

Chapter 1

Linear Mixed effects Models

1.1 Linear Mixed effects Models

A linear mixed effects (LME) model is a statistical model containing both fixed effects and random effects (random effects are also known as variance components). LME models are a generalization of the classical linear model, which contain fixed effects only. When the levels of factors are considered to be sampled from a population, and each level is not of particular interest, they are considered random quantities with associated variances. The effects of the levels, as described, are known as random effects. Random effects are represented by unobservable normally distributed random variables. Conversely fixed effects are considered non-random and the levels of each factor are of specific interest.

? introduced variance components models for use in genetical studies. Whereas an estimate for variance must take a non-negative value, an individual variance component, i.e. a component of the overall variance, may be negative.

The methodology has developed since, including contributions from ?, who extend the use of variance components into linear models, and ?, who introduced the ‘mixed model’ terminology and formally distinguished between mixed and random effects models. ? devised a methodology for deriving estimates for both the fixed effects and the

random effects, using a set of equations that would become known as ‘mixed model equations’ or ‘Henderson’s equations’. LME methodology is further enhanced by Henderson’s later works (1975). The key features of Henderson’s work provide the basis for the estimation techniques.

Wallerstein demonstrated that unique estimates of the variance components could be obtained using maximum likelihood methods. However these estimates are known to be biased ‘downwards’ (i.e. underestimated) , because of the assumption that the fixed estimates are known, rather than being estimated from the data. Wallerstein produced an alternative set of estimates, known as the restricted maximum likelihood (REML) estimates, that do not require the fixed effects to be known. Thusly there is a distinction the REML estimates and the original estimates, now commonly referred to as ML estimates.

Wallerstein provides a form of notation for notation for LME models that has since become the standard form, or the basis for more complex formulations. Due to computation complexity, linear mixed effects models have not seen widespread use until many well known statistical software applications began facilitating them. SAS Institute added PROC MIXED to its software suite in 1992 (1992). Wallerstein described how to compute LME models in the S-plus environment.

Using Laird-Ware form, the LME model is commonly described in matrix form,

$$y = X\beta + Zb + \epsilon \tag{1.1}$$

where y is a vector of N observable random variables, β is a vector of p fixed effects, X and Z are $N \times p$ and $N \times q$ known matrices, and b and ϵ are vectors of q and N , respectively, random effects such that $E(b) = 0$, $E(\epsilon) = 0$ and where D and Σ are positive definite matrices parameterized by an unknown variance component parameter vector θ . The variance-covariance matrix for the vector of observations y is given by $V = ZDZ' + \Sigma$. This implies $y \sim (X\beta, V) = (X\beta, ZDZ' + \Sigma)$. It is worth noting that V is an $n \times n$ matrix, as the dimensionality becomes relevant later on. The notation provided here is generic, and will be adapted to accord with complex formulations that will be encountered in due course.

1.1.1 Estimation

Estimation of LME models involve two complementary estimation issues'; estimating the vectors of the fixed and random effects estimates $\hat{\beta}$ and \hat{b} and estimating the variance covariance matrices D and Σ . Inference about fixed effects have become known as 'estimates', while inferences about random effects have become known as 'predictions'. The most common approach to obtain estimators are Best Linear Unbiased Estimator (BLUE) and Best Linear Unbiased Predictor (BLUP). For an LME model given by (1.1), the BLUE of $\hat{\beta}$ is given by

$$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y,$$

whereas the BLUP of \hat{b} is given by

$$\hat{b} = DZ'V^{-1}(y - X\hat{\beta}).$$

Estimation of the fixed parameters

The vector y has marginal density $y \sim N(X\beta, V)$, where $V = \Sigma + ZDZ'$ is specified through the variance component parameters θ . The log-likelihood of the fixed parameters (β, θ) is

$$\ell(\beta, \theta | y) = -\frac{1}{2} \log |V| - \frac{1}{2} (y - X\beta)' V^{-1} (y - X\beta), \quad (1.2)$$

and for fixed θ the estimate $\hat{\beta}$ of β is obtained as the solution of

$$(X'V^{-1}X)\beta = X'V^{-1}y. \quad (1.3)$$

Substituting $\hat{\beta}$ from (1.3) into $\ell(\beta, \theta | y)$ from (1.2) returns the *profile* log-likelihood

$$\begin{aligned} \ell_P(\theta | y) &= \ell(\hat{\beta}, \theta | y) \\ &= -\frac{1}{2} \log |V| - \frac{1}{2} (y - X\hat{\beta})' V^{-1} (y - X\hat{\beta}) \end{aligned}$$

of the variance parameter θ . Estimates of the parameters θ specifying V can be found by maximizing $\ell_P(\theta | y)$ over θ . These are the ML estimates.

For REML estimation the *restricted* log-likelihood is defined as

$$\ell_R(\theta \mid y) = \ell_P(\theta \mid y) - \frac{1}{2} \log |X'VX|.$$

The REML approach does not base estimates on a maximum likelihood fit of all the information, but instead uses a likelihood function derived from a data set, transformed to remove the irrelevant influences (?). Restricted maximum likelihood is often preferred to maximum likelihood because REML estimation reduces the bias in the variance component by taking into account the loss of degrees of freedom that results from estimating the fixed effects in β . Restricted maximum likelihood also handles high correlations more effectively, and is less sensitive to outliers than maximum likelihood. The problem with REML for model building is that the likelihoods obtained for different fixed effects are not comparable. Hence it is not valid to compare models with different fixed effects using a likelihood ratio test or AIC when REML is used to estimate the model. Therefore models derived using ML must be used instead.

Estimation of the random effects

The established approach for estimating the random effects is to use the best linear predictor of b from y , which for a given β equals $DZ'V^{-1}(y - X\beta)$. In practice β is replaced by an estimator such as $\hat{\beta}$ from (1.3) so that $\hat{b} = DZ'V^{-1}(y - X\hat{\beta})$. Pre-multiplying by the appropriate matrices it is straightforward to show that these estimates $\hat{\beta}$ and \hat{b} satisfy the equations in (??).

Algorithms for likelihood function optimization

Iterative numerical techniques are used to optimize the log-likelihood function and estimate the covariance parameters θ . The procedure is subject to the constraint that R and D are both positive definite. The most common iterative algorithms for optimizing the likelihood function are the Newton-Raphson method, which is the preferred method, the expectation maximization (EM) algorithm and the Fisher scoring methods.

The EM algorithm, introduced by ?, is an iterative technique for maximizing complicated likelihood functions. The algorithm alternates between performing an expectation (E) step and the maximization (M) step. The ‘E’ step computes the expectation of the log-likelihood evaluated using the current estimate for the variables. In the ‘M’ step, parameters that maximize the expected log-likelihood, found on the previous ‘E’ step, are computed. These parameter estimates are then used to determine the distribution of the variables in the next ‘E’ step. The algorithm alternatives between these two steps until convergence is reached.

The main drawback of the EM algorithm is its slow rate of convergence. Consequently the EM algorithm is rarely used entirely in LME estimation, instead providing an initial set of values that can be passed to other optimization techniques.

The Newton Raphson (NR) method is the most common, and recommended technique for ML and REML estimation. The NR algorithm minimizes an objective function defines as -2 times the log likelihood for the covariance parameters θ . At every iteration the NR algorithm requires the calculation of a vector of partial derivatives, known as the gradient, and the second derivative matrix with respect to the covariance parameters. This is known as the observed Hessian matrix. Due to the Hessian matrix, the NR algorithm is more time-consuming, but convergence is reached with fewer iterations compared to the EM algorithm. The Fisher scoring algorithm is an variant of the NR algorithm that is more numerically stable and likely to converge, but not recommended to obtain final estimates.

1.1.2 Formulation of the response vector

Information of individual i is recorded in a response vector \mathbf{y}_i . The response vector is constructed by stacking the response of the 2 responses at the first instance, then the 2 responses at the second instance, and so on. Therefore the response vector is a $2n_i \times 1$ column vector. The covariance matrix of \mathbf{y}_i is a $2n_i \times 2n_i$ positive definite matrix $\mathbf{\Omega}_i$.

Consider the case where three measurements are taken by both methods A and B ,

\mathbf{y}_i is a 6×1 random vector describing the i th subject.

$$\mathbf{y}_i = (y_i^{A1}, y_i^{B1}, y_i^{A2}, y_i^{B2}, y_i^{A3}, y_i^{B3})'$$

The response vector \mathbf{y}_i can be formulated as an LME model according to Laird-Ware form.

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i$$

$$\mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$$

$$\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_i)$$

Information on the fixed effects are contained in a three dimensional vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$. For computational purposes β_2 is conventionally set to zero. Consequently $\boldsymbol{\beta}$ is the solutions of the means of the two methods, i.e. $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$. The variance covariance matrix \mathbf{D} is a general 2×2 matrix, while \mathbf{R}_i is a $2n_i \times 2n_i$ matrix.

1.1.3 Decomposition of the response covariance matrix

The variance covariance structure can be re-expressed in the following form,

$$\text{Cov}(\mathbf{y}_i) = \boldsymbol{\Omega}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i.$$

\mathbf{R}_i can be shown to be the Kronecker product of a correlation matrix \mathbf{V} and $\mathbf{\Lambda}$. The correlation matrix \mathbf{V} of the repeated measures on a given response variable is assumed to be the same for all response variables. Both ? and ? use the identity matrix, with dimensions $n_i \times n_i$ as the formulation for \mathbf{V} . ? remarks that, with repeated measures, the response for each subject is correlated for each variable, and that such correlation must be taken into account in order to produce a valid inference on correlation estimates. ? proposes various correlation structures may be assumed for repeated measure correlations, such as the compound symmetry and autoregressive structures, as alternative to the identity matrix.

However, for the purposes of method comparison studies, the necessary estimates are currently only determinable when the identity matrix is specified, and the results in ? indicate its use.

For the response vector described, ? presents a detailed covariance matrix. A brief summary shall be presented here only. The overall variance matrix is a 6×6 matrix composed of two types of 2×2 blocks. Each block represents one separate time of measurement.

$$\mathbf{\Omega}_i = \begin{pmatrix} \mathbf{\Sigma} & \mathbf{D} & \mathbf{D} \\ \mathbf{D} & \mathbf{\Sigma} & \mathbf{D} \\ \mathbf{D} & \mathbf{D} & \mathbf{\Sigma} \end{pmatrix}$$

The diagonal blocks are Σ , as described previously. The 2×2 block diagonal matrix in $\mathbf{\Omega}$ gives Σ . Σ is the sum of the between-subject variability \mathbf{D} and the within subject variability $\mathbf{\Lambda}$.

$\mathbf{\Omega}_i$ can be expressed as

$$\mathbf{\Omega}_i = \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + (\mathbf{I}_{n_i} \otimes \mathbf{\Lambda}).$$

The notation dim_{n_i} means an $n_i \times n_i$ diagonal block.