A short introduction to Linear Mixed Models

Xiang Ao

October 5, 2007

1 Linear Mixed Models

1.1 Random Intercept Model

Let's start with an example. Students learn in classes; classes are taught within school; schools are administered within school districts. A typical multilevel model (or mixed model; or sometimes called hierarchical models) would assign students to level 1, class level 2, school level 3 and district level 4. Ignoring the clustering structure of your data set will result in biased or inconsistent estimation of the coefficients and often downward-biased standard error estimation.

Let's first consider a two-level situation: students at level 1 and schools at level 2. Suppose our model for students' test scores is:

$$y_{ij} = \alpha_j + r_{ij}$$

where

$$r_{ij} \sim N(0, \sigma^2)$$

Here i is student; j is school index. We express a student's score in terms of the sum of an intercept for school (α_j) and a random error r_{ij} associated with student i at school j.

Now school level intercepts can be expressed as the sum of an overall mean μ and random deviations from that mean (u_j):

$$\alpha_j = \mu + u_j$$

where

$$u_j \sim N(0, \sigma_u^2).$$

Putting the two equations together we have a multilevel model:

$$y_{ij} = \mu + u_j + r_{ij}$$

where

$$r_{ij} \sim N(0, \sigma^2)$$

and

$$u_j \sim N(0, \sigma_u^2).$$

We can see that without the extra term u_j in the model, we basically have a simple mean model; or say we have a linear regression model with only a constant term as the regressor. By introducing u_i into the model, we are able to take account of the school effect in this case. That is, we model not only the grand mean, but also the mean within school. Also, we have the benefit of decomposing the total variance into two components: σ^2 and σ_u^2. This way we can explain how much of the variation comes from variation between schools and how much comes from variation within schools (between students). To be able to do these, we sacrifice by making an extra distribution assumption $u_j \sim N(0, \sigma_u^2)$.

Let's consider another example. Suppose we have patent prior art data. We have the proportion of prior art by examiners as the dependent variables. We have three levels: patent applicants, examiners, and technology. These three levels are crossed; they are not nested as students, schools, districts. Our model would be:

$$y_{ij} = \mu + u_j + v_k + w_l + r_{ijkl}$$

where

$$r_{ijkl} \sim N(0, \sigma^2)$$

and

$$u_j \sim N(0, \sigma_u^2),$$
$$v_k \sim N(0, \sigma_v^2),$$
$$w_l \sim N(0, \sigma_w^2).$$

1.2 Random Intercept and Random Slopes Model

We discussed about about the random intercept models; that is models which only allows the intercept to vary across groups. Now let's introduce the model with slopes varying across groups (levels).

Suppose we have predictors like student's family income in our test scores example:

$$y_{ij} = \alpha_j + \beta_j x_{ij} + r_{ij}$$
$$r_{ij} \sim N(0, \sigma^2)$$

$$\alpha_j = \mu + u_j$$
$$\beta_j = \eta + v_j$$

where

$$\begin{bmatrix} u_j \\ v_j \end{bmatrix} \sim N(\mathbf{0}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} \\ \sigma_{uv} & \sigma_v^2 \end{bmatrix})$$

In this model, we allow both the intercept and slopes of predictors vary across groups (levels). We make the assumption that the random components

of the intercept and slopes are jointly distributed as multivariate normal. σ_u^2 is the variance of level 2 residuals u_j from predicting level 1 intercept, σ_v^2 is the variance of level 2 residuals v_j from predicting level 1 slope. σ_{uv} is the covariance between u_j and v_j.

1.3 Variance Decomposition: ANOVA vs Mixed Models

Let's consider a simple model of intercept-only two-level model. One-way ANOVA and mixed model can be represented by the same model:

$$y_{ij} = \mu + u_j + r_{ij}$$

The difference in these two methods is: ANOVA is basically putting in group (level) dummies in the regression and mixed model is putting group-varying components into the error term.

ANOVA calculates the intraclass correlation by calculating between-group mean square error and within-group mean square error. Intraclass correlation is ratio of between-group variance and total variance (the sum of between-group variance and within-group variance). A linear mixed model treat u_i as part of the variance of the model; then it estimates the variance-covariance matrix.

For a model with only one group variable, the difference between variance decomposition by ANOVA and mixed model may not be huge. However, for a model with multiple group variables, such as

$$y_{ij} = \mu + u_j + v_k + w_l + r_{ijkl}.$$

If the three group variables are not orthogonal, then the variance decomposition by ANOVA depends on the order they are put in the model, while in a mixed model, the order is totally irrelevant.

1.4 General Linear Mixed Models

In general, a linear mixed model (LMM) is defined by

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{e}$$

where

$$e \sim N(0, R)$$

$$\gamma \sim N(0, G)$$

, and e and γ are uncorrelated.

This model has two parts: $\mathbf{X}\beta$ is the fixed effect part, and $\mathbf{Z}\gamma$ is the random effect part.

In this model, $E(\mathbf{y}) = \mathbf{X}\beta$ and $\text{Var}(\mathbf{y}) = \text{Var}(\mathbf{Z}\gamma) + \text{Var(e)} = \mathbf{ZGZ}' + \mathbf{R}$.

This means that in the linear mixed models, the unconditional mean is determined by the fixed effect part only.

The other way to formulate this is:

$$\mathbf{y}|\gamma \sim \mathbf{N}(\mathbf{X}\beta + \mathbf{Z}\gamma, \mathbf{R})$$

where

$$\gamma \sim N(0, G)$$

.

2 Generalized Linear Mixed Models

Generalized Linear Mixed Models (GLMM) are used to model non-normal data or normal data with correlations or heteroskadasticities. Generalized Linear Models (GLM) deal with data with distributions that belong to exponential family, such as Logit, Poisson. GLMM is an expansion of GLM to incorporate normally distributed random effects.

We know GLM usually models the mean of the distribution: $g(\mathrm{E}[y_i])) = x_i'\beta$; that is, some function $g()$ of the mean of the distribution is a linear function of x_i's. This function $g()$ is called link function. In the case of logit, it is the logit function. In the case of Poisson model, it is the log function.

Now the difference between GLM and GLMM is that GLMM has an extra term in it:

$$g(\mathrm{E}[y_i \mid \gamma])) = x_i'\beta + z'\gamma$$

where γ is normally distributed with mean 0 and variance G.

In Linear Mixed Models, we have $\mathrm{E}[y_i \mid \gamma] = \mathrm{E}[y_i]$. With a linear model, $\mathrm{E}[y_i \mid \gamma] = x_i'\beta + z'\mathrm{E}[\gamma] = x_i'\beta$ since $\gamma \sim N(0, G)$. With a non-linear model like GLMM,

$$\mathrm{E}[y_i \mid \gamma]) = g^{-1}(x_i'\beta + z'\gamma),$$

where $z'\gamma$ remains since g^{-1} is usually a non-linear function.

The other way to formulate this is by two steps:

First, the fixed and random effects are combined to form a linear predictor

$$\eta = \mathbf{X}\beta + \mathbf{Z}\gamma.$$

Second, if we have a linear mixed model, then we simply need to add in a normally distributed error term:

$$\mathbf{y} = \eta + \mathbf{e} = \mathbf{X}\beta + \mathbf{Z}\gamma + \mathbf{e}.$$

Equivalently, we have

$$\mathbf{y}|\gamma \sim \mathbf{N}(\mathbf{X}\beta + \mathbf{Z}\gamma, \mathbf{R})$$

where

$$\gamma \sim N(0, G)$$

.

Now if we don't have $\mathbf{y}$ normally distributed, then we have to "link" the independent variable to the linear predictor η, therefore the name "link function".

$$g(\mathbf{y}) = \eta$$

We have

$$\mathbf{y}|\gamma \sim \mathbf{N}(\mathbf{g^{-1}}(\eta), \mathbf{R}).$$

It says that the conditional distribution of $\mathbf{y}$ given η is a normal distribution, with mean as a function of the linear combination of fixed and random effects.

For example, say we have a log link (for example, in a Poisson model we have a log link), then

$$\mathrm{E}[\mathbf{y}] = \mathbf{E}[\exp(\mathbf{X}'\beta + \mathbf{Z}'\gamma)] = \exp(\mathbf{X}'\beta)\ \mathbf{E}[\exp(\mathbf{Z}'\gamma)],$$

which means the unconditional expectation of $\mathbf{y}$ is not only a function of fixed effects, but also random effects.

3 Estimate the models

In general, GLMM is estimated in the maximum likelihood framework,

$$L = \int \prod_i f_{y_i|\gamma}(y_i|\gamma) f_\gamma(\gamma) d\gamma,$$

Basically, this says that we need to integrate out the random effect part to get the unconditional likelihood function. Unfortunately usually we don't have an analytic solution for this, since random effects can be high-dimensional. Three possible solutions are proposed:

- Penalized Quasi-Likelihood. SAS Glimmix uses this approach. This approach uses a second order approximation of the log density, based on Taylor expansion. Essentially it determines the marginal log likelihood as that of an approximate linear mixed models. This approach lacks of a "true" log likelihood, but computationally feasible in many occasions.

- Integration by Gauss-Hermite Quadrature. This is essentially doing numerical integration to estimate the true likelihood function. This approach has the highest precision but often not feasible if there are many random effects, since it's very computationally intensive. SAS NLMIXED, Stata's xtnlmixed uses this approach.

- Laplace approximation. This approach is relatively fast, but not as accurate. R's lme4 package uses it currently for GLMM's.