MATH3091 Computer Lab 4

22 Feb 2022

Preliminary

This week we will carry out data analysis using linear mixed models (LMMs). This requires an adds-on R package called `lme4`, which provides functions to automatically fit and analyse LMMs.

We first install and load the `lme4` package.

```
install.packages("lme4")
library(lme4)
```

In `lme4`, the linear mixed modelling is implemented by the `lmer` function. Very similar to `lmer` for linear models, it takes as its first two arguments specifying the model formula and the data with which to evaluate the formula, i.e.

```
lmer(formula, data=NULL,...)
```

The `fomula` here is used to describe the linear model, similar to that in the `lm` function. In this case it includes both fixed-effects and random-effects terms. The `fomula` take the form `resp∼expr`, where `resp` determines the response variable and `expr` is an expression of explanatory variables. For example, formulas in the `lmer` function should contain random-effects terms as follows:

```
resp ∼ FEexpr + (REexpr1 | factor1) ...
```

where `FEexpr` is an expression determining the fixed-effects, and `(REexpr1 | factor1)`is an expression determining a random-effect, with groups determined by `factor1`.

In `lmer`, the second argument, `data`, is optional but recommended and is usually the name of an R data frame, this is also similar to that in the `lm` function.

The `mathachieve` dataset

The dataset `mathachieve` that we will analyse today is from the 1982 "High School and Beyond" survey, and pertain to 7185 high-school students from 160 schools. The response variable variable `mathach` is the student's score on a math-achievement test, and there are 3 explanatory variables of our interest: `cses`, the adjusted socioeconomic status of the student's family; `meanses`, the average socioeconomic status for students in each school; and `sector`, a factor coded `Catholic` or `Public` for the type of student's school.

The variable `School` is an identification number for the student's school. The schools define groups — it is unreasonable to assume that students in the same school are independent of one-another. Note that `sector` is a school-level variable and hence is identical for all students in the same school.

Let's import the date file into R:

```
mathachieve <- read.csv("mathachieve.csv")
```

Fit the data with LMMs

Recall the general form of LMM in the lecture notes, which is:

$$Y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{u}_{ij}^T \boldsymbol{\gamma}_i + \epsilon_{ij}, \qquad j = 1, \ldots, n_i;\ i = 1, \ldots, m, \tag{1}$$

where $\boldsymbol{\gamma}_i \sim N(\boldsymbol{0}, \boldsymbol{D})$ and $\epsilon_{ij} \sim (0, \sigma^2)$.

Consider a linear model consists of two hierarchical levels: First, within schools, we have the regression of math achievement on the individual-level covariate `cses`; Then, at the school level, we will entertain the possibility that the math achievement depend upon sector and upon the average level covariate `meanses` in the schools. Here the hierarchical model is created due to we group the data by schools. Therefore, including the intercept terms, we can write the following model:

$$\texttt{mathach}_{ij} = \beta_0 + \beta_1 \texttt{meanses}_i + \beta_2 \texttt{sector}_i + \beta_3 \texttt{cses}_{ij} + \gamma_{i0} + \gamma_{i1} \texttt{cses}_{ij} + \epsilon_{ij} \tag{2}$$

where the β's are fixed effects, while the γ's are random effects, $m = 160$ is the number of groups (schools). The change from equation (1) to equation (2) is purely notational.

As a result, to implement the LMMs we need to estimate the following 3 unknown parameters of interest:

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3), \quad \sigma^2, \quad \text{and} \quad \boldsymbol{D} = \begin{pmatrix} d_{11}, & d_{12} \\ d_{21}, & d_{22} \end{pmatrix}$$

Recall that earlier we claimed to use `lmer` function random-effects terms should be of the form (`expr | factor`). Therefore, to fit the linear mixed models here we just need to simply organise the formula as requested,

```
math.lme.1 <- lmer(mathach ~ meanses + sector +  cses + (cses | school),
                   data=mathachieve )
```

Similar to the linear regression, you can check the model fit by

```
summary(math.lme.1)
```

Notice that the formula for the random effects includes only the term `cses`; as in a linear-model formula, a random intercept is implied unless it is explicitly excluded (by specifying `-1` in the random formula).

In the outpout, for random effects it displays estimates of the variance and covariance parameters for the random effects. Therefore, we have $\hat{d}_{11} = 2.3851$ and $\hat{d}_{22} = 0.7004$. The term labelled `Residual` is the estimate of σ^2, where we obtain that $\hat{\sigma}^2 = 36.7098$.

> **Question:** How would you estimate $\hat{d}_{12}$?

The table of fixed effects is similar to output from `lm`, where we have $\hat{\beta}_0 = 13.4356$, $\hat{\beta}_1 = 5.2463$, $\hat{\beta}_2 = -1.3722$, and $\hat{\beta}_3 = 2.1950$.

The panel labelled `Correlation` gives the estimated sampling correlations among the fixed-effect coefficient estimates, which are not usually of direct interest. Very large correlations (close to 1 in absolute value), however, are indicative of an ill-conditioned model.

> **Question:** How would you run above linear mixed model with two additional interaction terms `meanses:cses` and `cses:sector` as fixed effects?

> **Question:** How would you run above linear mixed model without the intercept term in random effect?

> **Question:** How would you run above linear mixed model with only the intercept term appear in random effect?

REML or MLE

By default, we estimate the parameters (fixed effects and $\boldsymbol{D}$) using the restricted maximum likelihood (REML) method. If we want to switch to the MLE estimate, we just need to specify `RMEML=FALSE`, for example:

```
math.lme.1.mle <- lmer(mathach ~ meanses + sector +  cses + (cses | school),
                      data=mathachieve, REML=FALSE )
summary(math.lme.1.mle)
```

This will have a slightly different estimate.

Model comparison

We can use AIC or BIC socre, or ANOVA (generliased likelihood ratio test) to compare different models. Try

```
AIC(math.lme.1, math.lme.2, math.lme.3, math.lme.4, math.lme.1.mle )
```

and

```
anova(math.lme.1, math.lme.2, math.lme.3, math.lme.4, math.lme.1.mle)
```

> **Question:** Do you know what are the degrees of freedom in above models? Why?