_i$ individuals, we re-write equation 1 as a multivariate model:

$$\mathbf{Y}_i \quad \mu_i + \mathbf{Z}_{\mathbf{a}_i}\mathbf{a}_i + \mathbf{Z}_{\mathbf{c}_i}\mathbf{c}_i + \mathbf{Z}_{\quad i}\varepsilon_i$$

where $\mathbf{Y}_i$ is an $n_i \times 1$ vector of observed phenotypes, μ_i is an $n_i \times 1$ vector of means, $\mathbf{a}_i$, $\mathbf{c}_i$, and ε_i are $n_i \times 1$ vectors of additive genetic, common environment and individual-specific random effects with $n_i \times n_i$ design matrices $\mathbf{Z}_{\mathbf{a}_i}$, $\mathbf{Z}_{\mathbf{c}_i}$ and $\mathbf{Z}_{\varepsilon_i}$ respectively.

Setting aside the specification of $\mathbf{Z}_{\mathbf{a}_i}$ for the moment, let $\mathbf{Z}_{\mathbf{c}_i} = \mathbf{1}_{n_i}$, the $n_i \times 1$ vector of all ones and $c_{ij} = c_i$ for all i so that individuals within families share a single, common random effect representing the shared family environment. More generally, $\mathbf{Z}_{\mathbf{c}_i}$ will be a matrix indicating which related individuals share a common environment. Also, let $\mathbf{Z}_{\varepsilon_i}$ be the $n_i \times n_i$ identity matrix, $\mathbf{I}_{n_i}$. Our revised specification with these assumptions is therefore

$$\mathbf{Y}_i = \mu_i + \mathbf{Z}_{\mathbf{a}_i}\mathbf{a}_i + \mathbf{1}_{n_i}c_i + \mathbf{I}_{n_i}\varepsilon_i, \tag{2}$$

implying

$$\begin{aligned}\mathrm{Var}(\mathbf{Y}_i) &= \mathbf{Z}_{\mathbf{a}_i}\mathrm{Var}(\mathbf{a}_i)\mathbf{Z}_{\mathbf{a}_i}' + \mathbf{1}_{n_i}\mathrm{Var}(c_i)\mathbf{1}_{n_i}' + \mathrm{Var}(\varepsilon) \\ &= \mathbf{Z}_{\mathbf{a}_i}\sigma_a^2\mathbf{I}_{n_i}\mathbf{Z}_{\mathbf{a}_i}' + \mathbf{1}_{n_i}\sigma_c^2\mathbf{I}_{n_i}\mathbf{1}_{n_i}' + \sigma_\varepsilon^2\mathbf{I}_{n_i} \\ &= \sigma_a^2\mathbf{Z}_{\mathbf{a}_i}\mathbf{Z}_{\mathbf{a}_i}' + \sigma_c^2\mathbf{J}_{n_i} + \sigma_\varepsilon^2\mathbf{I}_{n_i}\end{aligned} \tag{3}$$

where $\mathbf{J}_{n_i}$ is the $n_i \times n_i$ matrix of all ones. The entries $k_{jj'}$, of the matrix $\mathbf{K}_i = \mathbf{Z}_{\mathbf{a}_i}\mathbf{Z}_{\mathbf{a}_i}'$ correspond to the kinship coefficients [4] that represent the relatedness of individuals j and j' within the same family. We can then re-state the problem of model specification as the requirement that the additive genetic design matrix $\mathbf{Z}_{\mathbf{a}_i}$ satisfies the condition $\mathbf{Z}_{\mathbf{a}_i}\mathbf{Z}_{\mathbf{a}_i}' = \mathbf{K}_i$ for known $\mathbf{K}_i$.

Since $\mathbf{K}_i$ can be interpreted as a correlation matrix, it is symmetric positive-definite and can therefore be decomposed uniquely into the product of a lower triangular matrix (the *Cholesky triangle*) and its transpose. The product $\mathbf{Z}_{\mathbf{a}_i}\mathbf{Z}_{\mathbf{a}_i}'$ is then the *Cholesky decomposition* of $\mathbf{K}_i$. We show in the next section that estimates of the linear predictor are invariant to the choice of matrix decomposition.

To illustrate the Cholesky decomposition of the kinship matrix, we now consider two example models, each consisting of a single nuclear family. First, let $\mathbf{Y}_i = (Y_{i1}, Y_{i2})$ be a vector of continuously-valued phenotype data from a pair of full siblings. From equation 3 we have

$$\mathrm{Var}(\mathbf{Y}_i) = \mathbf{K}_{\mathrm{sib}}\sigma_a^2 + \mathbf{J}_{\mathrm{sib}}\sigma_c^2 + \mathbf{I}_{\mathrm{sib}}\sigma_\varepsilon^2,$$

where

$$\mathbf{K}_{\mathrm{sib}} = \begin{pmatrix} 1 & 1/2 \\ 1/2 & 1 \end{pmatrix}$$

so that the phenotypic correlation between full siblings due to shared genetic factors is 1/2 since they will, on average, share half of their genetic material. The Cholesky triangle of $\mathbf{K}_{\mathrm{sib}}$ is then

$$\mathbf{Z}_{\mathrm{sib}} = \begin{pmatrix} 1 & 0 \\ 1/2 & \sqrt{3}/2 \end{pmatrix}$$

so that the system of equations in 2 reduces to

$$Y_{i1} = \mu_{i1} + a_{i1} + c_i + \varepsilon_{i1}$$

$$Y_{i2} = \mu_{i2} + \frac{1}{2}a_{i1} + \frac{\sqrt{3}}{2}a_{i2} + c_i + \varepsilon_{i2}.$$

For a nuclear family of biologically unrelated parents and two children (arbitrarily ordered as father, mother, child 1 and child 2), the kinship matrix $\mathbf{K}_{\mathrm{nuc}}$ for the within-family correlation structure implied by shared genetic factors is

$$\mathbf{K}_{\mathrm{nuc}} = \begin{pmatrix} 1 & 0 & 1/2 & 1/2 \\ 0 & 1 & 1/2 & 1/2 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1/2 & 1/2 & 1/2 & 1 \end{pmatrix}$$

which implies that the corresponding random effects design matrix $\mathbf{Z}_{\text{nuc}}$ is

$$\mathbf{Z}_{\text{nuc}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1/2 & 1/2 & 1/\sqrt{2} & 0 \\ 1/2 & 1/2 & 0 & 1/\sqrt{2} \end{pmatrix}.$$

so that the system of equations in 2 becomes

$$Y_{i1} = \mu_{i1} + a_{i1} + c_i + \varepsilon_{i1}$$

$$Y_{i2} = \mu_{i2} + a_{i2} + c_i + \varepsilon_{i2}$$

$$Y_{i3} = \mu_{i3} + \frac{1}{2}a_{i1} + \frac{1}{2}a_{i2} + \frac{1}{\sqrt{2}}a_{i3} + c_i + \varepsilon_{i3}$$

$$Y_{i4} = \mu_{i4} + \frac{1}{2}a_{i1} + \frac{1}{2}a_{i2} + \frac{1}{\sqrt{2}}a_{i4} + c_i + \varepsilon_{i4},$$

which produces the correct within-family correlation structure.

For a non-continuously-valued phenotype, we specify a generalized linear model for $\mathbf{Y}$ using a link function $g()$ to relate the expected value of $\mathbf{Y}$ to the linear predictor of fixed and random effects:

$$g(E(\mathbf{Y}_i)) = \mu_i + \mathbf{Z}_{\mathbf{a}_i}\mathbf{a}_i + \mathbf{1}_{n_i}c_i. \quad (4)$$

McCulloch and Searle [11] show that a Taylor expansion of the link function around $E(\mathbf{Y}_i)$ approximately follows a linear mixed model.

Choice of matrix decomposition

Before illustrating the use of the matrix specification above to analyze data from multiple families, we show that predictions from the fitted model are invariant to the choice of the matrix decomposition. Consider two alternative specifications of the model in equation 2

$$\mathbf{Y}_i = \mu_i + \mathbf{Z}_{\mathbf{a1}_i}\mathbf{a1}_i + \mathbf{1}_{n_i}c_i + \mathbf{I}_{n_i}\varepsilon_i$$

$$\mathbf{Y}_i = \mu_i + \mathbf{Z}_{\mathbf{a2}_i}\mathbf{a2}_i + \mathbf{1}_{n_i}c_i + \mathbf{I}_{n_i}\varepsilon_i.$$

Note that $\mathbf{a1}_i$ and $\mathbf{a2}_i$ are alternative specifications of the random effect vector $\mathbf{a}_i$. The equations above imply

$$\text{Var}(\mathbf{Y}_i) = \sigma^2_{a1}\mathbf{Z}_{\mathbf{a1}_i}\mathbf{Z}'_{\mathbf{a1}_i} + \sigma^2_c\mathbf{J}_{n_i} + \sigma^2_\varepsilon\mathbf{I}_{n_i}$$

$$= \sigma^2_{a2}\mathbf{Z}_{\mathbf{a2}_i}\mathbf{Z}'_{\mathbf{a2}_i} + \sigma^2_c\mathbf{J}_{n_i} + \sigma^2_\varepsilon\mathbf{I}_{n_i}$$

where $\text{Var}(\mathbf{ak}_i) = \sigma^2_{ak}$ for $\mathbf{k} = 1,2$, which implies that $\mathbf{Z}_{\mathbf{a1}_i}\mathbf{Z}'_{\mathbf{a1}_i} = \mathbf{Z}_{\mathbf{a2}_i}\mathbf{Z}'_{\mathbf{a2}_i}$ if $\sigma^2_{a1} = \sigma^2_{a2} = \sigma^2_a$. Consider now the fitted value at the family level taking expectations at the individual level, namely $E(\mathbf{Y}_i) = \mu_i + \mathbf{Z}_{\mathbf{ak}_i}\mathbf{ak}_i + \mathbf{1}_{n_i}\mathbf{c}_i$. Denoting

$$\mathbf{V}_i = \text{Var}(\mathbf{Y}_i) = \sigma^2_a\mathbf{Z}_{\mathbf{ak}_i}\mathbf{Z}'_{\mathbf{ak}_i} + \sigma^2_c\mathbf{J}_{n_i} + \sigma^2_\varepsilon\mathbf{I}_{n_i} = \sigma^2_a\mathbf{K}_i + \sigma^2_c\mathbf{J}_{n_i} + \sigma^2_\varepsilon\mathbf{I}_{n_i}$$

and using results stated in [12] we have

$$E(\mathbf{ak}_i \mid \mathbf{Y}_i) = \text{Var}(\mathbf{ak}_i)\mathbf{Z}'_{\mathbf{ak}_i}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mu_i)$$

$$= \sigma^2_a\mathbf{Z}'_{\mathbf{ak}_i}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mu_i),$$

which implies that

$$E(\mathbf{Z}_{\mathbf{ak}_i}\mathbf{ak}_i \mid \mathbf{Y}_i) = \mathbf{Z}_{\mathbf{ak}_i}\sigma^2_a\mathbf{Z}'_{\mathbf{ak}_i}\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mu_i)$$

$$= \sigma^2_a\mathbf{K}_i\mathbf{V}_i^{-1}(\mathbf{Y}_i - \mu_i).$$

Thus $E(\mathbf{Z}_{\mathbf{ak}_i}\mathbf{ak}_i \mid \mathbf{Y}_i)$ does not depend on the choice of decomposition. Again using results stated in [12] we also have

$$\text{Var}(\mathbf{ak}_i \mid \mathbf{Y}_i) = [\mathbf{Z}'_{\mathbf{ak}_i}(\sigma^2_\varepsilon\mathbf{I}_{n_i})^{-1}\mathbf{Z}_{\mathbf{ak}_i} + (\sigma^2_a)^{-1}]^{-1},$$

and therefore

$$\text{Var}(\mathbf{Z}_{\mathbf{ak}_i}\mathbf{ak}_i \mid \mathbf{Y}_i) = \mathbf{Z}_{\mathbf{ak}_i}[\mathbf{Z}'_{\mathbf{ak}_i}(\sigma^2_\varepsilon\mathbf{I}_{n_i})^{-1}\mathbf{Z}_{\mathbf{ak}_i} + (\sigma^2_a)^{-1}]^{-1}\mathbf{Z}'_{\mathbf{ak}_i}$$

$$= \sigma^2_\varepsilon\mathbf{I}_{n_i} + \sigma^2_a\mathbf{Z}_{\mathbf{ak}_i}\mathbf{Z}'_{\mathbf{ak}_i}$$

$$= \sigma^2_\varepsilon\mathbf{I}_{n_i} + \sigma^2_a\mathbf{K}_i.$$

Therefore $\text{Var}(\mathbf{Z}_{\mathbf{ak}_i}\mathbf{ak}_i \mid \mathbf{Y}_i)$ does not depend on the choice of matrix decomposition for the design matrix of the genetic random effects contribution to the linear predictor.

Implementing the Model Specification

The model specification described above can be implemented using, for example, R and WinBUGS by the following procedure:

1. Create the *kinship* matrix for all families, which results in a $N \times N$ block diagonal sparse matrix, with N being the total number of individuals in the data. Here we assume that the data can be partitioned into families (possibly extended families or "pedigrees") and that members in the same family appear as consecutive records in the dataset. The diagonal blocks are the within-family *kinship* matrices and the off-diagonal blocks are zero (implying outcomes for individuals in different families are uncorrelated). This computation can be done in *R* using the makekinship command from the kinship package [6]. Note that makekinship does not distinguish between monozygous (MZ) and dizygous (DZ) twins, so the resulting kinship matrix needs to be amended to reflect this distinction.

2. Obtain $\mathbf{Z}$ by computing the transpose of the Cholesky decomposition of the kinship matrix, which can also be done in R. This results in a "triangular diagonal" matrix, that is, a block diagonal sparse matrix where the blocks are lower triangular matrices.

3. Set up a three level hierarchical model in WinBUGS. For a continuously-valued phenotype, level one (the individual level, i.e. the j^{th} member of family i) is specified as

$$Y_{ij} \sim N(\mu_{ij}, \sigma^2_\varepsilon),$$

where

$$\mu_{ij} = \mathbf{x}_{ij}\beta + \mathbf{z}_{aij}\mathbf{a} + c_{ij} \quad (5)$$

Thus the mean μ for the j^{th} person of family i depends on (i) a vector of fixed-effects regression coefficients β and the corresponding fixed-effects design matrix $\mathbf{X}$ (has columns x_{ij}); (ii) the $N \times 1$ vector of genetic random effects $\mathbf{a}$ (with $a_{ij} \sim \text{N}(0, \sigma^2_a)$) and the corresponding $N \times N$ block diagonal sparse matrix of family decompositions $\mathbf{Z}_\mathbf{a}$ (which has columns z_{aij}) and (iii) the common environment random effect c_{ij}, where $c_{ij} = c_i$ for all i and $c_i \sim \text{N}(0, \sigma^2_c)$. For a general categorical phenotype, the model is specified as the linear model in equation 5 for

$g(E(Y_{ij}))$, where g is the link function. In a full probability (i.e. Bayesian) framework, prior distributions would be specified for β, σ_a^2, σ_c^2 and σ_ε^2.

It is straightforward to implement this process in standard statistical software, for example, by creating an R source file that prepares the data (steps 1 & 2), performs the analysis in WinBUGS via R2WinBUGS [13] (step 3) and outputs the results as dataframes and tables.

Evaluation of the Model Specification

Analysis of simulated data

To assess empirically the validity of the proposed model specification, we conducted a small simulation study. Data were simulated (in R) for 790 independent nuclear families (two parents and one or more children, possibly including MZ or DZ twins), where the family sizes and compositions were chosen to mimic those from the Victorian Family Heart Study (VFHS) [14].

Three phenotypes were simulated: (i) a continuously-valued phenotype from the linear mixed model in equation 2 with $\mu_i = 0$, $\sigma_a = 4$, $\sigma_c = 3$ and $\sigma_\varepsilon = 2$; (ii) a binary phenotype from the generalized linear mixed model specified in equation 4 with $g(E(\mathbf{Y}_i)) = \log(E(\mathbf{Y}_i)/(1-E(\mathbf{Y}_i)))$ (i.e the *logit* transformation of a proportion) and $\mu_i = -3.70\mathbf{1}_{n_i}$, corresponding approximately to an overall proportion of $E(\mathbf{Y}_i) = 0.25$; and (iii) a binary phenotype generated in the same way as (ii) above but with $g(E(\mathbf{Y}_i)) = \Phi^{-1}(E(\mathbf{Y}_i))$ where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution (i.e. the *probit* transformation of a proportion).

The simulated data were analyzed by fitting the models from which they were generated, producing three analyses per simulated dataset. The proposed model specification was implemented in WinBUGS (via R2WinBUGS, [13]) in R. For all models in WinBUGS, a single chain was specified and uniform prior distributions (U(0,30)) were specified for σ_a^2, σ_c^2 and σ_ε^2. The linear model was run for 5,000 iterations (2,500 burn-in), the logistic model was run for 10,000 iterations (5,000 burn-in) and the probit model was run for 20,000 iterations (10,000 burn-in). For the binary phenotypes the cut function was used to prevent any "orphan" random effects (those that appear in the linear predictor of a single individual and are not shared by any other family members) from contributing to the posterior distribution of the corresponding variance component. This was done since these random effects mimic those associated with a residual error term (which no longer exists in a GLMM specification) and can lead to upwardly biased estimates [1]. For all analyses, the posterior medians and 95% posterior intervals for the variance components were computed. This simulation-estimation process was repeated 10 times for all phenotypes.

The results from the simulation-estimation procedure are displayed in Figure 1. For the linear model all displayed posterior summaries were close to the target values (the median of the posterior medians for σ_a, σ_c and σ_ε, were 4.14 (target value 4), 3.04 (3) and 1.90 (2), respectively). The estimates from the logistic model were consistent with the nominal values, with only one posterior interval for σ_a and one posterior interval for σ_c excluding the target values (the median of the posterior medians for σ_a and σ_c were 4.05 (4) and 2.82 (3), respectively). The estimates from the probit model were somewhat less consistent with the nominal values, with four posterior intervals for σ_a and three posterior intervals for σ_c excluding the target values (although the median of the posterior medians for σ_a and σ_c were 3.79 and 2.90, respectively, which were close to the respective target values of 4 and 3). On average it took 58 seconds per analysis for the linear models, 5.15 minutes for the logistic models and 6.36 minutes for the probit models.

Analysis of data from the victorian family heart study

To illustrate the proposed model specification in an application, data from the Victorian Family Heart Study (VFHS) were analyzed in WinBUGS. The Victorian Family Heart Study was established to investigate the causes of familial patterns in cardiovascular risk factors. The study consisted of adult families recruited in Melbourne, Australia, where each family consisted of both parents and at least one natural adult offspring. For full details of the study see [14].

Two phenotypes were analyzed: systolic blood pressure (SBP) taken lying down (continuously-valued) and high blood pressure (HBP), defined as a systolic blood pressure reading of >140 mm Hg or a diastolic blood pressure reading of >90 mm Hg. SBP was analyzed using a linear model and HBP was analyzed using a logistic model. In both models, age (dichotomised as <35 years or ≥ 35 years) and sex were included as fixed-effect covariates, and the additive genetic and common environment variance components were estimated (the linear model also included the residual component of variance). Descriptive statistics of the phenotypes and covariates included in the analyses are given in Table 1. In addition, three chains were specified for each model, and convergence was declared when the Gelman-Rubin R-hat statistic was ≤ 1.1. The analyses were also run in Stata to provide a comparison with a maximum likelihood method. Stata was chosen for the maximum likelihood analysis since it comes with pre-packaged routines allowing easy implementation of the proposed model specification for both continuously-valued and categorical phenotypes (xtmixed was used for SBP and xtmelogit was used for HBP, where 6 integration points were specified for xtmelogit). The results from these analyses are displayed in Table 2.

The results were similar between WinBUGS and Stata, however Stata took substantially longer to achieve estimates for the logistic model.

Discussion

In this work we developed a general method for specifying GLMMs for data from families of varying size and composition. We evaluated the method by analyzing simulated and real data and found that it performed well for parameter estimation and run times for both continuously-valued and binary phenotypes. The model specification is particularly suited to MCMC methods that require a "mechanistic" description of a linear predictor containing both fixed and random effects.

For the simulated continuously-valued phenotypes the proposed model specification delivered unbiased parameter estimates. Using the specification in WinBUGS for the continuously-valued phenotype from the Victorian Family Heart Study (systolic blood pressure) gave similar results to xtmixed in Stata. The analysis did, however, take much longer in WinBUGS - iterative estimation routines for linear mixed models typically converge to the maximum likelihood estimates quickly, so a simulation-based approach to parameter estimation will under-perform in comparison. The MCMC framework does, however, allow for flexibility in model specification, such as the opportunity for

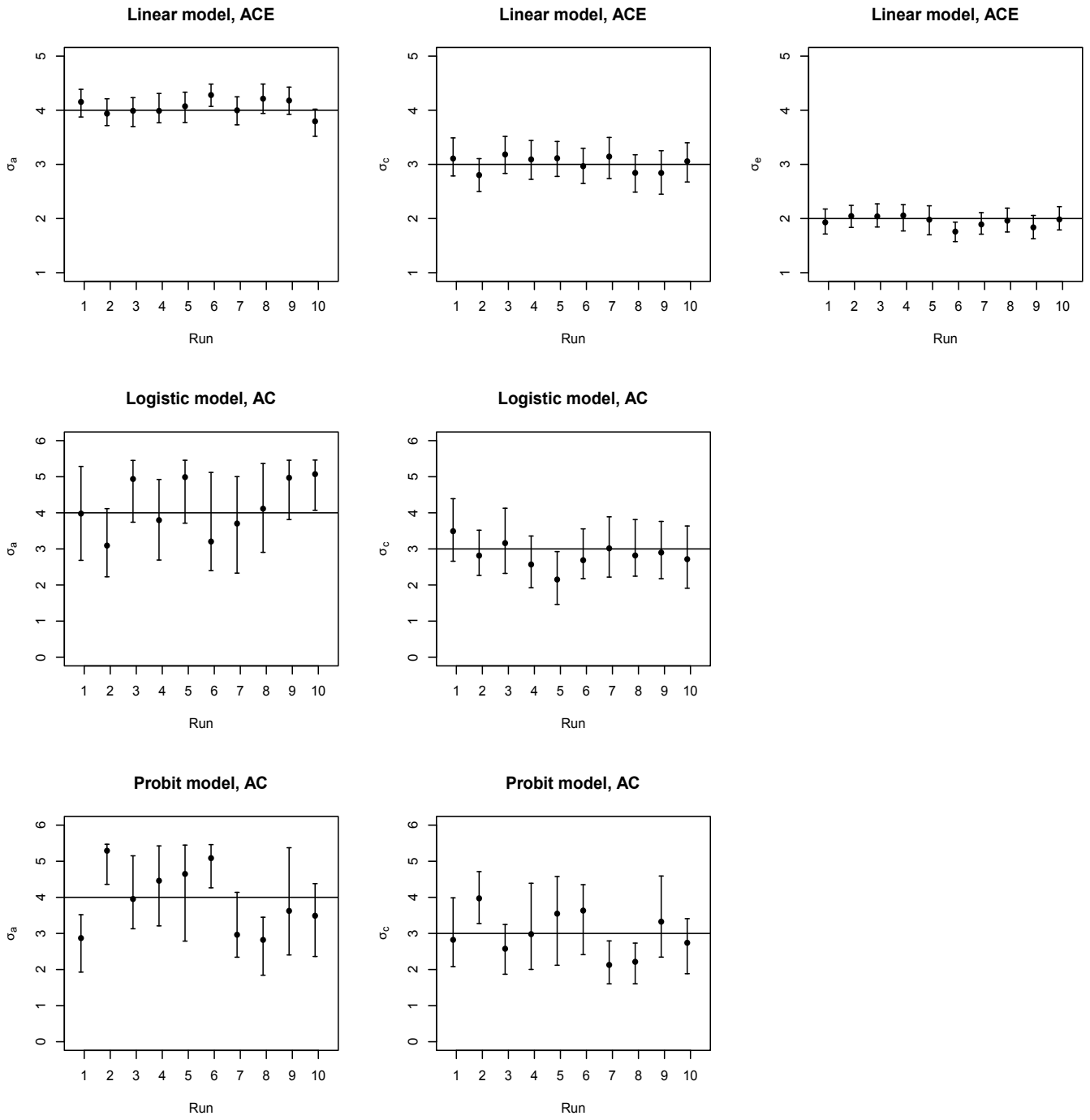

Figure 1: Posterior medians and 95% posterior intervals from the simulation-estimation procedure for the linear, logistic and probit models.

the shared environment component of variance to be dependent on the value of other covariates, something that is difficult to achieve in linear modelling routines such as xtmixed in Stata.

For the simulated binary phenotypes the proposed model specification also produced parameter estimates close to the target values, but the coverage of the target parameters by the nominal 95% posterior intervals was not as good as it was for the continuously-valued phenotype. The analysis of the binary phenotype from the Victorian Family Heart Study (high blood pressure) that employed our model specification using WinBUGS yielded similar estimates to xtmelogit in Stata, however the analysis took substantially less time in WinBUGS. The xtmelogit routine uses adaptive quadrature, which can be slow, especially when there are many random effects, a large number of observations and several integration points [15]. Even with the use of multiple processors, it is unlikely that the processing time in Stata could be reduced from days to hours.

It is straightforward to extend the proposed model specification to accommodate multi-category phenotypes. Details on the specification

		SBP* lying down (mm Hg)		HBP †
	No.	Mean (SD ‡)	Range	Percent
Offspring				
Female	767	114 (9.86)	86-149	1.22
Male	698	122 (11.2)	96-167	6.68
Parents				
Female	790	126 (16.4)	87-198	20.5
Male	790	132 (16.6)	96-215	27.5

*SBP = systolic blood pressure
† HBP = high blood pressure
‡ SD = standard deviation

Table 1: Descriptive statistics of phenotypes by generation and sex in the Victorian Family Heart Study.

		σ_a	σ_c	σ_e	
Model	Package	(95% interval)*	(95% interval)*	(95% interval)*	Run time
Linear	WinBUGS	6.1	4.7	11.8	2.4 minutes †
		(4.3, 7.6)	(3.6, 5.7)	(11.2, 12.2)	
Linear	Stata	5.8	4.8	11.9	8 seconds
		(4.1, 8.0)	(3.8, 6.0)	(11.2, 12.6)	
Logistic	WinBUGS	2.1	1.2	n/a	24.3 minutes ‡
		(1.2, 3.0)	(0.7, 1.7)	n/a	
Logistic	Stata	1.7	1.1	n/a	25 days
		(0.9, 3.2)	(0.8, 1.6)	n/a	

*95% posterior intervals for WinBUGS; 95% confidence intervals for Stata
† Results based on 3,000 iterations (1,500 burn-in)
‡ Results based on 10,000 iterations (5,000 burn-in)

Table 2: Results from analyzing continuously-valued and binary phenotypes from the Victorian Family Heart Study.

of logistic regression models that incorporate random effects that can reproduce within-family correlation structures for ordinal outcomes due to shared environmental and genetic risk factors are described in Zaloumis [16]. In addition, Ellis et al. [17] analyzed data from the VFHS to investigate risk factors for male pattern baldness (a four-category ordinal outcome variable) using generalized linear mixed models similar to those proposed here.

We successfully reduced the processing time in WinBUGS by stating explicitly the multiplication of the columns of the design matrix $\mathbf{Z}_a$ by the a_{ij}'s. This contrasts with what appears to be a more convenient coding strategy relying on the inprod (inner product) command. Results from a small subset of simulated data (continuously-valued phenotype, 300 families) showed this choice of coding took at least three times as long to achieve convergence of estimates. In addition, it is possible to further reduce the processing time in WinBUGS (and Stata) by "stacking" the diagonal blocks of $\mathbf{Z}_a$ on top of one another, so that the number of columns of this "stacked" matrix is equal to the number of people in the largest observed family. For families that don't have as many members as the largest family, the remaining columns can be filled in with zeros. This "stacking" approach is appropriate if families are assumed to be unrelated (as they were in the simulation-estimation procedure and the analysis of the VFHS data) and the software allows one to specify a hierarchical model on families where the $\mathbf{a}_i$ random effects vectors are independent among families.

Although we have shown in this paper that our approach to specifying GLMMs for family data is analytically sound, it does not overcome the computational difficulties of estimating variance components with categorical phenotypes [1]. It is possible to circumvent these problems with careful specification of the model, but a large number of iterations of the Gibbs sampler will be required (i.e. tens if not hundreds of thousands) and accurate estimation will require a substantial number of families (i.e. hundreds if not thousands). Therefore when using our proposed approach to specifying mixed models for family data with categorical phenotypes, a large number of families will be needed for adequate estimation of the variance components, particularly σ_a^2, and the Gibbs sampler will have to be run for several thousands of iterations.

Our proposed general approach to specifying GLMMs for family data avoids the nuisance of having to specify a model that is dataset-specific, which can be tedious and time-consuming. This is an improvement on current approaches, which are tied to particular phenotypes and/or family compositions. The method can be easily implemented in freely available software and was particularly suited to the MCMC framework. The method of specification proposed in this paper should be considered when analyzing family data, particularly when outcomes are categorical, prior information on parameters needs to be incorporated and/or non-standard specifications of the genetic model are of interest.

Acknowledgments

The authors thank Prof Murray Aitkin for his insightful suggestions regarding the implementation of MCMC when fitting GLMMs.

References

1. Burton PR, Tiller KJ, Gurrin LC, Cookson WO, Musk AW, et al. (1999) Genetic Variance Components Analysis for Binary Phenotypes Using Generalized Linear Mixed Models (GLMMs) and Gibbs Sampling. Genet Epidemiol 17: 118-140.
2. Scurrah KJ, Palmer LJ, Burton PR (2000) Variance Components Analysis for Pedigree-Based Censored Survival Data Using Generalized Linear Mixed Models (GLMMs) and Gibbs Sampling in BUGS. Genet Epidemiol 19: 127-148.
3. Rabe-Hesketh S, Skrondal A, Gjessing HK (2008) Biometrical Modeling of Twin and Family Data Using Standard Mixed Model Software. Biometrics 64: 280-288.
4. Lange K, Westlake J, Spence MA (1976) Extensions to pedigree analysis. III. Variance components by the scoring method. Ann Hum Genet 39: 485-491.
5. Lange K, Weeks D, Boehnke M (1988) Programs for pedigree analysis: MENDEL, FISHER, and dGENE. Genet Epidemiol 5: 471-472.
6. Atkinson B, Therneau T (2009) kinship: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. R package version 1.1.3.
7. Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 62: 1198-1211.
8. Statacorp (2009) Stata Statistical Software: Release 11. College Station, TX, USA.
9. Lunn DJ, Thomas A, Best N, Spiegelhalter D (2000) WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility. Stat Comput 10: 325-337.
10. Fisher R (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edinb 52: 399-433.
11. McCulloch C, Searle S (2001) Generalized, Linear and Mixed Models. Wiley Series in Probability and Statistics. New York, New York: John Wiley and Sons, Inc.
12. Robinson GK (1991) That BLUP is a good thing: The estimation of random effects. Statist Sci 6: 15-32.
13. Sturtz S, Ligges U, Gelman A (2005) R2WinBUGS: A Package for Running WinBUGS from R. J Stat Softw 12: 1-16.

14. Harrap S, Stebbing M, Hopper J, Hoang H, Giles G (2000) Familial Patterns of Covariation for Cardiovascular Risk Factors in Adults: The Victorian Family Heart Study. Am J Epidemiol 152: 704-715.

15. Rabe-Hesketh S, Skrondal A (2008) Multilevel and Longitudinal Modeling Using Stata. Stata Press, 2nd ed., College Station, TX, USA.

16. Zaloumis SG (2011) A statistical model for ordinal categorical family data [PhD Dissertation]. Medicine, Dentistry and Health Sciences Physiology, The University of Melbourne.

17. Ellis JA, Scurrah KJ, Cobb JE, Zaloumis SG, Duncan AE, et al. (2007) Baldness and the androgen receptor: the AR polyglycine repeat polymorphism does not confer susceptibility to androgenetic alopecia. Hum Genet 121: 451-457.

This article was originally published in a special issue, **Advances in Markov Chain Monte Carlo Methods and Survival Analysis** handled by Editor(s). Dr. Faming Liang, Texas A&M University, USA; Dr. Nengjun Yi, University of Alabama at Birmingham, USA; Dr. Wenqing He, University of Western Ontario, Canada; Dr. Liuquan Sun, Institute of Applied Mathematics, Academy of Mathematics and Systems Science, China