3. Linear Mixed Models

Specification of Linear Mixed Models

Using the hierarchical notation of Laird and Ware (1982), we can express the linear mixed model as

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \tag{1}$$

where $Y_i = (n_i \times 1)$ response vector, $X_i = (n_i \times p)$ design matrix for fixed effects, $\beta = (p \times 1)$ regression coefficients for fixed effects, $Z_i = (n_i \times q)$ design matrix for random effects, $b_i = (q \times 1)$ random effects, and $\epsilon_i = (n_i \times 1)$ error vector for $i = 1, \cdots, m$.

Distributional assumptions: b_i and ϵ_i are independent with

$$b_i \overset{iid}{\sim} N(0, D),$$
$$\epsilon_i \overset{iid}{\sim} N(0, \sigma^2 I).$$

Note that D is a $q \times q$ matrix that does not depend on i. Under these assumptions, Y_i has a multivariate

normal distribution:

$$Y_i \sim N\left(X_i\beta, V(\alpha)\right) \tag{2}$$

where $V(\alpha) = Z_i D Z_i^T + \sigma^2 I$ and α denotes the variance component parameters.

- D must be symmetric and positive definite (all its eigenvalues are strictly positive).

- Suppose that we have one covariate and $X_i = Z_i = (1, X_i)$ (random intercept and random slope model), then we can write

$$\begin{aligned} & Y_i = X_i(\beta + b_i) + \epsilon_i, \\ \text{or} \qquad & Y_i = X_i\beta_i + \epsilon_i \end{aligned}$$

  where $\beta_i \sim N(\beta, D)$ and $D = \begin{pmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{pmatrix}$, with $d_{00} = var(\beta_{i0})$, $d_{11} = var(\beta_{i1})$, and $d_{01} = d_{10} = cov(\beta_{i0}, \beta_{i1})$.

- The error ϵ are assumed to be iid normally distributed with variance σ^2. This assumption will be relaxed later.

- The columns of matrix Z_i are typically a subset of the columns in X_i. In particular, $Z_i = 1_i$ corresponds to the random intercept model.

- The marginal mean for Y is the same as in the marginal general linear model:

$$\begin{aligned} E(Y_i) &= E(E(Y_i|b_i)) = E(X_i\beta + Z_i b_i) \\ &= X_i\beta. \end{aligned}$$

  Thus, the interpretation of the regression coefficients β is the same. This will no longer hold for nonlinear models where $E(Y_i|b_i)$ is not a linear function of β and b_i.

Multilevel Mixed Effects Models

The hierarchical specification of linear mixed model can be easily extended to multiple nested levels, to accommodate, for example, longitudinal measurements from subjects in the same clinical center, family, or community. Let $k = 1, \cdots, K$ be index for the group and $i = 1, \cdots, m_k$ be for individuals in group k, $j = 1, \cdots, n_i$ for observational times for individuals i. The model can be written as

$$Y_{ki} = X_{ki}\beta + Z_{ki}b_k + Z_{ki}b_{ki} + \epsilon_{ki}, \tag{3}$$

where $b_k \sim N(0, D_1)$, $b_{ki} \sim N(0, D_2)$, and $\epsilon_{ki} \sim N(0, \sigma^2 I)$.

- The first-level random effects b_k of length q_1 are assumed to be independent for different k (groups).
- The second-level random effects b_{ki} of length q_2 are assumed to be independent for different k (groups) or i (individuals), and of the b_k.
- ϵ_{ki} are independent for k, i, and the random effects.

- Some people would call this a three-level model as in

$$
\begin{aligned}
Y_i &\sim N(X_i\beta_i, \sigma^2 I), \\
\beta_i &\sim N(Z_{i1}\gamma_{i1}, \tau_1^2 I), \\
\gamma_{i1} &\sim N(Z_{i2}\gamma_{i2}, \tau_2^2 I).
\end{aligned}
$$

Estimation for LME

- In the original paper, an EM (expectation-maximization) algorithm was used for estimation. Nowadays the estimation is often done based on the marginal model using numeric optimization to maximize the likelihood (ML) or restricted likelihood (REML).

- In general linear model specification, we only require $V(\alpha)$ to be symmetric and positive definite. So often the maximization is one in this slightly larger parameter space. Not very well specified (overspecified) models can result in instability in the estimates of the variance parameter.

- In particular, lme() uses a mixed EM and Newton-Raphson iterations. Whereas SAS PROC MIXED uses Newton-Raphson. For small data sets with large models, it may be wise to monitor the convergence and change some optimization parameters when necessary. Better still, avoid fitting over-elaborated models.

Inference for Fixed Effects in LME

- For a contrast matrix L, to test the null hypothesis $H_0 : L\beta = 0$ vs $H_0 : L\beta \neq 0$, the Wald statistic is

$$(\hat{\beta} - \beta)^T L^T \left\{ L \left(\sum_i X_i^T V_i(\hat{\alpha})^{-1} X_i \right) L^T \right\}^{-1} L(\hat{\beta} - \beta) \overset{Approx.}{\sim} \chi^2_{df},$$

  where $df = rank(L)$.

- Wald test tends to be anti-conservative. For small samples, the F test statistic is used. Assume $V(\alpha) = \sigma^2 W(\alpha)^{-1}$. Then

$$\begin{aligned} F &= \frac{(\hat{\beta} - \beta)^T L^T \left\{ L \left(\sum_i X_i^T \hat{V}_i(\hat{\alpha})^{-1} X_i \right)^{-1} L^T \right\}^{-1} L(\hat{\beta} - \beta)}{rank(L)} \\ &= \frac{(\hat{\beta} - \beta)^T L^T \left\{ L \left(\sum_i X_i^T W_i(\hat{\alpha}) X_i \right)^{-1} L^T \right\}^{-1} L(\hat{\beta} - \beta)}{rank(L)\hat{\sigma^2}} \\ &\overset{Approx.}{\sim} F_{df1, df2} \end{aligned}$$

  where $\hat{\sigma^2} = \frac{y^T [W - WX(X^T WX)^- X^T W] y}{m - rank(X)}$, $df_1 = rank(L)$ and df_2 =degree of freedom estimated from the data.

- Empirical variance estimates for $\hat{\beta}$ can also be used (PROC MIXED, not supported in lme()).

- The Wald test is conditional in that the parameter estimation is done once and the test for the certain coefficients is done conditional on the esimates for the other, nuisance parameters in the model.

- For nested models, likelihood ratio tests can also be used (there the estimation has to be done using ML instead of REML). But Pinheiro and Bates (2000) showed that the LRT tests tend to be anticonservative (p-value too small) for small samples.

- Confidence intervals for regression coefficients can be constructed using the t distribution.

- Methods based on multivariate $t-$ distribution and bootstrp are in development.

Inference for variance parameters in LME

- The MLEs for regression coefficients β and variance parameter α are asymptotically uncorrelated.

- Variance estimates for $\hat{\alpha}$ are computed using the inverse observed information matrix $(-\ddot{l})$.

- Confidence intervals for variance parameters can be tricky because it is often of interest to calculate CIs on the original parameters in (σ^2, D) and not the marginal parameters α.

- When the true parameter value is on the boundary of the parameter space (i.e., $\sigma = 0$), Wald test is not valid.

- Likelihood ratio test can be used to compare nested models with different variance parameters. Both ML and REML can be used.

- However, when one of the model sets some parameters at 0, which is at the boundary of the parameter space, the degrees of freedom for the LRT needs to be adjusted (Stram and Lee, 1994;

Self and Liang, 1987). The unadjusted LRT tends to be too conservative (p-value too large).

- A random intercept model with $var(b_i) = \tau^2$. Testing the null hypothesis that the random intercept is not needed is equivalent to $H_0 : \tau = 0$ vs $H_1 : \tau > 0$. The LRT statistic has a mixture distribution that puts 0.5 mass on 0 and 0.5 mass at χ^2_1, or $\sim 0.5\chi^2_0 + 0.5\chi^2_1$.
- A random intercept and slope model with

$$var(b_i) = \begin{pmatrix} \tau_0^2 & \rho\tau_0\tau_1 \\ \rho\tau_0\tau_1 & \tau_1^2 \end{pmatrix}.$$

  Under the null hypothesis $H_0 : \tau_1 = 0$, the correlation ρ degenerates. Therefore, the distribution of LRT statistic is a 1:1 mixture of χ^2_1 and χ^2_2.
- More generally, when comparing a model with $q+1$ (correlated) random effects with the model with q (correlated) random effects, the distribution of the LRT statistic is a 1:1 mixture of χ^2_q and χ^2_{q+1}.
- When comparing models with $q + k$ and q correlated random effects where $k > 1$, the

distribution of the LRT statistic is not well understood.

- There is no general rule to reliably come up with the distribution of LRT statistic. In nlme library, the naive approach (no adjustment) is implemented in the two-argument version of anova.
- In Fitzmaurice, Laird, and Ware (2004), they recommended using 0.1 significant level instead of 0.05 to compensate some of the conservativeness, that leads in overly simple models.

- Information criteria can be used to compare non-nested models. They are based on the likelihoods with a penalty term that is larger for models with larger number of parameters.

  - Let n_{par} denote the total number of parameters (fixed and random effects), and $N = \sum_{i=1}^{m} n_i$, then the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are defined

as

$$AIC = -2l(\hat{\theta}|y) + 2n_{par},$$
$$BIC = -2l(\hat{\theta}|y) + n_{par}\log(N).$$

- The REML version of AIC and BIC replace $l(\hat{\theta}|y)$ with $l_R(\hat{\theta}|y)$ and N with $N - p$ where p is the number of fixed effects parameters (Again, ML should be used to compare models with different fixed effects).
- Models with the smaller AIC or BIC are better.
- Information criteria are more flexible than likelihood ratio test but they only provide a "rule-of-thumb" and not a formal statistical significance test.
- Different criteria can lead to different conclusions.

Inference about the Random Effects

The random effects b_i are random variables, not parameters. Technically, we predict the random effects, not estimate them.

BLUP

- For arbitrary vectors s and t,
  the best linear unbiased predictors (BLUP) $\tilde{\beta}$ and $\tilde{b}$ minimize the prediction error

$$E\left\{(s^T X\tilde{\beta} + t^T Z\tilde{b}) - (s^T X\beta + t^T Zb)\right\}^2,$$

  subject to the unbiasedness condition

$$E(s^T X\tilde{\beta} + t^T Z\tilde{b}) = E(s^T X\beta + t^T Zb).$$

  It can be shown that the solutions are

$$BLUP(\beta) = \tilde{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y.$$
$$BLUP(b) = \tilde{b} = DZ^T V^{-1}(y - X\tilde{\beta}),$$

  where $V = ZDZ^T + \sigma^2 I$.

  NOTE

  – $\tilde{\beta}$ is identical to the general least squares estimator which is also the best linear unbiased estimator (BLUE).

– The BLUP for b is the best linear predictor (BLP)

$$BLP(b) = DZ^T V^{-1}(y - X\beta)$$

with β replaced with $\hat{\beta}$.

– D and V have to be estimated. i.e., via ML or REML. In practice, the BLUPs are replaced by the estimated BLUPs or EBLUPs,

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y,$$
$$\hat{b} = \hat{D} Z^T \hat{V}^{-1}(y - X\hat{\beta}).$$

- One simple (ad-hoc) justification of BLUP is due to Henderson et al. (1959) that involves making the distributional assumption

$$Y|b \sim N(X\beta + Zb, R),$$
$$b \sim N(0, D),$$

and maximize the likelihood of the (y, b) over the unknown β and b. This leads to the criterion

$$(y - X\beta - Zb)^T R^{-1}(y - X\beta - Zb) + b^T D^{-1} b.$$

This shows that BLUP estimation of (β, b) involves general least squares with a penalty term. Hence, it is related to ridge regression.

Empirical Bayes and Shrinkage

From a Bayesian perspective, the posterior distribution of b given the data y is (dependence on the parameters θ is suppressed)

$$f(b|y) = \frac{f(y|b)f(b)}{\int f(y|b)f(b)db}$$

which can be shown to be a multivariate normal distribution. Thus, b can be estimated using the posterior mean

$$\begin{aligned} \hat{b}(\theta) &= E(b|y) \\ &= \int bf(b|y)db \\ &= DZ^TV^{-1}(y - X\beta). \end{aligned}$$

In practice, unknown parameters β and α are replaced by their ML or REML estimates. The resulting estimates for the random effects are called empirical Bayes (EB) estimates.

Consider the prediction of the response for i

$$\begin{aligned}
\hat{Y}_i &= X_i\hat{\beta} + Z_i\hat{b}_i \\
&= X_i\hat{\beta} + Z_i D Z_i^T V_i^{-1}(y_i - X_i\hat{\beta}) \\
&= (I_{n_i} - Z_i D Z_i^T V_i^{-1})X_i\hat{\beta} + Z_i D Z_i^T V_i^{-1} y_i \\
&= \Sigma_i V_i^{-1} X_i\hat{\beta} + (I_{n_i} - \Sigma_i V_i^{-1}) y_i,
\end{aligned}$$

where the residual variance is $\Sigma_i = V_i - Z_i D Z_i^T$.

Note:

- It can be interpreted as a weighted average of the population mean $X\hat{\beta}$ and the observed data y_i.

- Larger weights ar given in the overall mean if the residual variability Σ_i is large in comparison with the between-subject variability $Z_i D Z_I^T$.

- In Bayesian literature (Carlin and Louis, 1996) it is referred to as shrinkage. The observed data are shrunk toward the prior mean which is $X_i\beta$ since the prior mean of the random effects was zero.

Normality Assumption

- The random effects are assumed to be normally distributed. When that assumption is violated, the inference about marginal model and especially the fixed effects are still valid.

- However, the EB estimates of the random effects may be highly affected by their distributional assumption. In particular, heterogeneity in the population random effect b may not be preserved in the shrunk $\hat{b}$.

- Checking the normality assumption in tricky. In particular, histograms of the $\hat{b}_i$ are not useful because the $\hat{b}_i$ are not identically distributed and they have smaller variance than the population b_i because of the shrinkage.

Extending Linear Mixed Models

- The covariance matrix for the random effects D is generally assumed to be simply a symmetric positive

definite matrix. Sometime it is desirable to use other type of matrices such as an identity matrix, a diagonal matrix, etc. In particular, the two-level random effects model can be written as a one-level random effects with block compound symmetry covariance matrix.

- The covariance matrix for the errors can be made generally

$$Y_i|b_i \sim N(X_i\beta + Zb_i, \sigma^2\Lambda_i)$$

where Λ_i are positive-definite matrices parameterized by a fixed λ parameter.

- DHLZ further decomposed the error variance $\sigma^2\Lambda_i$ into a serial correlation and a measurement error components, the latter being $\tau^2 I$.

- Pinheiro and Bates (2000) decomposed

$$\Lambda_i = B_i C_i B_i$$

where B_i is diagonal and C_i is a correlation matrix. To ensure uniqueness, all elements in B_i are required to be positive. Thus,

$$var(\epsilon_{ij}) = \sigma^2[B_i]^2_{jj},$$
$$corr(\epsilon_{ij}, \epsilon_{ik}) = [C_i]_{jk}.$$

This decomposed Λ_i into a variance structure component and a correlation structure component.

Correlation Structure

- Compound Symmetry

$$C_i = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}_{n_i \times n_i}$$

- Banded (width 1)

$$C_i = \begin{pmatrix} 1 & \rho & 0 & \cdots & 0 \\ \rho & 1 & \rho & \cdots & 0 \\ 0 & \rho & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}_{n_i \times n_i}$$

- Exponential (a special case of discrete time AR(1) model) and Gaussian correlated models.

- Autoregressive-moving average (ARMA) models (time series).

- Spatial correlation (eg., CAR).

Variance Structure

A general variance function model is

$$var(\epsilon_{ij}|b_i) = \sigma^2 g^2(\mu_{ij}, v_{ij}, \delta)$$

where $\mu_{ij} = E(y_{ij}|b_i)$, v_{ij} is a vector of covariates, δ is a vector of variance parameters and g is the variance function. For example,

$$var(\epsilon_{ij}|b_i) = \sigma^2|v_{ij}|^{2\delta}.$$

- In practice μ_{ij} is replaced by $\hat{\mu}_{ij}$. The estimation is done by iterating the following steps until convergence.
  - given $\beta^{(t)}$ and $\delta^{(t)}$, estimate $\mu_{ij}^{(t)}$.
  - given $\mu_{ij}^{(t)}$, estimate $\beta^{(t+1)}$ and $\delta^{(t+1)}$.