Introduction to Mixed Models

Before introducing a mixed model, we will consider a model that we have previously seen in ANOVA:

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \text{ for } i = 1, ..., k \text{ and } j = 1, ..., n_i,$$

where the μ_i terms are group means and we have the usual error assumptions that the ε_{ij} are independent and $\varepsilon_{ij} \sim N(0, \sigma^2)$. For example we might be studying the high school calculus test scores at five different schools. If we are only interested in these five schools, then the above **fixed-effects model** is appropriate.

On the other hand, if these five schools represent a random sample of schools from a population of schools, then schools is a **random effect** and our model may be written as:

$$Y_{ij} = \mu + m_i + \varepsilon_{ij}, \text{ for } i = 1, ..., k \text{ and } j = 1, ..., n_i.$$

Here μ is the overall mean and the m_i are random variables describing the deviation from the mean. We assume that the m_i and ε_{ij} are independent random variables with $m_i \sim N(0, \sigma_m^2)$ and $\varepsilon_{ij} \sim N(0, \sigma^2)$. For the calculus example above, μ is the mean of all of the test scores and m_i is the random variable giving the deviation from the mean for the i^{th} high school. Models that incorporate both fixed and random effect terms are called **mixed models**. The model above is one simple example of a mixed model, but other well-known mixed model examples include **random coefficient models** and **repeated measures models**. In the model above the only parameters to be estimated are μ, σ_m^2, and σ^2. The m_i values are not parameters but realizations of a random variable, however it is often of interest to give predicted values for these terms.

The Matrix Formulation of the General Linear Mixed Model

The fixed-effects model for the sample is:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon,$$

where the errors ε satisfy:

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

The mixed model for the sample can be written as:

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

where $\mathbf{Z}$ is a design matrix for the random effects and $\mathbf{u}$ is the vector of random effects. The random vectors $\mathbf{u}$ and ε are assumed to be independent, and satisfy:

$$\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}) \text{ and } \varepsilon \sim N(\mathbf{0}, \mathbf{R}).$$

Note that the covariance matrix of ε for the general mixed model is a generalization of the covariance matrix for the error vector for the standard fixed effects model above. The mixed model can be alternately expressed in terms of either the conditional distribution of $\mathbf{y}$ given $\mathbf{u}$:

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

or in terms of the marginal distribution of $\mathbf{y}$:

$$\mathbf{y} \sim N(\mathbf{X}\beta, \mathbf{V}) \text{ where } \mathbf{V} = var(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}.$$

Estimation for the General Linear Mixed Model

Estimation is typically done in two steps: first for the covariance parameters, then secondly for the fixed-effect parameters and random-effect predictions using the estimated covariance parameters. The two most popular methods for estimation of the covariance parameters are maximum-likelihood and restricted maximum-likelihood (REML). Maximum likelihood estimators of the covariance parameters tend to be biased (as we have seen for σ^2 in the general linear model) with the bias becoming worse for increasing numbers of covariance parameters and for small samples. The REML method was developed to address this bias problem. One way to understand the bias of ML estimators of covariance parameters is to attribute the bias to a failure to adjust for estimated fixed-effect parameters. Suppose that all covariance parameters are contained in the vector θ, and that β are the fixed effects. The likelihood function for the data will be written as $L(\beta, \theta)$. The REML likelihood function is then defined as:

$$L_R(\theta) = \int L(\beta, \theta) \, d\beta.$$

More specifically we have

$$L_R(\theta) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{V}|}} \int \exp[-(\mathbf{y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)/2]\ d\beta,$$

where the matrix $\mathbf{V}$ is a function of the covariance parameters θ. Wood (2006) shows that by adding $\mathbf{X}\widehat{\beta} - \mathbf{X}\widehat{\beta}$ (where $\widehat{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$) to each $(\mathbf{y} - \mathbf{X}\beta)$ term then expanding and simplifying leads to

$$L_R(\theta) = \frac{\exp[-(\mathbf{y} - \mathbf{X}\widehat{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\beta})/2]}{\sqrt{(2\pi)^n|\mathbf{V}|}} \sqrt{\frac{(2\pi)^p}{|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}|}},$$

yielding a log REML likelihood function of:

$$l_R(\theta) = -\frac{n-p}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\widehat{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\widehat{\beta}).$$

Once θ has been estimated, then the fixed-effects parameters are estimated as:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{V}(\widehat{\theta})^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}(\widehat{\theta})^{-1}\mathbf{y}.$$

The random-effects terms $\mathbf{u}$ can be predicted by:

$$\widehat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}(\widehat{\theta})^{-1}(\mathbf{y} - \mathbf{X}\widehat{\beta})$$

Inference for Parameters in the General Linear Mixed Model

Likelihood-ratio tests can be conducted for nested mixed models, although when using REML methods the models must share the same set of fixed-effect parameters. General linear hypotheses can be tested in a manner analogous to that for the (fixed-effect) general linear model, but the resulting test statistics may be only have approximate t or F distributions.

References

Littell, R.C., Milliken, G.A., Stroup, W.W., Wolfinger, R.D., and Schabenberger, O. 2006. SAS for Mixed Models, Second Edition, Cary, NC: SAS Institute Inc.

Pinheiro, J.C. and Bates, D.M. 2000. Mixed-Effects Models in S and S-PLUS, New York: Springer.

Wood, Simon. 2006. Generalized Additive Models: An Introduction with R, Boca Raton: Chapman & Hall/CRC.