Mixed Effects Models & Longitudinal Data Analysis

Jiming Jiang

University of California, Davis, USA

March 2005

1. Linear Mixed Models

1.1. Introduction

• What is a mixed model?

• Random effects

a. Unobsered (random) variables of practical interest

b. Way of modelling the correlations among the observations

• Some References: Rao and Kleffe (1988), Robinson (1991), Searle *et al.* (1992), Khuri *et al.* (1998), McCulloch and Searle (2000), among others.

1.2. Some real-life examples

1. Effect of air pollution episodes on children

• Source: Laird and Ware (1982)

• Data: 200 school children examined under normal conditions, then during an air pollution alert and on three successive weeks following the alert. Data is longitudinal.

• Objective: FEV_1 = volume of air exhaled in the first second of a forced exhalation depressed during the alert?

• A simple linear mixed model: $y_{ij} = \beta_j + \alpha_i + \epsilon_{ij}$, $i = 1, \ldots, m$, $j = 1, \ldots, 5$, where β_j is the mean FEV_1 for the jth observational time, α_i is a random effect associated with the ith child.

2. Prediction of maize single-cross

• Source: Bernardo (1996) reported results of BLUP of single-cross performance and genetic relationship among parental inbreds.

• Data: Grain yield, moisture, stalk lodging and root lodging obtained for 2043 maize single crosses evaluated in the multilocation testing program of Limagrain Genetics, from 1990 to 1994.

• Objective: Robustness of BLUP for identifying superior single crosses when estimates of genetic relationship among inbreds are erroneous.

• A linear mixed model:

$$y = X\beta + Z_0 c + Z_1 g_1 + Z_2 g_2 + Zd + e,$$

where y is a vector of observed performace for a given trait, c is a vector of check effects; g_j is a vector of general combining ability effects of Group j, $j = 1, 2$; d is a vector of specific combining ability effects; and e is a vector of residual effects.

The covariance matrices of c, g_1, g_2, d and e are modeled according to the genetic relationships.

3. Small area estimation of income

• Small area estimation

• Source for the current example: Fay and Herriot (1979)

• Objective: Estimation of per-cpita income (PCI) for small places from the 1970 Census of Population and Housing.

• Data: Income was collected on the basis of a 20 percent sample. However, of the estimates required, more than one-third, or approximately 15,000, are for places with population of fewer than 500 persons.

• Fay and Herriot proposed the following linear mixed model, which is later known as the Fay-Herriot model:

$$y_i = x_i'\beta + v_i + e_i,$$

where y_i is the natural logarithm of the sample estimate of PCI for the ith place; x_i is a vector of known covariates related to the place; β is a vector of unknown regression coefficients; v_i is a random effect associated with the place; and e_i represents the sampling error.

It is assumed that v_i and e_i are distributed independently such that $v_i \sim N(0, A)$, $e_i \sim N(0, D_i)$, where A is unknown but D_i's are known.

1.3. Linear mixed model

- Linear regression: $y = X\beta + \epsilon$

- Linear mixed model: $y = X\beta + Z\alpha + \epsilon$

Example 1.1. y_{ij} = obs. from the i subject collected at time t_j; α_i = random effect associated with the ith individual, $i = 1, \ldots, m$, $j = 1, \ldots, k$.

A linear mixed model may be expressed as

$$y_{ij} = x'_{ij}\beta + \alpha_i + \epsilon_{ij},$$

where ...

The correlation between any two obs. from the same individual is $\sigma^2/(\sigma^2 + \tau^2)$; while obs. from different individuals are uncorrelated.

1.4. Types of linear mixed models

• Gaussian (linear) mixed model

More generally, $y \sim N\{X\beta, V(\theta)\}$, where θ is a vector of dispersion parameters.

• Non-Gaussian linear mixed model

Alternatively, the models may be classified as

• Mixed ANOVA model

$$y = X\beta + Z_1\alpha_1 + \cdots + Z_s\alpha_s + \epsilon,$$

where $\alpha_1, \ldots, \alpha_s, \epsilon$ are indep.;
the components of α_j are i.i.d. with mean 0 and variance σ_j^2, $1 \leq j \leq s$;
the components of ϵ are i.i.d. with mean 0 and variance σ_0^2. Hence $\theta = (\sigma_0^2, \sigma_1^2, \ldots, \sigma_s^2)'$.

• Longitudinal Model

$$y_i = X_i\beta + Z_i\alpha_i + \epsilon_i, \quad i = 1, \ldots, m,$$

where y_i = a vector of obs. from the ith subject; α_i = a subject-specific vector of random effects; ϵ_i = a vector of errors.

Example 1.2. (Growth curve model)

$$y_{ij} = \xi_i + \eta_i x_{ij} + \zeta_{ij} + \epsilon_{ij},$$

where ξ_i = a random intercept;
η_i = a random slope;
x_{ij} = a known covariate;
ζ_{ij} = a serial correlation;
ϵ_{ij} = a measurement error.

It is assume that ξ_i and η_i are jointly normally distributed with means μ_1, μ_2, variances σ_1^2, σ_2^2, corr. ρ; $\zeta_{ij} \sim$ AR(1); ϵ_{ij} indep. $\sim N(0, \tau^2)$.

Example 1.3. (Normal longitudinal model) $y_i \sim N\{X_i\beta, V_i(\theta)\}$, $i = 1, \ldots, m$. For example, in Example 1.2, $\theta = (\tau^2, \sigma_1^2, \sigma_2^2, \rho, \ldots)'$.

1.5. Estimation of variance components

- Why variance components?

- Earlier methods: ANOVA (Henderson 1953), MINQUE (Rao 1972), I-MINQUE, etc.

Advantages: Relatively easy to compute, not requiring normality.

Disadvantages: ANOVA est. are inefficient; MINQUE depends on the initial values; both est. can fall outside the parameter space.

• Maximum likelihood

If normality is assumed, the efficient estimators of the variance components are the MLE. But not until Hartley and Rao (1967). Asymptotic properties of the MLE were studied by Miller (1977).

Bias of the MLE: Neyman and Scott (1948).

Example 1.4. (The Neyman-Scott problem)

$$y_{ij} = \mu_i + \epsilon_{ij}, i = 1, \ldots, m, j = 1, 2,$$

where μ_{ij} are unknown and $\epsilon_{ij} \sim N(0, \sigma^2)$. The problem of interest is to estimate σ^2.

• Restricted Maximum likelihood (REML)

Thompson (1962), Patterson and Thompson (1971).

Example 1.4 (continied).

$z_i = y_{i1} - y_{i2} \sim N(0, 2\sigma^2)$. Estimate σ^2 by ML based on z_i, $i = 1, \ldots, m$.

In general, assume that rank$(X) = p$. Let A be a $n \times (n-p)$ matrix of full rank such that $A'X = 0$ ($n = \dim(y)$). The REML est. of the variance components are the MLE based on $z = A'y$.

Some nice properties of REML:

1. Not dependent on A;

2. can be derived from different point of view (e. g., Harville 1974, Barndorff-Nielson 1983, Verbyla 1994, Heyde 1994, Jiang 1996);

3. no loss of information in estimating the variance components (e. g., Patterson and Thompson 1971, Harville 1977, Jiang 1996);

4. consistency and asymptotic normality even without normality and with $p \to \infty$ (Jiang 1996, 1997a).

1.6. Inference about non-Gaussian linear mixed models

• Normality assumption is likely to be violated in practice (e.g., Lange and Ryan 1989).

• Difficulties of inference without normality

• Gaussian (restricted) likelihood inference:
e.g., Richardson and Welsh (1994),
Jiang (1996, 1997a), Heyde (1994, 1997).

• Asymptotic Covariance Matrix (ACM)

θ - the vector of variance components, $\hat{\theta}$ - the (Gaussian) REML estimator.

The ACM of $\widehat{\theta}$ is given by $\Sigma = \mathcal{I}_2^{-1}\mathcal{I}_1\mathcal{I}_2^{-1}$, where

$$\begin{aligned} \mathcal{I}_1 &= \mathrm{Var}\left(\frac{\partial l_{\mathrm{R}}}{\partial \theta}\right), \\ \mathcal{I}_2 &= \mathrm{E}\left(\frac{\partial^2 l_{\mathrm{R}}}{\partial \theta \partial \theta'}\right) \end{aligned}$$

l_{R} is the (Gaussian) restricted log-likelihood.

• Usually, $\mathcal{I}_2$ involves only the variance components.

• But $\mathcal{I}_1$ involves higher (3rd, 4th) moments of the random effects and errors. $\mathcal{I}_1$ is called the quasi information matrix (QUIM).

The main problem is to estimate the QUIM.

- Estimated and Observed Information

In the i.i.d. case,

$$\mathrm{Var}\left(\frac{\partial l}{\partial \theta}\right) = \sum_{i=1}^{n} \mathrm{E}\left(\frac{\partial l_i}{\partial \theta}\right)^2 = \mathrm{E}\left\{\sum_{i=1}^{n}\left(\frac{\partial l_i}{\partial \theta}\right)^2\right\}.$$

Estimated:

$$\mathrm{Var}\left(\frac{\partial l}{\partial \theta}\right)\Big|_{\theta=\hat{\theta}}.$$

Observed:

$$\sum_{i=1}^{n}\left(\frac{\partial l_i}{\partial \theta}\right)^2\Big|_{\theta=\hat{\theta}}.$$

Linear mixed model:

- Why can't we estimate?

- Why don't we observe?

• Partially Observed Information

Example 1.5.

$$y_{kl} = \mu + u_k + v_l + e_{kl},$$

$k = 1, \ldots, m$, $l = 1, \ldots, n$, where
μ is an unknown mean,
u_k, v_l are random effects,
and e_{kl} is an error such that
u_k's are i.i.d. with mean 0 and variance σ_1^2,
v_l's are i.i.d. with mean 0 and variance σ_2^2,
e_{kl}'s are i.i.d. with mean 0 and variance σ_0^2,
and u, v, e are independent.

Let $\lambda = \sigma_0^2$ and $\gamma_j = \sigma_j^2/\sigma_0^2$, $j = 1, 2$.

Consider, for example, $\mathrm{var}(\partial l_R/\partial\lambda)$, which is a diagonal element of the QUIM $\mathrm{Var}(\partial l_R/\partial\theta)$, where $\theta = (\lambda, \gamma_1, \gamma_2)'$.

It can be shown that $\mathrm{var}(\partial l_R/\partial\lambda) = S_1 + S_2$, where $S_1 = \mathrm{E}(\cdots)$ with

$$\begin{aligned}\cdots &= (a_0 + a_1 + a_2)\sum_{i,j} u_{ij}^4 \\ &\quad -a_1\sum_i\left(\sum_j u_{ij}\right)^4 - a_2\sum_j\left(\sum_i u_{ij}\right)^4,\end{aligned}$$

a_r, $r = 0, 1, 2$ and S_2 depend only on λ, γ_1 and γ_2. Conclusion:

- S_1 can be estimated by $\widehat{S}_1$, which is $\cdots$ with θ replaced by $\widehat{\theta}$, the MLE, say;
- S_2 can be estimated by $\widehat{S}_2$, which is S_2 with the variance components replaced by the MLE;
- and $\mathrm{var}(\partial l_R/\partial\lambda)$ estimated by $\widehat{S}_1 + \widehat{S}_2$.

The estimator is called partially observed information matrix, or POQUIM, because part of it is observed and part of it estimated.

In general, we have

Theorem 1 (Jiang 2004). For any non-Gaussian linear mixed model, we have $\mathcal{I} = (\mathcal{I}_{jk})_{1\le j,k\le s}$, where $\mathcal{I}_{1,jk} = \mathcal{I}_{1,1,jk} + \mathcal{I}_{1,2,jk}$,

$$\mathcal{I}_{1,1,jk} = \mathsf{E}\left\{\sum_{\cdots} c_{jk}(i_1,\ldots,i_4)u_{i_1}\cdots u_{i_4}\right\},$$

where $\cdots$ represent an index set, $u_i = y_i - x_i'\beta$, and $c_{jk}(i_1,\ldots,i_4)$ and $\mathcal{I}_{1,2,jk}$ depend only on θ.

An estimator of $\mathcal{I}_{1,jk}$ is then $\hat{\mathcal{I}}_{1,1,jk} + \hat{\mathcal{I}}_{1,2,jk}$, where

$$\hat{\mathcal{I}}_{1,1,jk} = \sum_{\cdots} \hat{c}_{jk}(i_1,\ldots,i_4)\hat{u}_{i_1}\cdots\hat{u}_{i_4}$$

($\hat{c}$ and $\hat{u}$ represent the corresponding quantities with θ replaced by $\hat{\theta}$), and $\hat{\mathcal{I}}_{1,2,jk}$ is $\mathcal{I}_{1,2,jk}$ with θ replaced by $\hat{\theta}$.

See Jiang (2004, *Ann. Statist.*, in press) for more detail.

1.7. Prediction of random effects

C. R. Henderson pioneered the prediction of random effects, or mixed effects, with his early work in animal breedings (Henderson 1948). See, e. g., Robinson (1991), Ghosh & Rao (1994) for review and applications.

• A mixed effect may be expressed as $\eta = b'\beta + a'\alpha$, where a, b are known vectors. If β and θ are known, the best predictor under normality is $\tilde{\eta}$, which is η with α replace by

$$\tilde{\alpha} = \mathrm{E}(\alpha|y) = GZ'V^{-1}(y - X\beta),$$

where $G = \mathrm{Var}(\alpha)$ and $V = \mathrm{Var}(y)$.

• Since β is unknown in practice, it is customary to replace it by

$$\tilde{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

which is the MLE of $\widehat{\beta}$ under normality, provided that θ is known. The result is the best linear unbiased predictor, or BLUP.

Example 1.6. Robinson (1991) used the following example to illustrate the calculation of BLUP. Consider a linear mixed model for the first lactation yields of dairy cows with sire additive genetic merits being treated as random effects and herd effects being treated as fixed effects. The herd effects are represented by β_j, $j = 1, 2, 3$ and sire effects by α_i, $i = 1, 2, 3, 4$, which correspond to sires A, B, C, D. Suppose that the data is given below.

Herd	1	1	2	2	2
Sire	A	D	B	D	D
Yield	110	100	110	100	100
Herd	3	3	3	3	
Sire	C	C	D	D	
Yield	110	110	100	100	

Suppose that $R = I$ and $G = 0.1I$. The BLUP is given by

$$\tilde{\alpha} = (0.40, 0.52, 0.76, -1.67)'.$$

• Different derivations of BLUP: Henderson (1950, 1973), Harvill (1990), Jiang (1997b).

• The expression of BLUP involves θ, the vector of variance components, which is typically unknown. It is customary to replace θ by a consistent estimator, $\hat{\theta}$. The resulting predictor is often called empirical BLUP, or EBLUP. Denoted by $\hat{\eta}$.

• Properties of EBLUP:

1. Unbiasedness: Kackar and Harville (1981) showed that, if $\hat{\theta}$ is an even and translation invariant estimator (e. g., ANOVA, ML and REML est.), and the data is normal, EBLUP remains unbiased.

2. Existence of moments: Jiang (2000).

3. Asymptotic properties: Jiang (1998).

4. EBLUP and Empirical Bayes (EB): Harville (1991)

• MSE of EBLUP

Applications: Small area estimation (SAE; e. g., Ghosh and Rao 1994, Rao 2000).

Studies: Kackar and Harville (1984), Prasad and Rao (1990). The latter gave second order approximation and estimation of the MSE of EBLUP for two important special linear mixed models used in SAE, the Fay-Herriot model (Fay and Herriot 1979) and nested error regression model (Battese *et al.* 1988).

Recently, Das *et al.* (2004) extended Prasad and Rao's work to general linear mixed models.

• A jackknife method (Jiang *et al.* 2002):

Denote $\mathrm{MSE}(\tilde{\eta})$ by $b(\theta)$. The jackknife estimator of $\mathrm{MSE}(\hat{\eta})$ is given by

$$\widehat{\mathrm{MSE}}(\hat{\eta}) = \widehat{\mathrm{MSAE}}(\hat{\eta}) + \widehat{\mathrm{MSE}}(\tilde{\eta}),$$

where

$$\widehat{\mathrm{MSAE}}(\hat{\theta}) = \frac{m-1}{m} \sum_{i=1}^{m} (\hat{\eta}_{-i} - \hat{\eta})^2 ,$$

$$\widehat{\mathrm{MSE}}(\tilde{\eta}) = b(\hat{\theta}) - \frac{m-1}{m} \sum_{i=1}^{m} \left\{ b(\hat{\theta}_{-i}) - b(\hat{\theta}) \right\}.$$

Here m = the number of clusters (e.g., number of small areas), $\hat{\theta}_{-i}$ denotes an M-estimator of θ using the data without the ith cluster (e.g., the ith small area), and $\hat{\eta}_{-i}$ the EBLUP of η in which the <u>fixed</u> parameters are estimated using the data without the ith cluster. We have

$$\mathrm{E}\{\widehat{\mathrm{MSE}}(\hat{\eta})\} = \mathrm{MSE}(\hat{\eta}) + o(m^{-1}).$$

1.8. Other types of inference

• Bayesian inference: e. g., Hill (1965), Tiao and Tan (1965, 1966), Gianola and Fernando (1986), Gelman *et al.* (2003).

• Tests in linear mixed models: e. g., Khuri *et al.* (1998), Jiang (2003, 2004).

• Confidence intervals: Burdick and Graybill (1992).

• Prediction intervals: Jeske and Harville (1988), Jiang and Zhang (2002).

• Mixed model diagnostics

Informal model checking: Dempster and Ryan (1985), Lange and Ryan (1989), Calvin and Sedransk (1991);

Formal model checking: Jiang *et al.* (2001), Jiang (2001).

• Mixed model selection: Jiang and Rao (2003).