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

Kris M Jamsen1, Sophie G Zaloumis1,2, Katrina J Scurrah1,3 and Lyle C Gurrin1

1Centre for Molecular, Environmental, Genetic and Analytic Epidemiology, Melbourne School of Population Health, The University of Melbourne, Victoria 3010, Australia
2Department 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 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 decomposition 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...
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:

\[
Y_i = \mu_i + a_i + c_i + \epsilon_i,
\]

(1)

where \(Y_i\) is the observed phenotype (continuously-valued) for the
\(j^{th}\) member of the \(i^{th}\) family, \(\mu_i\) 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, \(a_i \sim N(0, \sigma_a^2)\) is the additive genetic random effect,
\(c_i \sim N(0, \sigma_c^2)\) is a common environment random effect (typically we
take \(c_i = \epsilon_i\) for all \(i\)) and \(\epsilon_i \sim N(0, \sigma_e^2)\) 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. \(\text{Var}(Y_i) = \sigma_a^2 + \sigma_c^2 + \sigma_e^2\). This preliminary
model specifies the marginal, univariate distribution of the random
effects, \(a, c, \epsilon\), 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 \(a, c, \epsilon\). For a family \(i\) with \(j = 1, 2, \ldots, n_i\) individuals, we
re-write equation 1 as a multivariate model:

\[
Y_i \sim \mu_i + Z_{a_i}a_i + Z_{c_i}c_i + Z_{\epsilon_i}\epsilon_i,
\]

where \(Y_i\) is an \(n_i \times 1\) vector of observed phenotypes, \(\mu_i\) is an \(n_i \times 1\) vector
of means, \(a, c, \epsilon\) are \(n_i \times 1\) vectors of additive genetic, common
environment and individual-specific random effects with \(n_i \times n_i\) design
matrices \(Z_{a_i}, Z_{c_i}\) and \(Z_{\epsilon_i}\) respectively.

Setting aside the specification of \(Z_{\epsilon_i}\) for the moment, let \(Z_{a_i} = I_{n_i}\),
the \(n_i \times 1\) vector of all ones and \(c_i = \epsilon_i\) for all \(i\) so that individuals within
families share a single, common random effect representing the shared
family environment. More generally, \(Z_{a_i}\) will be a matrix indicating
which related individuals share a common environment. Also, let \(Z_{c_i}\)
be the \(n_i \times n_i\) identity matrix, \(I_{n_i}\). Our revised specification with these
assumptions is therefore

\[
Y_i = \mu_i + Z_{a_i}a_i + I_{n_i}c_i + I_{n_i}\epsilon_i,
\]

(2)

implying

\[
\text{Var}(Y_i) = Z_{a_i}\text{Var}(a_i)Z_{a_i} + I_{n_i}\text{Var}(c_i)I_{n_i} + \text{Var}(\epsilon)
\]

\[
= Z_{a_i}\sigma_a^2I_{n_i} + I_{n_i}\sigma_c^2I_{n_i} + \sigma_e^2I_{n_i}
\]

(3)

where \(J_{n_i}\) is the \(n_i \times n_i\) matrix of all ones. The entries \(k_{ij}\), of the matrix
\(K_{n_i} = Z_{a_i}Z_{a_i}'\) correspond to the kinship coefficients [4] that represent
the relatedness of individuals \(i\) 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 \(Z_{a_i}\) satisfies the condition
\(Z_{a_i}Z_{a_i}' = K_{n_i}\) for known \(K_{n_i}\).

Since \(K_{n_i}\) 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 \(Z_{a_i}Z_{a_i}'\) is then the Cholesky decomposition of
\(K_{n_i}\). 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 \(Y_i = (Y_{a_i}, Y_{c_i})\) be a vector of continuously-valued
phenotype data from a pair of full siblings. From equation 3 we have

\[
\text{Var}(Y_i) = K_{n_i}\sigma_a^2 + J_{n_i}\sigma_c^2 + I_{n_i}\sigma_e^2,
\]

where

\[
K_{n_i} = \begin{pmatrix}
1 & 1/2 \\
1/2 & 1
\end{pmatrix}
\]

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 \(K_{n_i}\) is then

\[
J_{n_i} = \begin{pmatrix}
1 & 0 \\
1/2 & \sqrt{1/2}
\end{pmatrix}
\]

so that the system of equations in 2 reduces to

\[
\begin{align*}
Y_{a_i} &= \mu_{a_i} + a_i + c_i + \epsilon_i \\
Y_{c_i} &= \mu_{c_i} + 1/2a_i + \sqrt{1/2}c_i + \epsilon_i.
\end{align*}
\]

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

\[
K_{n_s} = \begin{pmatrix}
1 & 0 & 1/2 & 1/2 \\
0 & 1 & 1/2 & 1/2 \\
1/2 & 1/2 & 1 & 1/2 \\
1/2 & 1/2 & 1/2 & 1
\end{pmatrix}
\]
which implies that the corresponding random effects design matrix $Z_{true}$ is

$$Z_{true} = \begin{pmatrix}
1 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 \\
1/2 & 1/2 & 1/\sqrt{2} & 0 \\
1/2 & 1/2 & 0 & 1/\sqrt{2}
\end{pmatrix},$$

so that the system of equations in 2 becomes

$$Y_i = \mu_i + a_i + c_i + \epsilon_i,$$

$$Y_i = \mu_i + a_i + c_i + \epsilon_i,$$

$$Y_i = \mu_i + a_i + c_i + \epsilon_i,$$

$$Y_i = \mu_i + a_i + c_i + \epsilon_i,$$

which produces the correct within-family correlation structure.

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

$$g(E(Y_i)) = \mu_i + Z_{true} a_i + 1_i c_i.$$  

McCulloch and Searle [11] show that a Taylor expansion of the link function around $E(Y_i)$ 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 model in equation 2 are alternative specifications of the random effects contribution to the linear predictor.

Thus the mean $E(Y_i)$ does not depend on the choice of decomposition. Again using results stated in [12] we also have

$$\text{Var}(a_i | Y_i) = [Z_{sh_i}(\sigma^2_{1_s})^{-1}Z_{sh_i} + (\sigma^2_{e})^{-1}]^{-1},$$

and therefore

$$\text{Var}(a_i | Y_i) = [Z_{sh_i}(\sigma^2_{1_s})^{-1}Z_{sh_i} + (\sigma^2_{e})^{-1}]^{-1}Z_{sh_i}$$

$$= \sigma^2_{1_s} + \sigma^2_{e}Z_{sh_i},$$

Therefore $\text{Var}(a_i | Y_i)$ does not depend on the choice of matrix decomposition for the design matrix of the genetic random effects.

### 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 monoygous (MZ) and dizygous (DZ) twins, so the resulting kinship matrix needs to be amended to reflect this distinction.

2. Obtain $Z$ 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

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

where

$$\mu_{ij} = x_{ij} \beta + Z_{ih} \alpha + c_{ij}.$$  

(5)

Thus the mean $\mu$ for the $j$th person of family $i$ depends on (i) a vector of fixed-effects regression coefficients $\beta$ and the corresponding fixed-effects design matrix $X$ (has columns $x_{ij}$); (ii) the $N \times 1$ vector of genetic random effects $a$ (with $a_{ij} \sim N(0, \sigma^2_{a})$) and the corresponding $N \times N$ block diagonal sparse matrix of family decompositions $Z_a$ (which has columns $z_{ij}$); and (iii) the common environment random effect $c_{ij}$ (where $c_{ij} \sim \epsilon_{ij}$ for all $i$ and $c_{ij} \sim N(0, \sigma^2_{c})$). For a general categorical phenotype, the model is specified as the linear model in equation 5 for
g(E(Y)), where g is the link function. In a full probability (i.e. Bayesian) framework, prior distributions would be specified for β, σa², σc² and σe².

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 μ = 0, σe = 4, σa = 3 and σc = 2; (ii) a binary phenotype from the generalized linear mixed model specified in equation 4 with g(E(Y)) = log(E(Y))/1-E(Y)) (i.e. the logit transformation of a proportion) and μ = -3.701, corresponding approximately to an overall proportion of E(Y) = 0.25; and (iii) a binary phenotype generated in the same way as (ii) above but with g(E(Y)) = Φ⁻¹(E(Y)) 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 σa², σc² and σe². 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 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
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
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 $Z_i$ by the $a_j$'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 $Z_i$ 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 $a_i$ 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 $\sigma_e^2$, 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**


