Mixed models in R using the lme4 package
Part 4: Theory of linear mixed models

Douglas Bates

8^{th} International Amsterdam Conference
on Multilevel Analysis
<Bates@R-project.org>

2011-03-16

Outline

Outline

Outline

Outline

Definition of linear mixed models

- As previously stated, we define a linear mixed model in terms of two random variables: the n-dimensional $\mathcal{Y}$ and the q-dimensional $\mathcal{B}$
- The probability model specifies the conditional distribution

$$(\mathcal{Y}|\mathcal{B}=\boldsymbol{b}) \sim \mathcal{N}\left(\boldsymbol{Z}\boldsymbol{b}+\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right)$$

  and the unconditional distribution

$$\mathcal{B} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right).$$

  These distributions depend on the parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ.
- The probability model defines the *likelihood* of the parameters, given the observed data, $\boldsymbol{y}$. In theory all we need to know is how to define the likelihood from the data so that we can maximize the likelihood with respect to the parameters. In practice we want to be able to evaluate it quickly and accurately.

Properties of Σ_θ; generating it

- Because it is a variance-covariance matrix, the $q \times q$ Σ_θ must be symmetric and *positive semi-definite*, which means, in effect, that it has a "square root" — there must be another matrix that, when multiplied by its transpose, gives Σ_θ.
- We never really form Σ; we always work with the *relative covariance factor*, Λ_θ, defined so that

$$\Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^\mathsf{T}$$

  where σ^2 is the same variance parameter as in $(\mathcal{Y}|\mathcal{B} = b)$.
- We also work with a q-dimensional "spherical" or "unit" random-effects vector, $\mathcal{U}$, such that

$$\mathcal{U} \sim \mathcal{N}\left(0, \sigma^2 I_q\right), \; \mathcal{B} = \Lambda_\theta \mathcal{U} \Rightarrow \mathsf{Var}(\mathcal{B}) = \sigma^2 \Lambda_\theta \Lambda_\theta^\mathsf{T} = \Sigma.$$

- The linear predictor expression becomes

$$Z b + X \beta = Z \Lambda_\theta u + X \beta$$

The conditional mean $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}$

- Although the probability model is defined from $(\mathcal{Y}|\mathcal{U}=\boldsymbol{u})$, we observe $\boldsymbol{y}$, not $\boldsymbol{u}$ (or $\boldsymbol{b}$) so we want to work with the other conditional distribution, $(\mathcal{U}|\mathcal{Y}=\boldsymbol{y})$.
- The joint distribution of $\mathcal{Y}$ and $\mathcal{U}$ is Gaussian with density

$$\begin{aligned} f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u}) &= f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}|\boldsymbol{u})\, f_{\mathcal{U}}(\boldsymbol{u}) \\ &= \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{n/2}}\,\frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{q/2}} \\ &= \frac{\exp(-\left[\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2+\|\boldsymbol{u}\|^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}} \end{aligned}$$

- $(\mathcal{U}|\mathcal{Y}=\boldsymbol{y})$ is also Gaussian so its mean is its mode. I.e.

$$\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}} = \arg\min_{\boldsymbol{u}} \left[\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2+\|\boldsymbol{u}\|^2\right]$$

Minimizing a penalized sum of squared residuals

- An expression like $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2+\|\boldsymbol{u}\|^2$ is called a *penalized sum of squared residuals* because $\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2$ is a sum of squared residuals and $\|\boldsymbol{u}\|^2$ is a penalty on the size of the vector $\boldsymbol{u}$.
- Determining $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}$ as the minimizer of this expression is a *penalized least squares* (PLS) problem. In this case it is a *penalized linear least squares problem* that we can solve directly (i.e. without iterating).
- One way to determine the solution is to rephrase it as a linear least squares problem for an extended residual vector

$$\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}=\arg\min_{\boldsymbol{u}}\left\|\begin{bmatrix}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}\end{bmatrix}-\begin{bmatrix}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta\\ \boldsymbol{I}_q\end{bmatrix}\boldsymbol{u}\right\|^2$$

This is sometimes called a *pseudo-data* approach because we create the effect of the penalty term, $\|\boldsymbol{u}\|^2$, by adding "pseudo-observations" to $\boldsymbol{y}$ and to the predictor.

Solving the linear PLS problem

- The conditional mean satisfies the equations

$$\left(\boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta^\mathsf{T}+\boldsymbol{I}_q\right)\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}=\boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}).$$

- This would be interesting but not very important were it not for the fact that we actually can solve that system for $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}$ even when its dimension, q, is very, very large.
- Because $\boldsymbol{Z}$ is generated from indicator columns for the grouping factors, it is sparse. $\boldsymbol{Z}\boldsymbol{\Lambda}_\theta$ is also very sparse.
- There are sophisticated and efficient ways of calculating a sparse Cholesky factor, which is a sparse, lower-triangular matrix $\boldsymbol{L}_\theta$ that satisfies

$$\boldsymbol{L}_\theta\boldsymbol{L}_\theta^\mathsf{T}=\boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta+\boldsymbol{I}_q$$

  and, from that, solving for $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}$.

The sparse Choleksy factor, $\boldsymbol{L}_\theta$

- Because the ability to evaluate the sparse Cholesky factor, $\boldsymbol{L}_\theta$, is the key to the computational methods in the `lme4` package, we consider this in detail.
- In practice we will evaluate $\boldsymbol{L}_\theta$ for many different values of $\boldsymbol{\theta}$ when determining the ML or REML estimates of the parameters.
- As described in Davis (2006), §4.6, the calculation is performed in two steps: in the *symbolic decomposition* we determine the position of the nonzeros in $\boldsymbol{L}$ from those in $\boldsymbol{Z\Lambda}_\theta$ then, in the *numeric decomposition*, we determine the numerical values in those positions. Although the numeric decomposition may be done dozens, perhaps hundreds of times as we iterate on $\boldsymbol{\theta}$, the symbolic decomposition is only done once.

A fill-reducing permutation, $\boldsymbol{P}$

- In practice it can be important while performing the symbolic decomposition to determine a *fill-reducing permutation*, which is written as a $q \times q$ permutation matrix, $\boldsymbol{P}$. This matrix is just a re-ordering of the columns of $\boldsymbol{I}_q$ and has an orthogonality property, $\boldsymbol{P}\boldsymbol{P}^{\mathsf{T}} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{P} = \boldsymbol{I}_q$.
- When $\boldsymbol{P}$ is used, the factor $\boldsymbol{L}_\theta$ is defined to be the sparse, lower-triangular matrix that satisfies

$$\boldsymbol{L}_\theta \boldsymbol{L}_\theta^{\mathsf{T}} = \boldsymbol{P}\left[\boldsymbol{\Lambda}_\theta^{\mathsf{T}} \boldsymbol{Z}_\theta^{\mathsf{T}} \boldsymbol{Z} \boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q\right] \boldsymbol{P}^{\mathsf{T}}$$

- In the `Matrix` package for *R*, the `Cholesky` method for a sparse, symmetric matrix (class `dsCMatrix`) performs both the symbolic and numeric decomposition. By default, it determines a fill-reducing permutation, $\boldsymbol{P}$. The `update` method for a Cholesky factor (class `CHMfactor`) performs the numeric decomposition only.

The conditional density, $f_{\mathcal{U}|\mathcal{Y}}$

- We know the joint density, $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})$, and

$$f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y}) = \frac{f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})}{\int f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})\,d\boldsymbol{u}}$$

  so we almost have $f_{\mathcal{U}|\mathcal{Y}}$. The trick is evaluating the integral in the denominator, which, it turns out, is exactly the likelihood, $L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma^2|\boldsymbol{y})$, that we want to maximize.

- The Cholesky factor, $\boldsymbol{L}_\theta$ is the key to doing this because

$$\boldsymbol{P}^\mathsf{T}\boldsymbol{L}_\theta\boldsymbol{L}_\theta^\mathsf{T}\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}} = \boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}).$$

  Although the `Matrix` package provides a one-step `solve` method for this, we write it in stages:

  Solve $\boldsymbol{L}\boldsymbol{c}_{\boldsymbol{u}} = \boldsymbol{P}\boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ for $\boldsymbol{c}_{\boldsymbol{u}}$.
  Solve $\boldsymbol{L}^\mathsf{T}\boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{c}_{\boldsymbol{u}}$ for $\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}$ and $\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}$ as $\boldsymbol{P}^\mathsf{T}\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}$.

Evaluating the likelihood

- The exponent of $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})$ can now be written

$$\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{\Lambda}_\theta\boldsymbol{u}\|^2+\|\boldsymbol{u}\|^2=r^2(\boldsymbol{\theta},\boldsymbol{\beta})+\|\boldsymbol{L}^{\mathsf{T}}\boldsymbol{P}(\boldsymbol{u}-\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}})\|^2.$$

  where $r^2(\boldsymbol{\theta},\boldsymbol{\beta})=\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{U}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}\|^2+\|\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}}\|^2$. The first term doesn't depend on $\boldsymbol{u}$ and the second is relatively easy to integrate.
- Use the change of variable $\boldsymbol{v}=\boldsymbol{L}^{\mathsf{T}}\boldsymbol{P}(\boldsymbol{u}-\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}})$, with $d\boldsymbol{v}=\operatorname{abs}(|\boldsymbol{L}||\boldsymbol{P}|)\,d\boldsymbol{u}$, in

$$\int\frac{\exp\left(\frac{-\|\boldsymbol{L}^{\mathsf{T}}\boldsymbol{P}(\boldsymbol{u}-\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}})\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{q/2}}\,d\boldsymbol{u}$$

$$=\int\frac{\exp\left(\frac{-\|\boldsymbol{v}\|^2}{2\sigma^2}\right)}{(2\pi\sigma^2)^{q/2}}\,\frac{d\boldsymbol{v}}{\operatorname{abs}(|\boldsymbol{L}||\boldsymbol{P}|)}=\frac{1}{\operatorname{abs}(|\boldsymbol{L}||\boldsymbol{P}|)}=\frac{1}{|\boldsymbol{L}|}$$

  because $\operatorname{abs}|\boldsymbol{P}|=1$ and $\operatorname{abs}|\boldsymbol{L}|$, which is the product of its diagonal elements, all of which are positive, is positive.

Evaluating the likelihood (cont'd)

- As is often the case, it is easiest to write the log-likelihood. On the deviance scale (negative twice the log-likelihood) $\ell(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y}) = \log L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y})$ becomes

$$-2\ell(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y}) = n\log(2\pi\sigma^2) + \frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{\sigma^2} + \log(|\boldsymbol{L}_\theta|^2)$$

- We wish to minimize the deviance. Its dependence on σ is straightforward. Given values of the other parameters, we can evaluate the conditional estimate

$$\widehat{\sigma^2}(\boldsymbol{\theta},\boldsymbol{\beta}) = \frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}$$

  producing the *profiled deviance*

$$-2\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}\right)\right]$$

- However, an even greater simplification is possible because the deviance depends on $\boldsymbol{\beta}$ only through $r^2(\boldsymbol{\theta},\boldsymbol{\beta})$.

Profiling the deviance with respect to $\boldsymbol{\beta}$

- Because the deviance depends on $\boldsymbol{\beta}$ only through $r^2(\boldsymbol{\theta},\boldsymbol{\beta})$ we can obtain the conditional estimate, $\widehat{\boldsymbol{\beta}}_\theta$, by extending the PLS problem to

$$r_\theta^2 = \min_{\boldsymbol{u},\boldsymbol{\beta}} \left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{\Lambda}_\theta \boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 \right]$$

  with the solution satisfying the equations

$$\begin{bmatrix} \boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I}_q & \boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{X} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}} \\ \widehat{\boldsymbol{\beta}}_\theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_\theta^\mathsf{T}\boldsymbol{Z}^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{y}. \end{bmatrix}$$

- The profiled deviance, which is a function of $\boldsymbol{\theta}$ only, is

$$-2\tilde{\ell}(\boldsymbol{\theta}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r_\theta^2}{n}\right)\right]$$

Solving the extended PLS problem

- For brevity we will no longer show the dependence of matrices and vectors on the parameter $\boldsymbol{\theta}$.
- As before we use the sparse Cholesky decomposition, with $\boldsymbol{L}$ and $\boldsymbol{P}$ satisfying $\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}} = \boldsymbol{P}(\boldsymbol{\Lambda}_\theta^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I})$ and $\boldsymbol{c_u}$, the solution to $\boldsymbol{L}\boldsymbol{c_u} = \boldsymbol{P}\boldsymbol{\Lambda}_\theta^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{y}$.
- We extend the decomposition with the $q \times p$ matrix $\boldsymbol{R}_{ZX}$, the upper triangular $p \times p$ matrix $\boldsymbol{R}_X$, and the p-vector $\boldsymbol{c}_{\boldsymbol{\beta}}$ satisfying

$$\begin{aligned}
\boldsymbol{L}\boldsymbol{R}_{ZX} &= \boldsymbol{P}\boldsymbol{\Lambda}_\theta^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X}\\
\boldsymbol{R}_X^{\mathsf{T}}\boldsymbol{R}_X &= \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} - \boldsymbol{R}_{ZX}^{\mathsf{T}}\boldsymbol{R}_{ZX}\\
\boldsymbol{R}_X^{\mathsf{T}}\boldsymbol{c}_{\boldsymbol{\beta}} &= \boldsymbol{X}^{\mathsf{T}}\boldsymbol{y} - \boldsymbol{R}_{ZX}^{\mathsf{T}}\boldsymbol{c_u}
\end{aligned}$$

so that

$$\begin{bmatrix} \boldsymbol{P}^{\mathsf{T}}\boldsymbol{L} & \boldsymbol{0} \\ \boldsymbol{R}_{ZX}^{\mathsf{T}} & \boldsymbol{R}_X^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{L}^{\mathsf{T}}\boldsymbol{P} & \boldsymbol{R}_{ZX} \\ \boldsymbol{0} & \boldsymbol{R}_X \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Lambda}_\theta^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta + \boldsymbol{I} & \boldsymbol{\Lambda}_\theta^{\mathsf{T}}\boldsymbol{Z}^{\mathsf{T}}\boldsymbol{X} \\ \boldsymbol{X}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\Lambda}_\theta & \boldsymbol{X}^{\mathsf{T}}\boldsymbol{X} \end{bmatrix}.$$

Solving the extended PLS problem (cont'd)

- Finally we solve

$$\boldsymbol{R}_X\widehat{\boldsymbol{\beta}}_\theta = \boldsymbol{c}_{\boldsymbol{\beta}}$$
$$\boldsymbol{L}^{\mathsf{T}}\boldsymbol{P}\boldsymbol{\mu}_{\mathcal{U}|\mathcal{Y}} = \boldsymbol{c}_{\boldsymbol{u}} - \boldsymbol{R}_{ZX}\widehat{\boldsymbol{\beta}}_\theta$$

- The profiled REML criterion also can be expressed simply. The criterion is

$$L_R(\boldsymbol{\theta}, \sigma^2|\boldsymbol{y}) = \int L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2|\boldsymbol{y})\, d\boldsymbol{\beta}$$

  The same change-of-variable technique for evaluating the integral w.r.t. $\boldsymbol{u}$ as $1/\operatorname{abs}(|\boldsymbol{L}|)$ produces $1/\operatorname{abs}(|\boldsymbol{R}_X|)$ here and removes $(2\pi\sigma^2)^{p/2}$ from the denominator. On the deviance scale, the profiled REML criterion is

$$-2\tilde{\ell}_R(\boldsymbol{\theta}) = \log(|\boldsymbol{L}|^2) + \log(|\boldsymbol{R}_x|^2) + (n-p)\left[1 + \log\left(\frac{2\pi r_\theta^2}{n-p}\right)\right]$$

- These calculations can be expressed in a few lines of *R* code.

Summary

- For a linear mixed model, even one with a huge number of observations and random effects like the model for the grade point scores, evaluation of the ML or REML profiled deviance, given a value of $\boldsymbol{\theta}$, is straightforward. It involves updating $\boldsymbol{\Lambda}_\theta$, $\boldsymbol{L}_\theta$, $\boldsymbol{R}_{ZX}$, $\boldsymbol{R}_X$, calculating the penalized residual sum of squares, r^2_θ and two determinants of triangular matrices.
- The profiled deviance can be optimized as a function of $\boldsymbol{\theta}$ only. The dimension of $\boldsymbol{\theta}$ is usually very small. For the grade point scores there are only three components to $\boldsymbol{\theta}$.