Paper 253-28

An Introduction to the Analysis of Mixed Models

Dallas E. Johnson, Kansas State University, Manhattan, KS

ABSTRACT

This paper introduces the General Linear Mixed Model (GLMM) and compares various alternatives for estimating estimable functions of the model parameters and provides some of the advantages and disadvantages of each of the alternative methods of estimation. The paper also considers estimating the standard errors of the various estimates of estimable functions and how the estimates and their estimated standard errors can be used for statistical inference. Special cases of a GLMM include all types of split-plot experiments, repeated measures experiments, and various combinations of these. Examples will be given to illustrate how these kinds of experiments fit into a GLMM framework.

INTRODUCTION

The General Linear Mixed Model (GLMM) is defined by

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$$

where **y** is an nx1 observable data vector, β is a px1 vector of unknown parameters, **u** is a qx1 vector of unobservable random variables, **X** and **Z** are design matrices corresponding to the fixed and random effects, respectively, and ϵ is a vector of random errors. The basic assumptions are that

$$\epsilon \sim N(\mathbf{0}, \underline{\mathbf{R}}_{nxn}),$$
$$\mathbf{u} \sim N(\mathbf{0}, \underline{\mathbf{G}}_{qxq}),$$

and

$$COV(\epsilon, \mathbf{u}) = \mathbf{0}_{nxq}.$$

Note that

$$E(\mathbf{y}) = \mathbf{X}\beta$$

and

$$Cov(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R} = \mathbf{V} \text{ (say)}.$$

Thus

$$\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{V}).$$

ESTIMATORS OF $\ell'\beta$

Four different estimators of $\ell'\beta$ are introduced where $\ell'\beta$, is an estimable function of the parameters in β. That is, ℓ belongs to the column space of $\mathbf{X}'$.

Generalized Least Squares Estimators

If **G** and **R** are both known, then **V** is also known, and in this special case, the uniformly minimum variance unbiased (UMVU) estimate of any estimable function $\ell'\beta$ is $\ell'\hat{\beta}_V$ where

$$\hat{\beta}_V = (X'V^{-1}X)^- X'V^{-1}y \quad .$$

Furthermore, under the basic assumptions given above,

$$\ell'\hat{\beta} \sim N(\ell'\beta, \ell'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^-\ell).$$

This estimator of $\ell'\beta$ is called the *Generalized Least Squares Estimator*.

Estimated Generalized Least Squares Estimators

In practice, **G** and **R** are rarely known! However, suppose **G** and **R** can be estimated in some way by $\hat{\mathbf{G}}$ and $\hat{\mathbf{R}}$, respectively. Then $\ell'\beta$ could be estimated by $\ell'\hat{\beta}_{\hat{V}}$ where

$$\hat{\beta}_{\hat{V}} = (X'\hat{V}^{-1}X)^- X'\hat{V}^{-1}y \quad \text{and} \quad \hat{V} = \hat{Z}\hat{G}Z' + \hat{R} \quad .$$

The estimator $\ell'\hat{\beta}_{\hat{V}}$ is called the *Estimated Generalized Least Squares Estimator* (EGLS) of $\ell'\beta$.

In most balanced experiments $\ell'\hat{\beta}_{\hat{V}}$ is an unbiased estimator of $\ell'\beta$, but in unbalanced experiments $\ell'\hat{\beta}_{\hat{V}}$ cannot be shown to be an unbiased estimator of $\ell'\beta$.

Ordinary Least Squares Estimators

Another estimator of $\ell'\beta$., called the *Ordinary Least Squares Estimator* of $\ell'\beta$. is $\ell'\hat{\beta}_O$ where $\hat{\beta}_O = (X'X)^- X'y$. This estimator is always an unbiased estimator of $\ell'\beta$. In many balanced experiments

$$\ell'\hat{\beta}_{\hat{V}} = \ell'\hat{\beta}_O \quad .$$

When this is true, then $\ell'\hat{\beta}_{\hat{V}}$ is an unbiased estimator of $\ell'\beta$. It can also be shown that,

$$\ell'\hat{\beta}_O \sim N(\ell'\beta, \ell'(\mathbf{X}'\mathbf{X})^-(\mathbf{X}'\mathbf{V}\mathbf{X})(\mathbf{X}'\mathbf{X})^-\ell).$$

Even though $\ell'\hat{\beta}_{\hat{V}}$ is not unbiased for $\ell'\beta$ in unbalanced cases, one might expect that $\ell'\hat{\beta}_{\hat{V}}$ should be a reasonable estimator of $\ell'\beta$.

In most unbalanced cases one cannot determine the variance of $\ell'\hat{\beta}_{\hat{V}}$. However, note that

$$\text{Var}(\ \ell'\hat{\beta}_V\ ) = \ell'(X'V^{-1}X)^{-}\ell,$$

and so it would seem reasonable to approximate the variance of

$\ell'\hat{\beta}_{\hat{V}}$ with $\ell'(X'\hat{V}^{-1}X)^{-}\ell$ so that so that approximate

estimated standard error of $\ell'\hat{\beta}_{\hat{V}}$ is

$$s.e.(\ell'\hat{\beta}_{\hat{V}}) = \sqrt{\ell'(X'\hat{V}^{-1}X)^{-}\ell}\ .$$

GLM Estimators

Another possible estimator of $\ell'\beta$ is

$$\ell'\hat{\beta}_{GLM} + a'\hat{u}$$

where

$$\begin{bmatrix}\hat{\beta}_{GLM}\\ \hat{u}\end{bmatrix} = \begin{bmatrix}X'X & X'Z\\ Z'X & Z'Z\end{bmatrix}^{-}\begin{bmatrix}X'y\\ Z'y\end{bmatrix}$$

and where a is chosen so $\begin{bmatrix}\ell\\ a\end{bmatrix}$ belongs to the column

space of $\begin{bmatrix}X'X & X'Z\\ Z'X & Z'Z\end{bmatrix}$. This estimator is called the GLM

estimator as it is the estimator that one gets from SAS-GLM when the random effects are treated as fixed effects. The GLM estimator can be shown to be an unbiased estimator of $\ell'\beta$. One would like to choose a so that VAR($\ell'\hat{\beta}_{GLM} + a'\hat{u}$) is

minimized. The variance of $\ell'\hat{\beta}_{GLM} + a'\hat{u}$ is equal to

$$\begin{bmatrix}\ell\\ a\end{bmatrix}'[XZ]^{-}R\begin{bmatrix}X'\\ Z'\end{bmatrix}^{-}\begin{bmatrix}\ell\\ a\end{bmatrix} + \begin{bmatrix}\ell\\ a\end{bmatrix}'[XZ]^{-}ZGZ'\begin{bmatrix}X'\\ Z'\end{bmatrix}^{-}\begin{bmatrix}\ell\\ a\end{bmatrix}.$$

Choosing an appropriate **a** is easy in balanced cases, and in these cases SAS-GLM does the work for you. Choosing an appropriate **a** in unbalanced cases is much harder, and SAS-GLM rarely chooses an appropriate **a** for you. Usually, SAS - GLM gives you a message indicating that the LSMEANS are not estimable or if you are using ESTIMATE and/or CONTRAST options, you often get a message that the contrast is not estimable. In all cases, these happen because the criteria that SAS-GLM uses to select **a** does not provide an **a** such that

$\begin{bmatrix}\ell\\ a\end{bmatrix}$ belongs to the column space of $\begin{bmatrix}X'X & X'Z\\ Z'X & Z'Z\end{bmatrix}$. The

GLM estimator, $\ell'\hat{\beta}_{GLM} + a'\hat{u}$ of $\ell'\beta$ should also be a

reasonable estimator of $\ell'\beta$. Like the OLS estimator of $\ell'\beta$, it is unbiased.

To estimate the variance of the GLM estimator, one would need to estimate **R** and **G**, then Var($\ell'\hat{\beta}_{GLM} + a'\hat{u}$) is estimated

by $\begin{bmatrix}\ell\\ a\end{bmatrix}'[XZ]^{-}\hat{R}\begin{bmatrix}X'\\ Z'\end{bmatrix}^{-}\begin{bmatrix}\ell\\ a\end{bmatrix} + \begin{bmatrix}\ell\\ a\end{bmatrix}'[XZ]^{-}Z\hat{G}Z'\begin{bmatrix}X'\\ Z'\end{bmatrix}^{-}\begin{bmatrix}\ell\\ a\end{bmatrix}$.

In all cases, with any of these estimators, in order to make inferences about $\ell'\beta$, it is usually assumed that

$$\frac{\ell'\hat{\beta} - \ell'\beta}{s.e.(\ell'\hat{\beta})} \sim t(v)$$

for some appropriate value of v which usually has to be estimated in some way.

In order to compute the EGLS estimate of $\ell'\beta$, one must first be able to estimate **R** and **G**. While general estimates of **R** and **G** do not exist, there are many special cases in which one can estimate **R** and **G**.

Split-Plot Type Cases

In many cases the GLMM,

$$y_{nx1} = X_{nxp}\beta_{px1} + Z_{nxq}U_{qx1} + \epsilon_{nx1}$$

can be written as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \ldots + \mathbf{Z}_r\mathbf{u}_r + \epsilon \qquad (1)$$

where it can be assumed that $\mathbf{u}_i \sim$ independent $N(\mathbf{0}, \sigma_i^2\mathbf{I})$ for i=1,2,...r, $\epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2\mathbf{I})$, and that $COV(\mathbf{u}_i, \epsilon) = \mathbf{0}$ for all i. In this case, $\mathbf{V} = \sigma_1^2\mathbf{Z}_1\mathbf{Z}_1' + \sigma_2^2\mathbf{Z}_2\mathbf{Z}_2' + \ldots + \sigma_r^2\mathbf{Z}_r\mathbf{Z}_r' + \sigma_\epsilon^2\mathbf{I}$. Note that each $\mathbf{u}_i$ is a $q_i \times 1$ random vector with corresponding design matrix $\mathbf{Z}_i$ which is an $n \times q_i$ matrix. Here estimating **V** is accomplished by estimating σ_1^2, σ_2^2, ..., σ_r^2, and σ_ϵ^2, called variance components.

All split-plot type experiments fit into this framework. Repeated measures experiments also fit into this general framework when the repeated measures satisfy compound symmetry assumptions, and all repeated measures experiments can be written in the framework of model (1) simply by placing fewer restrictions on **R**=COV(ϵ).

Example 1: A Simple Split-Plot Model Suppose

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \delta_{k(i)} + \epsilon_{ijk}$$

for i=1,2,...,t; j=1,2,...,b; and k=1,2,...,n_i; where $\delta_{k(i)} \sim$ i.i.d. $N(0, \sigma_\delta^2)$, $\epsilon_{ijk} \sim$ i.i.d. $N(0, \sigma_\epsilon^2)$, and where all the $\delta_{k(i)}$'s and ϵ_{ijk}'s are independent. Then $\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}_1\mathbf{u}_1 + \epsilon$ where
$\mathbf{y}' = [y_{111}\ y_{112}\ldots y_{11n}\ \ y_{121}\ y_{122}\ldots y_{12n}\ \ \ldots\ \ y_{tb1}\ y_{tb2}\ \ldots\ y_{tbn}]$,
$\beta' = [\mu\ \ \tau_1\ \tau_2 \ldots \tau_t\ \ \beta_1\ \beta_2 \ldots \beta_b\ \ \gamma_{11}\ \gamma_{12} \ldots \gamma_{tb}]$,
$\mathbf{u}_1' = [\delta_{1(1)}\ \ \delta_{2(1)} \ldots \delta_{n(1)}\ \ldots\ \delta_{1(t)}\ \ \delta_{2(t)}\ \ldots\ \delta_{n(t)}]$,
$\epsilon' = [\epsilon_{111}\ \epsilon_{112}\ldots\epsilon_{11n}\ \ \epsilon_{121}\ \epsilon_{122}\ldots\epsilon_{12n}\ \ \ldots\ \ \epsilon_{tb1}\ \epsilon_{tb2}\ \ldots\ \epsilon_{tbn}]$, and **X** and $\mathbf{Z}_1$ are design matrices determined by the pattern of observed data. The ideal conditions for the errors in the split-plot model can be restated as $\mathbf{u}_1 \sim N(\mathbf{0}, \sigma_\delta^2\mathbf{I}_N)$ where $N=n_1+n_2+\ldots+n_t$,, $\epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_{Nb})$, and that $\mathbf{u}_1$ and ϵ are independent. For the simple split plot model, one must estimate σ_δ^2 and σ_ϵ^2 in order to estimate **V**.

Example 2: A Simple Repeated Measures Model. Suppose

$$y_{ijk} = \mu + \tau_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

for i=1,2,...,t; j=1,2,...,b; and k=1,2,...,n_i. Let $\epsilon_{ik}' = [\epsilon_{i1k}\ \epsilon_{i2k} \ldots \epsilon_{ibk}]$. Assume that the ϵ_{ik}'s are i.i.d. $N(\mathbf{0},\Sigma)$ for i=1,2,...,t; k=1,2,...,n_i. Thus ϵ_{ik} represents the bx1 vector of errors corresponding to the b repeated measures on the kth subject in the ith treatment group.

For the simple repeated measures experiment $\mathbf{Z}_1=\mathbf{0}$ which implies that $\mathbf{G}=\mathbf{0}$ and $\mathbf{V}=\mathbf{R}$. Here

$$\mathbf{R}=\mathrm{COV}(\epsilon)=\mathrm{COV}\left(\begin{bmatrix}\epsilon_{11}\\ \epsilon_{12}\\ \vdots\\ \epsilon_{tn_t}\end{bmatrix}\right)=\begin{bmatrix}\Sigma & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \Sigma & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \Sigma\end{bmatrix}, \text{ for}$$

$\mathbf{y}' = [y_{111}\ y_{121} \ldots y_{1b1}\ \ y_{112}\ y_{122} \ldots y_{1b2}\ \ \ldots\ \ y_{t1n}\ y_{t2n} \ldots\ y_{tbn}]$.

ESTIMATING G and R

Methods for estimating **G** and **R** are considered in this section.

Estimates of G and R in Split-plot Type Experiments.

For the general case,one has

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z_1u_1} + \mathbf{Z_2u_2} + \ldots + \mathbf{Z_ru_r} + \epsilon$$

where it can be assumed that

$$\mathbf{u_i} \sim \text{independent } N(\mathbf{0}, \sigma_i^2\,\mathbf{I}) \text{ for i=1,2,...r,}$$

$$\epsilon \sim N(\mathbf{0}, \sigma_\epsilon^2\,\mathbf{I}),$$

and that

$$\mathrm{COV}(\mathbf{u_i}, \epsilon) = \mathbf{0} \text{ for all i.}$$

In this case,

$$\mathbf{V}=\sigma_1^2\,\mathbf{Z_1Z_1}' + \sigma_2^2\,\mathbf{Z_2Z_2}' + \ldots + \sigma_r^2\,\mathbf{Z_rZ_r}' + \sigma_\epsilon^2\,\mathbf{I}$$

and estimating **V** is accomplished by estimating $\sigma_1^2, \sigma_2^2, \ldots \sigma_r^2$, and σ_ϵ^2.

Methods of Estimating the Variance Components

1. ANOVA (Method of Moment) Estimates

(a) Fit the GLMM by assuming that the random effects in the model are fixed effects.

(b) Next obtain the corresponding ANOVA table and compute the expected mean squares of the

observed mean squares in the ANOVA table under the true assumptions about the $\mathbf{u_i}$'s and ϵ.

(c) Equate the observed mean squares to their expected mean squares and solve the resulting system of equations for each of the variance components.

(d) Use the resulting solutions as the estimates of the variance components.

Advantages:

(a) The estimators of the variance components are unbiased.

(b) One can often approximate the degrees of freedom corresponding to the estimated standard errors of estimators of estimable functions of the fixed effects by using Satterthwaite's Method.

(c) SAS - GLM can produce the necessary information to perform these analyses.

Disadvantages:

(a) There is no unique way in which to form an ANOVA table when the data are not balanced.

(b) The procedure can produce negative estimates of the variance components which do not make sense.For the simple repeated measures model, one must estimate Σ in order to estimate **V**.

(c) If some of the expected mean squares of the random effects in the ANOVA table depend on fixed effects, the method cannot be applied. This problem can be avoided by placing all of the fixed effects in the model first followed by the random effects.

2. Maximum Likelihood Method

Find the values of $\beta, \sigma_1^2, \sigma_2^2, \ldots \sigma_r^2$, and σ_ϵ^2 that maximize the likelihood function over the parameter space.

Advantages

(a) Avoids negative estimates of the variance components.

Disadvantages

(a) Numerically intensive

(b) Resulting estimators are not unbiased.

(c) Solving the likelihood equations requires an iterative process which may or may not converge. Even when it converges, it may converge at a local maxima rather than a global maximum.

(d) Tends to underestimate the variance components.

(e) Distributional properties are not known except asymptotically.

3. Restricted Maximum Likelihood (REML)

REML estimators of the variance components are found by maximizing that part of the likelihood function that is invariant to fixed effects in the model.

The GLMM is given by

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Zu} + \epsilon\ .$$

Let **L** be a full row rank matrix that satisfies **LX**=**0** and such that rank(**L**) = n-rank(**X**) where n=the dimension of **y**.
Let $\mathbf{y}^* = \mathbf{Ly}$. Then

$$\mathbf{y}^* \sim N(\mathbf{0}, \sigma_1^2 \mathbf{LZ_1Z_1'L'} + \sigma_2^2 \mathbf{LZ_2Z_2'L'} + ... + \sigma_r^2 \mathbf{LZ_rZ_r'L'} + \sigma_\epsilon^2 \mathbf{I}).$$

The likelihood function formed from $\mathbf{y}^*$ depends only on the variance components; and the REML estimates of the variance components are those values of $\sigma_1^2, \sigma_2^2, ... \sigma_r^2$, and σ_ϵ^2 that maximize the restricted likelihood function formed from $\mathbf{y}^*$.

Advantages:

(a) Less numerically intensive than the Maximum Likelihood Method.

(b) The REML estimates and the ANOVA estimates agree when the data are balanced and all MM estimates of the variance components are non-negative.

(c) REML estimates tend to be less biased than the Maximum Likelihood Estimates.

Disadvantage:

(a) The distributional properties of these estimators are not known, except asymptotically.

Repeated Measures (Continued)

In the simple repeated measures experiment,

$\mathbf{y} = \mathbf{X}\beta + \epsilon$ where

$$\epsilon \sim N(\mathbf{0}, \begin{bmatrix} \Sigma & \mathbf{0} & . & . & \mathbf{0} \\ \mathbf{0} & \Sigma & . & . & \mathbf{0} \\ . & . & . & . & . \\ . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & . & . & \Sigma \end{bmatrix}).$$

If there are p repeated measures on each subject, then each Σ is a pxp matrix, provided that there are no missing data for any subject.

Here, in order to estimate **V**, one must estimate Σ. The estimate of Σ depends on whether one assumes any structure on Σ.

Reasonable Structures on Σ.

1. Unstructured Σ. Here it is assumed that

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & . & . & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & . & . & \sigma_{2p} \\ . & . & . & . & . \\ . & . & . & . & . \\ \sigma_{p1} & \sigma_{p2} & . & . & \sigma_{pp} \end{bmatrix}.$$

2. Compound Symmetry. Here it is assumed that

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & . & . & \rho \\ \rho & 1 & . & . & \rho \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho & \rho & . & . & 1 \end{bmatrix}.$$

3. Huyhn-Feldt Conditions. Here it is assumed that $\Sigma = \lambda I + \acute{\eta}\mathbf{j}' + \mathbf{j}\acute{\eta}'$ for some λ and some $\acute{\eta}$. This implies

$$\Sigma = \begin{bmatrix} \lambda+2\eta_1 & \eta_1+\eta_2 & . & . & \eta_1+\eta_p \\ \eta_2+\eta_1 & \lambda+2\eta_2 & . & . & \eta_2+\eta_p \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \eta_p+\eta_1 & \eta_p+\eta_2 & . & . & \lambda+2\eta_p \end{bmatrix}.$$

4. AR(1) . Here it is assumed that

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & . & \rho^{p-1} \\ \rho & 1 & \rho & . & \rho^{p-2} \\ \rho^2 & \rho & 1 & . & \rho^{p-3} \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ \rho^{p-1} & \rho^{p-2} & \rho^{p-3} & . & 1 \end{bmatrix}.$$

In each of the above structures for Σ as well as other possible structures , estimating **V** requires one to estimate the parameters in Σ. For the unstructured case and the compound symmetry case, Σ can be estimated either by method of moments (MoM), maximum likelihood (ML), or restricted maximum likelihood (REML). If there are no missing data on any subject, the estimates will agree. For all other cases, the parameters in Σ must be estimated by ML or REML methods.

Finally, in each of these cases, once estimates of the variance parameters are found, **V** can be estimated, and then $\ell'\beta$ is estimated by $\ell'\hat{\beta}_{\hat{V}}$ where $\hat{\beta}_{\hat{V}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^- \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y}$. The standard error of $\ell'\hat{\beta}_{\hat{V}}$ is estimated by $\sqrt{\ell'(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^-\ell}$.

CONTACT INFORMATION

Dallas E. Johnson
Department of Statistics
101 Dickens Hall
Kansas State University
Manhattan, KS 66506-0802
Work Phone: 785-532-0510
Fax: 785-532-7736
Email: dallas@stat.ksu.edu
Web: http://www.ksu.edu/stats/facultypages/johnson.htm