The Linear Mixed-Effects Probability Model

Christian Haas

A synopsis of the slides presented in the
2nd talk
of the seminary on **Mixed-Effects Models**.

Abstract

We have seen a few examples of different mixed-effect models and datasets. In this talk we now look at the formalized model, using matrix notation. We then want to fit our model to a given dataset, we thus look into the mathematical background of the computational techniques used in the lme4-Package in R.

1 General Model

Two random variables:
$\mathcal{Y}$: the n-dimensional response vector (visible, the data we we get)
$\mathcal{B}$: the q-dimensional vector of random effects (invisible)
with:

$$\mathcal{B} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_\theta). \tag{1}$$

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

Thus the *linear predictor* is:

$$\boldsymbol{Zb} + \boldsymbol{X\beta} \tag{3}$$

With the *model matrices* $\boldsymbol{Z}$ of dimension $n \times q$ and $\boldsymbol{X}$ of dimension $n \times p$, where p is the dimension of the *fixed-effects* parameter vector $\boldsymbol{\beta}$.
$\boldsymbol{\theta}$: the *variance-component parameter vector*; $\mathbf{\Sigma}_\theta$: the *variance-covariance matrix*; $\boldsymbol{\sigma}$: common scale parameter.
We then define:
$\mathbf{\Lambda}_\theta$: relative covariance factor (q x q),

$$\mathbf{\Sigma}_\theta := \sigma^2 \mathbf{\Lambda}_\theta \mathbf{\Lambda}_\theta^T \tag{4}$$

with the spherical random effects $\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 I_q)$, we get $\mathcal{B} = \mathbf{\Lambda}_\theta \mathcal{U}$.

We concentrate on $\mathbf{\Lambda}_\theta$ (not $\mathbf{\Sigma}_\theta$) and $\mathcal{U}$ (not $\mathcal{B}$).
(1) and (2) thus turn into:

$$\mathcal{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 I_q). \tag{5}$$

$$(\mathcal{Y}|\mathcal{U} = \boldsymbol{u}) \sim \mathcal{N}\left(\boldsymbol{Z}\mathbf{\Lambda}_\theta u + \boldsymbol{X\beta}, \sigma^2 \boldsymbol{I}_n\right) \tag{6}$$

and the *linear predictor* becomes:

$$\gamma = \boldsymbol{Z}\mathbf{\Lambda}_\theta u + \boldsymbol{X\beta} \tag{7}$$

the *conditional mean* of $\mathcal{Y}$, given $\mathcal{U} = \boldsymbol{u}$:

$$\mu = E[\mathcal{Y}|\mathcal{U} = \boldsymbol{u}] \tag{8}$$

Note: For a linear mixed model, we have $\mu = \gamma$.

2 Likelihood and its evaluation

Now we want to fit the model parameters $\boldsymbol{\theta}$, $\boldsymbol{\beta}$ and $\boldsymbol{\sigma}$. That is, we are given an observation y_{obs} and want to find the "best" (i.e. the most likely) estimates of those parameters, we can not measure.
The likelihood of those parameters, given the observed data, $\boldsymbol{y}_{obs}$, is the corresponding probability density of $\mathcal{Y}$, evaluated at $\boldsymbol{y}_{obs}$.

We mix up the usual steps (details in the *slides*) to calculate the likelihood and do it the following way:

- Determine joint density of $\mathcal{U}$ and $\mathcal{Y}$: $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})$
- Evaluate $f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y},\boldsymbol{u})$ at y_{obs}. ($\rightarrow$ intermediate function $h(u) := f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{obs},\boldsymbol{u})$)
- Integrate this function $h(u)$ along $\boldsymbol{u}$.

$h(u)$ is called the *unnormalized conditional density.* We understand why, when we see that:

$$f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y}_{obs}) = \frac{h(\boldsymbol{u})}{\int_{R^q} h(\boldsymbol{u})\, d\boldsymbol{u}} \tag{9}$$

Thus the likelihood becomes:

$$L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y}_{obs}) = \int_{R^q} f_{\mathcal{Y},\mathcal{U}}(\boldsymbol{y}_{obs},\boldsymbol{u})\, d\boldsymbol{u} = \int_{R^q} h(\boldsymbol{u})\, d\boldsymbol{u}. \tag{10}$$

We define the *conditional mode* of $\boldsymbol{u}$, given $\mathcal{Y} = \boldsymbol{y}_{obs}$:

$$\tilde{\boldsymbol{u}} := \arg\max_{\boldsymbol{u}} f_{\mathcal{U}|\mathcal{Y}}(\boldsymbol{u}|\boldsymbol{y}_{obs}) = \arg\max_{\boldsymbol{u}} h(\boldsymbol{u}) = \arg\max_{\boldsymbol{u}} f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}_{obs}|\boldsymbol{u}) f_{\mathcal{U}}(\boldsymbol{u}) \tag{11}$$

Looking at (5) and (6) we see that:

$$f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}|\boldsymbol{u}) = \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Z\Lambda}_\theta\,\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{n/2}} \tag{12}$$

$$f_{\mathcal{U}}(\boldsymbol{u}) = \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{q/2}} \tag{13}$$

And thus:

$$h(\boldsymbol{u}) = f_{\mathcal{Y}|\mathcal{U}}(\boldsymbol{y}_{obs}|\boldsymbol{u}) f_{\mathcal{U}}(\boldsymbol{u}) = \frac{\exp(-\left[\|\boldsymbol{y}_{obs}-\boldsymbol{X\beta}-\boldsymbol{Z\Lambda}_\theta\,\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}} \tag{14}$$

Taking the negative log density, we get:

$$-2\log(h(\boldsymbol{u})) = (n+q)\log(2\pi\sigma^2) + \frac{\|\boldsymbol{y}_{obs}-\boldsymbol{X\beta}-\boldsymbol{Z\Lambda}_\theta\,\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2}{\sigma^2} \tag{15}$$

So we get:

$$\tilde{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \|\boldsymbol{y}_{obs}-\boldsymbol{X\beta}-\boldsymbol{Z\Lambda}_\theta\,\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 \tag{16}$$

The expression to be minimized $\|\boldsymbol{y}_{obs}-\boldsymbol{X\beta}-\boldsymbol{Z\Lambda}_\theta\,\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2$ is called the *objective function*, here it is a *penalized residual sum of squares* (**PRSS**).
The minimizer $\tilde{\boldsymbol{u}}$ is called the *penalized least squares* (**PLS**) solution

We think of the **PRSS** criterion as a function of the parameters, given the data, ie.:

$$\boldsymbol{r}^2_{\theta,\beta} = \min_{\boldsymbol{u}} \left[\|\boldsymbol{y}_{obs}-\boldsymbol{X\beta}-\boldsymbol{Z}\Lambda_\theta\;\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2\right] \tag{17}$$

We can also minimize this expression wrt β. The minimum value we get is:

$$\boldsymbol{r}^2_{\theta} = \min_{\boldsymbol{u},\beta} \left[\|\boldsymbol{y}_{obs}-\boldsymbol{X\beta}-\boldsymbol{Z}\Lambda_\theta\;\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2\right] \tag{18}$$

$\tilde{\beta}$: *conditional estimate* of β as the value of β for which the minimum in (18) is attained.
We rephrase (16) using the so-called *pseudo-data approach* by adding *pseudo-observations.*
We get a linear least squares problem:

$$\tilde{\boldsymbol{u}} = \arg\min_{\boldsymbol{u}} \left\| \begin{bmatrix} \boldsymbol{y}_{obs} - \boldsymbol{X\beta} \\ \boldsymbol{0} \end{bmatrix} - \begin{bmatrix} \boldsymbol{Z\Lambda}_\theta \\ \boldsymbol{I}_q \end{bmatrix} \boldsymbol{u} \right\|^2 \tag{19}$$

whose solution satisfies:

$$(\Lambda_\theta^T \boldsymbol{Z}^T \boldsymbol{Z} \Lambda_\theta + \boldsymbol{I}_q)\tilde{\boldsymbol{u}} = \Lambda_\theta^T \boldsymbol{Z}^T (\boldsymbol{y}_{obs} - \boldsymbol{X\beta}) \tag{20}$$

We want fast evaluation of $\tilde{\boldsymbol{u}}$ for different inputs, so we form the *sparse Cholesky factor*, $\boldsymbol{L}_\theta$. It is a lower q x q matrix with:

$$\boldsymbol{L}_\theta \boldsymbol{L}_\theta^T = (\Lambda_\theta^T \boldsymbol{Z}^T \boldsymbol{Z} \Lambda_\theta + \boldsymbol{I}_q) \tag{21}$$

We want this matrix as sparse as possible and thus may permutate the columns of our data beforehands (formally applying a permutation matrix P).
The **PRSS** for general $\boldsymbol{u}$ can then be written as:

$$\|\boldsymbol{y}_{obs} - \boldsymbol{X\beta} - \boldsymbol{Z\Lambda}_\theta\, \boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 = r^2_{\theta,\beta} + \|\boldsymbol{L}_\theta^T(\boldsymbol{u} - \tilde{\boldsymbol{u}})\|^2 \tag{22}$$

Using (11), (14) and (22) we are now able to evaluate $L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y}_{obs})$ and get:

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}, \sigma | \boldsymbol{y}_{obs}) = \frac{\exp(-\frac{r^2_{\theta,\beta}}{2\sigma^2})}{(2\pi\sigma^2)^{n/2} |\boldsymbol{L}_\theta|}$$

So the *deviance* (negative twice the log-likelihood) is:

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

The maximum-likelihood estimates for the parameters are those, that minimize this deviance (a numerical problem) By using the dependances between the parameters as seen in the **PRSS** (we can find the minimizers of β and u for any given θ), we can reduce this to a function only of θ.
This is called the *profiled deviance*:

$$\tilde{d}(\theta | \boldsymbol{y}_{obs}) = 2\log|\boldsymbol{L}_\theta| + n\left[1 + \log\left(\frac{2\pi r^2_\theta}{n}\right)\right] \tag{23}$$

Now minimization of $\tilde{d}(\theta | \boldsymbol{y}_{obs})$ wrt θ determines the MLE $\tilde{\theta}$.
The MLEs for $\hat{\beta}$ and $\hat{\sigma}$ then are the corresponding conditional estimates evaluated at $\hat{\theta}$.
So we found all maximum-likelihood-estimators. That is we fitted our model to the data.