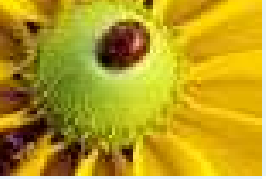
Mixed Model Prediction: Part I

Jiming Jiang and Thuan Nguyen

University of California, Davis, USA

and Oregon Health & Science University, USA



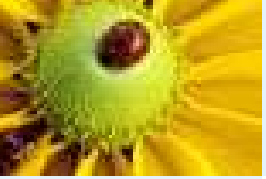
Introduction

- Mixed effects models, including linear mixed models, generalized linear mixed models, and nonlinear mixed effects models, are broadly used in practice.

  See, for example, Jiang (2007), McCulloch, Searle & Neuhaus (2008), Demidenko (2013), Rao & Molina (2015).

  These models are useful in modeling correlated data (e.g., longitudinal data analysis), estimating variance components (e.g., genetic studies), making subject-level inference (e.g., precision medicine), and making predictions (e.g., small area estimation).

  This is a field where frequentist and Bayesian approaches often find common grounds.



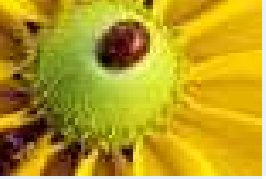
Examples

- *Example 1*: (Effect of air pollution on children) Source: Laird & Ware (1982; *Biometrics*).

  A study of ffect of air pollution episodes on pulmonary function in children.

  About 200 school children were examined (i) under normal conditions; then (ii) during an air pollution alert; then (iii) on three successive weeks following the alert.

  Correlation among the observations? and consequence?

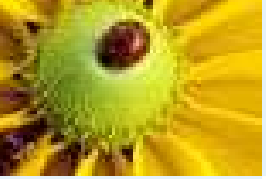


- Example 2. (Prediction of maize single-cross performance) Source: Bernardo (1996; *Crop Science*).

  A plant-genetic study. Grain yields, moisture, lodging and root lodging data were collected for 2043 maize single crosses evaluated in the multi-location testing program of Limagrain Genetics, 1991–1994.

  In most genetic studies, the observations are correlated due to the sharing of genetic effects.

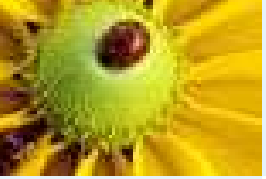
  In this case, the shared genetic effects include check effects, general combining ability group effects, and specific combining effects.



- Example 3. (Small area estimation) Sourse: Jiang & Lahiri (2006; *TEST*)

  The surveys are usually designed to produce reliable estimates of various characteristics of interest for large geographic areas.

  However, for effective planning of health, social and otherservices, and for apportioning government funds, there is a growing demand in producing similar estimates for small geographic areas and sub-populations.

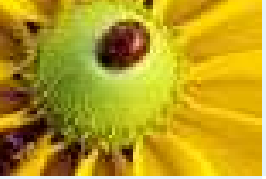


- For example, in a statewide telephone survey of sample size 4,300 in the state of Nebraska, only 14 observations were available to esrtimate the prevalence of alcohol abuse in Boone county.

  The situation is even worse for direct survey estimation of the prevalence for white female in the age-group of 25-44 in the county, as only one observation was available.

  By using small area estimation methods (e.g., Rao & Molina 2015), one is able to “borrow strength” from other small areas or sources, and therefore improve accuracy of the estimation.

  The borrowing of strength is typically through a mixed effects model (more later).



Linear mixed models

- A linear mixed model (LMM) may be viewed as adding random effects to a linear regression model.

  From a linear regression model:

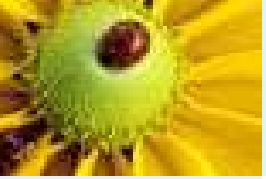
$$y = X\beta + \epsilon$$

  to a LMM:

$$y = X\beta + Z\alpha + \epsilon,$$

  where Z is a known (design) matrix, and α is a vector of random effects. In contast, the vector β is called fixed effects.

  It is typically assumed that the random effects have mean zero, i.e., $\mathrm{E}(\alpha) = 0$; in other words, if $\mathrm{E}(\alpha)$ is not zero, it is part of $X\beta$.

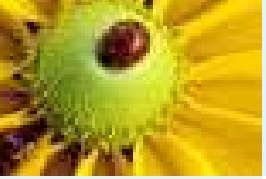


- Assumptions: (i) $\mathrm{E}(\alpha) = 0, \mathrm{E}(\epsilon) = 0$;

  (ii) $\mathrm{Var}(\alpha) = G, \mathrm{Var}(\epsilon) = R, \mathrm{Cov}(\alpha, \epsilon) = 0$;

  typically, the covariance matrices G, R depend on a vector, θ, of dispersion parameters, or variance components.

  (iv) quite often, normality of α and ϵ is assumed, implying that the data are normal.



Inference about LMM

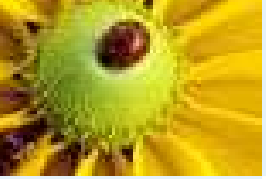
- Historical note: ANOVA estimation: $y'Ay = \mathrm{E}(y'Ay)$, where A is symmetric such that $AX = 0$.

  What A (or A's)?

  Balanced data: ANOVA table.

  Unbalanced data: Henderson's methods I, II, III.

  Nowadays, the ANOVA method is no longer popular. The standard methods of mixed model analysis are maximum likelihood (ML) and restricted maximum likelihood (REML).



- ML: First introduced by Hartley & Rao (1967). Under the normality assumption, the log-likelihood can be expressed as

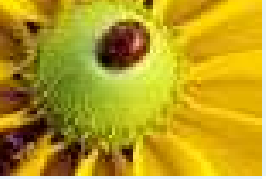
$$l(\beta,\theta) = c - \frac{1}{2}\{\log(|V|) + (y - X\beta)'V^{-1}(y - X\beta)\},$$

where ...

Likelihood equation:

$$\frac{\partial l}{\partial \beta} = X'V^{-1}y - X'V^{-1}X\beta = 0,$$

$$\begin{aligned}\frac{\partial l}{\partial \theta_r} &= \frac{1}{2}\left\{(y - X\beta)'V^{-1}\frac{\partial V}{\partial \theta_r}V^{-1}(y - X\beta)\right. \\ &\quad \left. -\mathrm{tr}\left(V^{-1}\frac{\partial V}{\partial \theta_r}\right)\right\} = 0, \;\; 1 \le r \le q.\end{aligned}$$

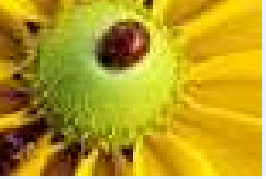


- Computation: Typically involving numerical solution to nonlinear equation system; available in software package (e.g., SAS, R).

  Other approaches: E-M algorithm (Dempster *et al.* 1977), etc.

  Note: Difference between solution to the ML equation & maximum likelihood estimator (MLE). They are not necessarily the same.

  Asymptotic properties of MLE: Miller (1977).



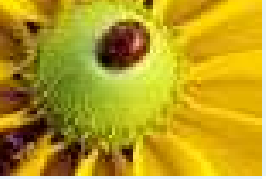
- REML: It is known that the MLE of the variance components can be seriously biased when the LMM involves a large number of fixed effects.

  Example 4 (Neyman-Scott problem). Suppose that there are m patients, and two measurements are collected from each patient.

  Suppose that the observations satisfy

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

  where $\mu_1, \ldots, \mu_m$ are unknown means; the ϵ_{ij}'s are independent $N(0, \sigma^2)$.

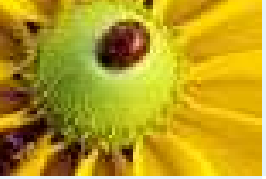


- The interest is to estimate σ^2, which may be interpreted as precision of the measurement.

  It can be shown that the MLE of σ^2 is inconsistent.

  Reason? Note that, with the ML procedure, one needs to estimate all the unknown parameters together.

  There are $m + 1$ unknown parameters; while the total sample size is $n = 2m$.



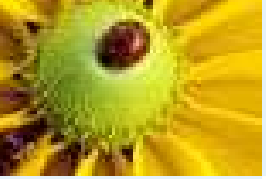
- Alternative approach?

  Let us take the differences, $z_i = y_{i1} - y_{i2}, i = 1, \ldots, m$.

  Note that, by taking the differences, the unknown means, μ_i, have "disappeared". In fact, the z_i's are independent $\sim N(0, 2\sigma^2)$.

  Now, the sample size is $n = m$ but there is only one unknown parameter. In fact, when applying the ML procedure based on $z_1, \ldots, z_m$, the resulting MLE is consistent.

  The latter procedure is a special case of REML, restricted (or residual) maximum likelihood.



- In general, consider the Gaussian LMM

(1) $$y = X\beta + Z\alpha + \epsilon,$$

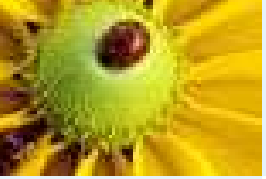
where, for simplicity (but w.l.o.g.), let X be $n \times p$ with $\text{rank}(X) = p$, $\alpha \sim N(0, G)$, $\epsilon \sim N(0, R)$, and α, ϵ indepenent.

Let A be an $n \times (n - p)$ matrix with

(2) $$\text{rank}(A) = n - p \ \text{ and } \ A'X = 0.$$

By multiplying A' on both sides of (1), one gets $z = A'y = A'(Z\alpha + \epsilon) \sim N(0, A'VA)$, where $V = R + ZGZ'$.

Note that the distribution of z depends only on θ, the vector of variance components involved in V.



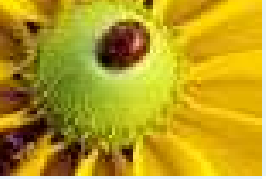
- Now apply the ML based on z to estimate θ. The result is called REML estimator.

  The log-likelihood function, known as restricted log-likelihood, has the expression

$$l_{\mathrm{R}} = c - \frac{1}{2}\{\log(|A'VA|) + z'(A'VA)^{-1}z\},$$

  where c is a constant.

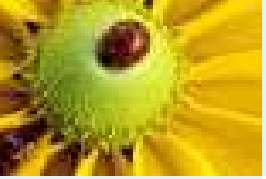
  Note that, although the derivation of l_{R} involves A, the REML estimator actually does not depend on the choice of A, so long as (2) is satisfied.



- Asymptotic behavior of REML estimator: Das (1979), Cressie & Lahiri (1993), Richardson & Welsh (1994), Jiang (1996, 1997a).

  In particular, Jiang (1996) showed that, even if the random effects and errors are not normal, but one still uses the normality based inference, the (Gaussian) REML estimators remain consistent and asymptotically normal.

  The asymptotic distributions of the REML or ML estimators are used in making inference about the fixed effects and variance components in LMM. See, for example, Jiang (2007, 2017).



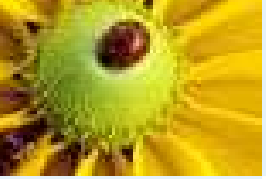
Example: Lambs data

- In one of the earlier applications of mixed effects models to animal genetics, Haville & Fenech (1985) published a study of mixed model analysis on lambs' weight data.

  The data consist of birth weights of 62 single-birth male lambs.

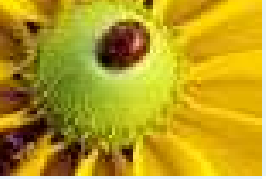
  The lambs were progenies of 23 rams, so that each lamb had a different dam.

  Another factor that was considered was the (distinct) population lines. There were two control lines and three selection lines.



Partial data

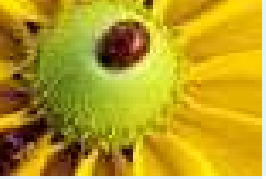
Obs.	6.2	13.0	9.5	10.1	11.4	11.8	12.9	13.1	10.4
Sire	11	12	13	13	13	13	13	13	14
Line	1	1	1	1	1	1	1	1	1
Age	1	1	1	1	1	2	3	3	1
Obs.	8.5	13.5	10.1	11.0	14.0	15.5	12.0	11.5	10.8
Sire	14	21	22	22	22	22	23	24	24
Line	1	2	2	2	2	2	2	2	2
Age	2	3	2	3	3	3	1	1	3



- Haville and Fenech considered the following LMM:

$$y_{ijk} = l_i + a_1 x_{ijk,1} + a_2 x_{ijk,2} + s_{ij} + e_{ijk},$$

$i = 1, \ldots, 5$, $j = 1, \ldots, n_i$, and $k = 1, \ldots, n_{ij}$, where y_{ijk} is the birth weight of the kth lamb who is a progeny of the jth sire (ram; father) in line i; l_i is the fixed line effect; $x_{ijk,1}, x_{ijk,2}$ are indicators of the age category of the dam (mother): 1-2 years, 2-3 years, and 3+ years; s_{ij} is the random sire effect; and e_{ijk} corresponds to environmental error.



- Analysis using the R package **nlme** (**lme4** does the same)

Restricted maximum likelihood (REML):

Effect	Line	Age	Est.	S.E.	t-value	$\Pr > \|t\|$
Line	1		10.5008	0.8070	13.01	$< .0001$
Line	2		12.2998	0.7569	16.25	$< .0001$
Line	3		11.0425	0.6562	16.83	$< .0001$
Line	4		10.2864	0.7882	13.05	$< .0001$
Line	5		10.9625	0.5438	20.16	$< .0001$
Age		1	-0.0097	0.5481	-0.02	0.9861
Age		2	-0.1651	0.6435	-0.26	0.7989

The estimates of σ_s^2 and σ_e^2 are 0.511 and 2.996, respectively.