Introduction to the Linear Mixed Model

Bahjat F. Qaqish
BIOS 667

The basic idea:
Subject-specific random effects (coefficients, parameters, intercepts and slopes)

Variance decomposition: Variance components, within and between
Within: R_i
Between: G
More on this follows below.

Why use mixed models?
(1) Interested in estimating the variance components.
(2) Interested in subject-specific predictions, i.e. prediction of the subject-specific random effects.

We'll use our established notation, $Y_{ij}, x_{ij}, Y_i, X_i, n_i, Y, X, K$.

Three levels of notation: one observation, one cluster, all clusters.

Notation for the conditional mean (given the random effects) of one observation:

$$\nu_{ij} := \mathrm{E}[Y_{ij}|b_i] = x_{ij}^\top \beta + z_{ij}^\top b_i$$

This is sometimes written in the less informative form

$$Y_{ij} = x_{ij}^\top \beta + z_{ij}^\top b_i + \epsilon_{ij}$$

Y_{ij} and ϵ_{ij} are scalars; ϵ_{ij} is a random variable (the within-subject error).
The covariate vectors x_{ij} are $p \times 1$.
The covariate vectors z_{ij} are $q \times 1$.
The vector b_i is $q \times 1$.

Notation for one cluster (e.g. one subject):

$$\nu_i := \mathrm{E}[Y_i|b_i] = X_i\beta + Z_i b_i$$

Again, this is sometimes written as

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i.$$

Y_i and ϵ_i are $n_i \times 1$ vectors.
X_i is $n_i \times p$, with $x_{ij}^\top$ in the jth row.
Z_i is $n_i \times q$, with $z_{ij}^\top$ in the jth row.
The *residual* or *within-subject* variance component is

$$\mathrm{cov}(Y_i|b_i) = R_i.$$

In many models, R_i is the same for all subjects. And in many models, R_i is diagonal (if it is thought that the random effects account for all the correlation within a cluster). The *between-subject* variance component is

$$\text{cov}(b_i) = G.$$

G is assumed to be the same for all subjects, at least in the initial models that will be considered, to keep things simple.

Notation for all the clusters:

$$\nu := \text{E}[Y|b] = X\beta + Zb$$

Again, this is sometimes written in the form

$$Y = X\beta + Zb + \epsilon.$$

Define $N := n_1 + \ldots + n_K$.
The response vector Y is $N \times 1$, obtained by vertically concatenating the vectors $Y_1, \ldots, Y_K$.
The error vector ϵ is $N \times 1$, obtained by vertically concatenating the vectors $\epsilon_1, \ldots, \epsilon_K$.
The covariate matrix X is $N \times p$, obtained by vertically concatenating the matrices $X_1, \ldots, X_K$.
The covariate matrix Z is $N \times Kq$, block-diagonal (but the blocks are not square matrices), with Z_i in the ith diagonal block. The vector of random effects b is $Kq \times 1$, obtained by vertically concatenating the vectors $b_1, \ldots, b_K$. There is a total of Kq scalar random effects.

$R := \text{cov}(Y|b)$ is block diagonal with ith block R_i.

Notice how X and Z are constructed differently. The reason for having Kq columns in Z is that each of the K subjects contributes q random effects. Even a subject with one observation ($n_i = 1$) contributes q random effects to the vector b.

The marginal picture (the induced or implied marginal model):
(All are obtained via the double-expectation formulae).

The mean and variance for one observation:

$$\mu_{ij} := \text{E}[Y_{ij}] = x_{ij}^\top \beta$$
$$\text{var}(Y_{ij}) = z_{ij}^\top G z_{ij} + R_{ijj}$$
$$\text{cov}(Y_{ij}, Y_{ik}) = z_{ij}^\top G z_{ik} + R_{ijk}$$

We'll always take $1 \leq j \neq k \leq n_i$.

The mean and covariance for one cluster (e.g. one subject):

$$\mu_i := \text{E}[Y_i] = X_i\beta$$

$$\Sigma_i := \text{cov}(Y_i) = Z_i G Z_i^\top + R_i.$$

The last equation is the actual expression of the total variance as the sum of between ($Z_i G Z_i^\top$) and within (R_i) components.

The mean and covariance for all the clusters:

$$\mu := \mathrm{E}[Y] = X\beta$$

$\Sigma := \mathrm{cov}(Y)$ is block diagonal with ith block Σ_i. Subjects are independent, so $\mathrm{cov}(Y_{ij}, Y_{kl})$ is zero if $i \neq k$.

Exercise: Write out the dimensions of all the above vectors and matrices for the TLC data, assuming x_{ij} consists of intercept, an indicator for group A and linear time, $p = 3$. Also assume a random intercept and a random slope for time, $q = 2$. There are 100 subjects, each with 4 observations.

Exercise: Write out the dimensions of all the above vectors and matrices for the MIT Growth and Development Study data set (data on percent body fat), assuming x_{ij} consists of intercept, time and time^2, $p = 3$. Also assume a random intercept and random slopes for time and time^2, $q = 3$.

Example 1:
A model with a random intercept.
$R_i = \sigma_w^2 I_{n_i \times n_i}$.
G is 1×1 , just one number, $G = [\sigma_b^2]$.
$Z_i = 1_{n_i \times 1}$, a column of n_i 1's.
$\Sigma_i = \sigma_b^2 11^\top + R_i$.
All diagonals of Σ_i equal $\sigma_b^2 + \sigma_w^2$.
All off-diagonals of Σ_i equal σ_b^2.
Exchangeable or CS covariance matrix Σ_i.
Exchangeable correlation.
Within-subject correlation: $j \neq k, \mathrm{corr}(Y_{ij}, Y_{ik}) = \sigma_b^2/(\sigma_b^2 + \sigma_w^2)$.
The ratio "between variance / total variance" is known as the *intra-class correlation*. It also has other names such as repeatability, reliability, reliability coefficient, etc.

Example 2:
A model with a random intercept and a random slope for time.
$R_i = \sigma_w^2 I_{n_i \times n_i}$.
G is 2×2.
Z_i has two columns; 1 and "time" values t_{ij}.

$$\nu_{ij} := \mathrm{E}[Y_{ij}|b_i] = x_{ij}^\top \beta + b_{i1} + b_{i2} t_{ij}.$$

x_{ij} would usually contain t_{ij}.

$$\mathrm{var}(Y_{ij}) = g_{11} + t_{ij}^2 g_{22} + 2t_{ij} g_{12} + \sigma_w^2.$$

For $j \neq k$,

$$\mathrm{cov}(Y_{ij}, Y_{ik}) = g_{11} + t_{ij} t_{ik} g_{22} + (t_{ij} + t_{ik}) g_{12}.$$

Clearly, $\mathrm{corr}(Y_{ij}, Y_{ik})$ is a complicated function of t_{ij}, t_{ik}, G and R_i.

The likelihood function:
If b_i is normal and Y_i given b_i is normal then, marginally, $Y \sim N_N(\mu, \Sigma)$. This defines the likelihood for the parameters that determine (β, R, G). A REML approach is also possible.

The dual interpretation of β:
(1) Conditional: Contrasts in ν.
(2) Marginal: Contrasts in μ.
This is specific to the *linear* mixed model, i.e. the identity link function.
We'll see later that it does not hold for other link functions such as logit, probit, etc.

Interpretation of R_i and G is best done by looking at:
(1) Variances (the diagonals of Σ_i)
(2) Correlations, $\mathrm{corr}(Y_{ij}, Y_{ik})$

Non-identifiability can arise quite easily. Extra care is needed. Example: Random intercept with a CS within-subject structure R_i.

Prediction of the subject-specific random effects:
BLUPs and EBLUPs:
The model implies that $(Y_i^\top, b_i^\top)^\top$ follows a multivariate normal distribution with $\mathrm{E}[Y_i] = \mu_i$, $\mathrm{E}[b_i] = 0$, $\mathrm{cov}(Y_i) = \Sigma_i$, $\mathrm{cov}(b_i) = G$, $\mathrm{cov}(Y_i, b_i^\top) = Z_i G$, $\mathrm{cov}(b_i, Y_i^\top) = G Z_i^\top$. This means that the conditional distribution of b_i given Y_i is multivariate normal with easily calculated $\mathrm{E}[b_i|Y_i]$ and $\mathrm{cov}(b_i|Y_i)$ (Exercise).

In developing predictors of b, most of the derivations are done assuming that G and R (and hence Σ) are known, but β is unknwon. At the very end, G and R are replaced by their estimates to produce what is called the *empirical* estimates and predictors.

First, the BLUE of β is the usual WLS estimator,

$$\hat{\beta} = (X^\top \Sigma^{-1} X)^{-1} X^\top \Sigma^{-1} Y.$$

This also defines $\hat{\mu} = X\hat{\beta}$.

Second, in

$$\mathrm{E}[b_i|Y_i] = G Z_i^\top \Sigma_i^{-1}(Y_i - \mu_i),$$

μ_i is replaced by its estimate to produce

$$\hat{b}_i = G Z_i^\top \Sigma_i^{-1}(Y_i - \hat{\mu}_i)$$

which is the best linear unbiased predictor (BLUP) of b_i based on Y_i. BLUP means that among all unbiased predictors of b_i that are linear in Y, it is the one with the smallest variance (and mean squared error).

Third, G and R (hence Σ) are replaced by their estimates to produce

$$\tilde{b}_i = \hat{G} Z_i^\top \hat{\Sigma}_i^{-1}(Y_i - \hat{\mu}_i)$$

which is known as the *empirical* BLUP, or EBLUP. Note that the EBLUP is neither linear nor unbiased.

The prediction error of the BLUP is $\hat{b}_i - b_i$. Its mean is zero, hence the prediction mean-squared error is its variance,

$$\mathrm{var}(\hat{b}_i - b_i) = G - Q_i^\top \left\{\Sigma_i - \mathrm{cov}(\hat{\mu}_i)\right\} Q_i,$$

where $Q_i := \Sigma_i^{-1} Z_i G$. This is the "prediction variance" (the "prediction standard error" is its square root) that is produced by most software (of course, with all unknown parameters replaced by estimates).

Q: Why not fit a model separately to each subject's data to estimate β_i and predict b_i, then take the average value of $\{\beta_i\}$ as an estimate of β?

Q: Why not fit a model using all data on all subjects, i.e. use WLS to regress Y on $[X, Z]$ with weight matrix R^{-1}? This should allow us to estimate β and predict b in one step?

Files: `tlc*, fat*`

Hypothesis testing for variance components:
Suppose that we have a random-effects model with $q+1$ random effects (per subject), and consider testing $H_0 : g_{q+1,q+1} = 0$ aganist $H_1 : g_{q+1,q+1} \neq 0$. Hypothesis H_0 represents a model with q random effects while H_1 represents the same model but with just *one* additional random effect. The likelihood ratio test can be used, but because H_0 is *on the boundary of the parameter space*, the "usual" theory does not hold. Under H_0, the large-sample distribution of the likelihood ratio test statistic is a 50:50 mixture of χ^2_q and χ^2_{q+1}. So, to compute the p-value, obtain a p-value from the χ^2_q distribution and another from the χ^2_{q+1} distribution, then take their average. If $q = 0$, take the p-value from the χ^2_1 distribution and divide it by 2 (i.e. the p-value from the χ^2_0 distribution is zero.)

For nested models that differ by more than one random effect (per subject), there is no simple answer. There is some theory, but the answers are quite complicated. See Li & cui (2016), Journal of Statistical Planning and Inference 178: 70-83.

Notes on SAS proc mixed:
The *model* statement describes X. The *repeated* statement describes the "within" part of the model, R. The *random* statement describes the "between" part of the model, Z and G.

If there is no repeated statement, $R_i = \sigma^2 I$ is assumed (i.e. constant variance and conditional independence within-subject).

In "repeated", a number of structures for R_i is available via the *type=* option; see the table in the manual.
Examples:
UN: unstructured
UN(1): diagonal, no structure otherwise (variance not constant over time)
UN(2): diagonal plus one band (above and below), no structure otherwise
UN(m): diagonal plus $m-1$ bands (above and below), no structure otherwise
CS: exchangeable correlation matrix with constant variance over time
CSH: exchangeable correlation matrix with varying variance over time
AR(1): first-order auto-regressive correlation structure with constant variance over time
ARH(1): first-order auto-regressive correlation structure with varying variance over time

If there is no random statement, $q = 0$ is assumed, i.e. no random effects. In this class, when there are random effects, $q \geq 1$, we will only use *type=un* in the random statement, i.e. G

is unstructured. The reason is that we should be able to reparametrize the model using any arbitrary full rank $q \times q$ matrix A to transform Z_i and b_i, i.e. write $Z_i b_i = (Z_i A^{-1})(A b_i) = (Z_i^*)(b_i^*)$. Now, $G^* = \text{var}(b_i^*) = \text{var}(A b_i) = A G A^\top$. Any special structure in G will almost certainly be destroyed and G^* will not have the same structure (unless A is trivial, such as $A = I$).

Q: The columns of Z_i are usually a subset of the columns of X_i. Why?