Selected Topics in Linear Mixed-Effects Models

Jürg Schelldorfer, Manuel Koller, Markus Kalisch

Seminar für Statistik, ETH Zürich

March 29, 2010

Table of Contents

ergoStool

$i = 1, \ldots, 9$ subjects
$j = 1, \ldots, 4$ different stools
Response y_{ij}: Effort required to arise from each stool
total $n = 36$ observations

```
'data.frame': 36 obs. of  3 variables:
 $ effort : num  12 15 12 10 10 14 13 12 7 14 ...
 $ Type   : Factor w/ 4 levels "T1","T2","T3",..: 1 2 3 4 1 2 3 4 1 2 ...
 $ Subject: Factor w/ 26 levels "A","B","C","D",..: 1 1 1 1 2 2 2 2 3 3 ...
```

$\longrightarrow$ R

Model Formulation 1

$$y_{ij} = \mu + \beta_j + b_i + \varepsilon_{ij} \quad i = 1, \ldots, 9 \quad j = 1, \ldots, 3 \tag{1}$$

with

$$b_i \sim \mathcal{N}_1(0, \sigma_b^2) \quad \varepsilon_{ij} \sim \mathcal{N}_1(0, \sigma^2) \quad \varepsilon_{ij} \perp b_i \quad \forall i, j$$

Model Formulation 2

$$\boldsymbol{y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\varepsilon}_i \quad i = 1, \dots, 9 \tag{2}$$

with

$$b_i \sim \mathcal{N}_1(0, \sigma_b^2) \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}_4(0, \sigma^2\boldsymbol{I}) \quad \varepsilon_i \perp b_i \quad \forall i, j$$

where
$\boldsymbol{y}_i \in \mathbb{R}^4$, $\boldsymbol{X}_i \in \mathbb{R}^{4\times 4}$, $\boldsymbol{Z}_i \in \mathbb{R}^{4\times 1}$, $\boldsymbol{\varepsilon}_i \in \mathbb{R}^4$.

$$\boldsymbol{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \end{bmatrix}, \boldsymbol{X}_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}, \boldsymbol{Z}_i = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \\ \varepsilon_{i4} \end{bmatrix}$$

$\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are the same for all subjects (in this example only).

Model Formulation 3

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon \tag{3}$$

with

$$\boldsymbol{b} \sim \mathcal{N}_q(0, \sigma_b^2 \boldsymbol{I}_{q\times q} = \Sigma_\theta) \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n}) \quad \varepsilon \perp b$$

and

$$\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n\times p}, \boldsymbol{Z} \in \mathbb{R}^{n\times q}, \Sigma_\theta = \sigma^2 \Lambda_\theta \Lambda_\theta^T$$

Remark:
This is the most general formulation.
Some models can not be written in Form (1) or (2)!

$\longrightarrow$ R

ScotsSec I

response: attainment scores of 3435 students in seconday school
covariates:
- primary: factor for each primary school with 148 levels
- secondary: factor for secondary school with 19 levels
- sex: sex of student
- verbal: verbal reasoning score on entry
- social: The student's social class from low to high social class.

$\longrightarrow$ R

ScotsSec I

With $n = 3435$, $p = 4$, $q = 167$ we can write the model as:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{Z}\boldsymbol{b} + \varepsilon \tag{4}$$

with

$$\boldsymbol{b} \sim \mathcal{N}_q(0, \Sigma_\theta) \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_{n\times n}) \quad \varepsilon \perp \boldsymbol{b}$$

and

$$\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{X} \in \mathbb{R}^{n\times p}, \boldsymbol{Z} \in \mathbb{R}^{n\times q}, \Sigma_\theta \in \mathbb{R}^{q\times q}$$

Data structure in lme4

Take home message:

Plot your data set in an appropriate way!

Let's look in R!

Confidence Intervals I

Goal:
Evaluate confidence intervals for the parameters.

Naive Approach:
Approximate the distribution of the parameters by a normal distribution and derive confidence intervals using an approximate standard error.

Confidence Intervals II

However:
Confidence Intervals for variance components can be heavily skewed!
$\rightarrow$ In general, the naive approach is not appropriate!

Since the distribution can not be well approximated by a normal distribution, it is not meaningful neither to determine confidence intervals nor to calculate p-values based on this assumption!

Idea:
Find a way to examine if the normal approximation is appropriate.

Profile Zeta Plot I

Suggestion:
Make a plot that shows the *sensitivity* of the model fit to changes in one particalur parameter.

$\longrightarrow$ R

Profile Zeta Plot II

Calculation of the Profile Zeta Plot:

1. Calculate the *globally optimal fit* $\rightarrow \mathcal{M}_0$
2. Fit the model with one parameter fixed at a specific value $\rightarrow \mathcal{M}_k$
3. Compare $\mathcal{M}_0$ and $\mathcal{M}_k$ by the LRT statistic t_k
4. Apply a *signed root transformation* to $t_k \rightarrow \zeta_k$
5. Draw a QQ Plot of $\zeta_0, \zeta_1, \ldots$

Profile Zeta Plot III

Interpretation:

- *Ideally it is a straight line.* Then perform inference based on the parameter's estimate, its standard error and quantiles of the standard normal distribution
- $\log(\sigma)$ is straight, so $\log(\sigma)$ has a good normal approximation.
- This does not hold neither for σ nor σ^2!!
- The CI for β_0 are wider than those based on a normal approximation.

Profile Pairs Plot I

Profile Zeta Plot:
shows the *sensitity* of the model to changes in parameters.

Profile Pairs Plot:
shows how the *parameters influence each other*.

$\longrightarrow$ R

Profile Pairs Plot I

Calculation:

1. Fix one parameter, i.e. σ_1. Calculate the conditional estimates of the other parameters σ and β_0. This gives the profile traces (vertical and horizontal lines).
2. Contour lines correspond to the marginal confidence intervals at different confidence levels.

What to learn from it

Interpretation:

- *Ideally there are ellipses.* Look at distortions from an elliptical shape.
- straight line: the conditional estimate of β_0 , given σ_1, is constant
- curved line: the conditional estimate of σ_1 given β_0 depends on β_0.
- small values of σ_1 inflate the estimate of $\log(\sigma)$ because the variability of the random effects gets transfered to variability in the error.
- We see the distortions from elliptical shape in the lower right part.

Key Tools in lme4

- Reduce the optimization problem to one involving θ only (profiling)
- use sparse matrix storage formats and sparse matrix computations
- The sparse choleski decomposition can easily be calculated

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

where $\boldsymbol{P}$ is a permuation matrix

REML and ML: Recap

Last talk:

- The ML and REML estimators in linear regression are

$$\hat{\sigma}^2_{ML} = \frac{RSS}{n} \quad \hat{\sigma}^2_{REML} = \frac{RSS}{n-p}$$

- The REML estimates of the variance components are less biased than the ML estimates in the linear mixed model setting.

Is that all to say?

REML

Let M_1 and M_2 be two nested models we want to compare.

Then (as a rule of thumb):
Models with different fixed-effects structes using REML should not be compared by a LRT. Use ML estimates in this case!

$\longrightarrow$ R

Tests at the boundary of the parameter space

Let's look at the Pastes Example:
$\longrightarrow$ R

So the test problem can be easily formulated... Test of interest:

$$H_0 : \sigma_2 = 0 \text{ versus } H_A : \sigma_2 > 0$$

and a LRT-test may be done in R...

...where we used `anova(fm3a,fm3)` using a χ^2_1 distribution.

But...

However:
We have to be cautious because the test statistic is not χ^2_1 distributed! The p-value is too conservative (i.e. too large)!

"Theoretical Result":
The asymptotic null distribution for the LRT is a mixture of a χ^2_k *and a* χ^2_{k+1} *distribution with equal weight* $1/2$*, where* k *is the number of correlated random effects.*

In the Pastes Example:

$$1/2\chi^2_0 + 1/2\chi^2_1$$

Sleepstudy: Data Set

$\longrightarrow$ R

Orthodont: Data set

$\longrightarrow$ R

Introduction

This analysis is a mixture of tools and concepts from the upcoming book of Douglas Bates and the book of West et. al.

Data set I

$n = 1190$ students sampled from 312 classroms in 107 schools.

- `sex` Sex of student
- `minority` 0=nonminority student, 1=minority student
- `mathkind` math score in the kindergarten
- `mathgain` change in student math scores from kindergarten to first grade
- `ses` Student socioeconomic status
- `yearstea` first-grade teacher's years of teaching experience
- `mathknow` teacher's mathematical knowledge
- `housepov` percentage of households in the neighbourhood of the school below the poverty level
- `mathprep` teacher's mathematics preparation
- `classid` identifying the classrooom (312 levels)
- `schoolid` identifying the school (107 levels)

Data set II

Some more information? - YES!!

- `mathgain` is the response variable
- schools and classrooms are randomly selected
- Student is nested in classroom and classroom in school

Three-level data set:

- students (Level 1)
- students are nested within classrooms (Level 2)
- classrooms are nested within schools (Level 3)

Three-Level Data

Allocate the covariates to the levels:

- (Level 1) `mathkind`, `sex`, `minority`, `ses`
- (Level 2) `classid`, `yearstea`, `mathprep`, `mathknow`
- (Level 3) `schoolid`, `housepov`

$\longrightarrow$ R

How to proceed?

Model Building Strategy

1. Start with a menas-only Level 1 including random effects from Level 2 and Level 3
2. Add Level 1 covariates
3. Add Level 2 covariates

1. Model

Model Formulation:
For $i = 1, \ldots, 107, j = 1, \ldots, 312$

$$y_{ijk} = \mu + u_i + v_{j(i)} + \epsilon_{ijk}$$

$\longrightarrow$ R

lme and lmer

- lme4 uses the full matrix approach
- lme4 can fit more general models than nlme
- lme4 can fit large data sets very fast (i.e. 378’047 test scores of 134’713 students in 3722 schools)