Fitting Linear Mixed-Effect Models in R

Yizeng Li

University of South Carolina

yizeng@email.sc.edu

April 30, 2019

A Quick Outline

What is LMM?

1) Linear mixed-effects models are extensions of linear regression models for data that are collected and summarized in groups. It consists of two parts, fixed effects and random effects.
2) The LLM form is

$$\underbrace{\mathbf{y}}_{n\times 1} = \underbrace{\mathbf{X}\beta}_{n\times p \;\; p\times 1 \text{ fixed effects}} + \underbrace{\mathbf{Z}\mathbf{b}}_{n\times q \;\; q\times 1 \text{ random effects}} + \underbrace{\epsilon}_{n\times 1 \text{ error}},$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{n\times n})$ and $\mathbf{b} \sim \mathcal{N}(0, \Sigma_{q\times q})$.

What is LMM?

1) Linear mixed-effects models are extensions of linear regression models for data that are collected and summarized in groups. It consists of two parts, fixed effects and random effects.
2) The LLM form is

$$\underbrace{\mathbf{y}}_{n\times 1} = \underbrace{\mathbf{X}\beta}_{n\times p \;\; p\times 1 \text{ fixed effects}} + \underbrace{\mathbf{Zb}}_{n\times q \;\; q\times 1 \text{ random effects}} + \underbrace{\epsilon}_{n\times 1 \text{ error}},$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_{n\times n})$ and $\mathbf{b} \sim \mathcal{N}(0, \Sigma_{q\times q})$.

How to fit LMM in R?

In the syntax of R's **lme4** package (Bates et al., 2015), we will use

LLM with random intercepts only

```
lmer(y ~ x + (1 | b), data=yourdataset)
```

Specifically, `(1 | b)` means that there is a single random factor which is constant within each level and its levels are given by the grouping factor `b`.

LLM with random intercepts and slopes

```
lmer(y ~ x + (x | b), data=yourdataset)
```

Specifically, `(x | b)` means that there is a single random factor which is linear with `x` within each level and its levels are given by the grouping factor `b`.

Dataset Format

The average reaction time per day for subjects in a sleep deprivation study. On day 0 the subjects had their normal amount of sleep. Starting that night they were restricted to 3 hours of sleep per night. The observations represent the average reaction time on a series of tests given each day to each subject. Note it is a **longitudinal study**, a research design that involves repeated observations of the same variables (e.g., `subject`) over short or long periods of time

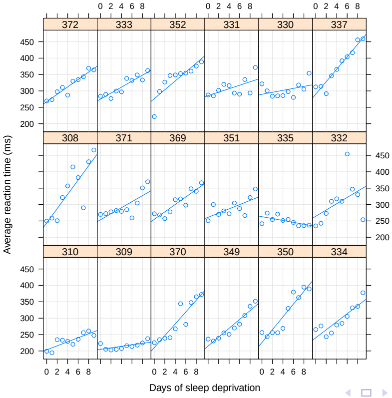
```
library(Matrix)
library(lme4)
library(lattice)

str(sleepstudy)

## 'data.frame': 180 obs. of  3 variables:
##  $ Reaction: num  250 259 251 321 357 ...
##  $ Days    : num  0 1 2 3 4 5 6 7 8 9 ...
##  $ Subject : Factor w/ 18 levels "308","309","310",..: 1 1 1 1 1 1 1 1 1
```

Data Visualization

```
print(xyplot(Reaction ~ Days | Subject, sleepstudy,
        aspect = "xy",layout = c(6,3),type = c("g","p","r"),
        index.cond = function(x,y) coef(lm(y ~ x))[1],
        xlab = "Days of sleep deprivation",
        ylab = "Average reaction time (ms)"))
```



Let's fit a LMM first!

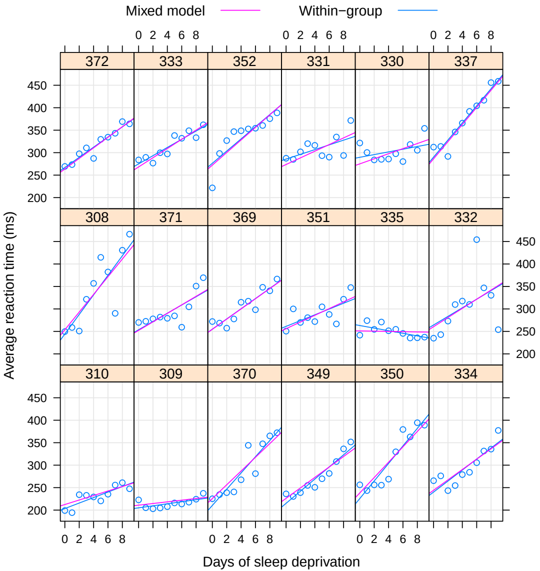
```
(fm1 <- lmer(Reaction ~ Days + (Days | Subject), sleepstudy))

## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (Days | Subject)
##    Data: sleepstudy
## REML criterion at convergence: 1743.628
## Random effects:
##  Groups   Name        Std.Dev. Corr
##  Subject  (Intercept) 24.737
##           Days         5.923   0.07
##  Residual             25.592
## Number of obs: 180, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##      251.41        10.47
```

Matrix expression

$$
\underbrace{\begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \\ 1 & 8 \\ 1 & 9 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 8 \\ 1 & 9 \end{pmatrix}}_{\mathbf{X}_{180\times 2}}
\underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\boldsymbol{\beta}_{2\times 1}}
+
\underbrace{\begin{pmatrix}
1 & 0 & 0 & 0 & \cdots & 0 \\
1 & 1 & 0 & 0 & \cdots & 0 \\
1 & 2 & 0 & 0 & \cdots & 0 \\
1 & 3 & 0 & 0 & \cdots & 0 \\
1 & 4 & 0 & 0 & \cdots & 0 \\
1 & 5 & 0 & 0 & \cdots & 0 \\
1 & 6 & 0 & 0 & \cdots & 0 \\
1 & 7 & 0 & 0 & \cdots & 0 \\
1 & 8 & 0 & 0 & \cdots & 0 \\
1 & 9 & 0 & 0 & \cdots & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & 1 & \cdots & 0 \\
0 & 0 & 1 & 2 & \cdots & 0 \\
0 & 0 & 1 & 3 & \cdots & 0 \\
0 & 0 & 1 & 4 & \cdots & 0 \\
0 & 0 & 1 & 5 & \cdots & 0 \\
0 & 0 & 1 & 6 & \cdots & 0 \\
0 & 0 & 1 & 7 & \cdots & 0 \\
0 & 0 & 1 & 8 & \cdots & 0 \\
0 & 0 & 1 & 9 & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 9
\end{pmatrix}}_{\mathbf{Z}_{180\times 36}}
\underbrace{\begin{pmatrix} b_{1,0} \\ b_{1,1} \\ b_{2,0} \\ b_{2,1} \\ b_{3,0} \\ b_{3,1} \\ b_{4,0} \\ b_{4,1} \\ b_{5,0} \\ b_{5,1} \\ \vdots \\ b_{17,0} \\ b_{17,1} \\ b_{18,0} \\ b_{18,1} \end{pmatrix}}_{\mathbf{b}_{36\times 1}}
=
\underbrace{\begin{pmatrix} \beta_0 \\ \beta_0+\beta_1 \\ \beta_0+2\beta_1 \\ \beta_0+3\beta_1 \\ \beta_0+4\beta_1 \\ \vdots \\ \beta_0+9\beta_1 \\ \beta_0 \\ \beta_0+\beta_1 \\ \beta_0+2\beta_1 \\ \beta_0+3\beta_3 \\ \vdots \\ \beta_0+9\beta_1 \\ \vdots \\ \beta_0+9b_1 \end{pmatrix}}_{\text{fitted fixed effects}}
+
\underbrace{\begin{pmatrix} b_{1,0} \\ b_{1,0}+b_{1,1} \\ b_{1,0}+2b_{1,1} \\ b_{1,0}+3b_{1,1} \\ b_{1,0}+4b_{1,1} \\ \vdots \\ b_{1,0}+9b_{1,1} \\ b_{2,0} \\ b_{2,0}+b_{2,1} \\ b_{2,0}+2b_{2,1} \\ \vdots \\ b_{2,0}+9b_{2,1} \\ \vdots \\ b_{18,0}+9b_{18,1} \end{pmatrix}}_{\text{fitted random effects}}
= \mathbf{Y}_{180\times 1}
$$

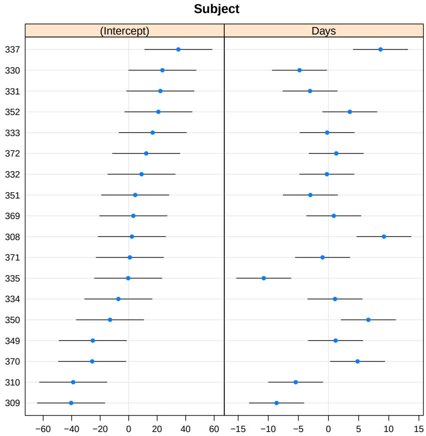
Fitted vs Observed



95% Confidence Interval for random effects

Note some of the prediction intervals for the random effects overlap zero.

```
dotplot(ranef(fm1,cond=TRUE),
scales = list(x = list(relation = 'free')))[["Subject"]]
```



Fit a LMM with uncorrelated random effects

The estimated correlation between random intercept and random slope (0.07) in `fm1` model is quite small. We could consider a model with uncorrelated random effects.

```
(fm2 <- lmer(Reaction ~ Days + (1 | Subject) +(0+Days | Subject),
             sleepstudy))
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
##    Data: sleepstudy
## REML criterion at convergence: 1743.669
## Random effects:
##  Groups    Name        Std.Dev.
##  Subject   (Intercept) 25.050
##  Subject.1 Days         5.989
##  Residual              25.565
## Number of obs: 180, groups:  Subject, 18
## Fixed Effects:
## (Intercept)         Days
##      251.41        10.47
```

Compare two models

1) Because the large p-value indicates that we would not reject `fm2` in favor of `fm1`, we prefer the more parsimonious `fm2`.
2) This conclusion is consistent with the AIC (Akaike's Information Criterion) and the BIC (Bayesian Information Criterion) values for which "smaller is better".

```
anova(fm2, fm1)

## refitting model(s) with ML (instead of REML)

## Data: sleepstudy
## Models:
## fm2: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
## fm1: Reaction ~ Days + (Days | Subject)
##     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## fm2  5 1762.0 1778.0 -876.00   1752.0
## fm1  6 1763.9 1783.1 -875.97   1751.9 0.0639      1     0.8004
```

Compare two models

1) Because the large p-value indicates that we would not reject `fm2` in favor of `fm1`, we prefer the more parsimonious `fm2`.
2) This conclusion is consistent with the AIC (Akaike's Information Criterion) and the BIC (Bayesian Information Criterion) values for which "smaller is better".

```
anova(fm2, fm1)

## refitting model(s) with ML (instead of REML)

## Data: sleepstudy
## Models:
## fm2: Reaction ~ Days + (1 | Subject) + (0 + Days | Subject)
## fm1: Reaction ~ Days + (Days | Subject)
##     Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## fm2  5 1762.0 1778.0 -876.00   1752.0
## fm1  6 1763.9 1783.1 -875.97   1751.9 0.0639      1     0.8004
```

Pros and Cons of LMM

1) Compared with LMM, the existing linear regression could be employed to fit the data. However, we can not ignore the reaction time of each day for subject is not independent, and the sample size is quickly reduced if we run multiple linear analyses within each subject. On the other hand, LMM utilized all data, even when we have low sample sizes, structured data and many covariates, in fitting.
2) A potential disadvantage of LMM is that people are not quite familar with LMM, in particular when there are non-linear trends appearring in the data. Also approximations (no closed form) usually have to be used in estimating parameters of models. Nevertheless, LMM offer a powerful and flexible tool for analysis of longitudinal data.
3) We perform how to fit LLM in R, how to visualize the data and compare models using ANOVA test in R. Due the limited time, further resisual analysis and assumptions validation are not included in this report. The results support the model well.

Pros and Cons of LMM

1) Compared with LMM, the existing linear regression could be employed to fit the data. However, we can not ignore the reaction time of each day for subject is not independent, and the sample size is quickly reduced if we run multiple linear analyses within each subject. On the other hand, LMM utilized all data, even when we have low sample sizes, structured data and many covariates, in fitting.
2) A potential disadvantage of LMM is that people are not quite familar with LMM, in particular when there are non-linear trends appearring in the data. Also approximations (no closed form) usually have to be used in estimating parameters of models. Nevertheless, LMM offer a powerful and flexible tool for analysis of longitudinal data.
3) We perform how to fit LLM in R, how to visualize the data and compare models using ANOVA test in R. Due the limited time, further resisual analysis and assumptions validation are not included in this report. The results support the model well.

Pros and Cons of LMM

1) Compared with LMM, the existing linear regression could be employed to fit the data. However, we can not ignore the reaction time of each day for subject is not independent, and the sample size is quickly reduced if we run multiple linear analyses within each subject. On the other hand, LMM utilized all data, even when we have low sample sizes, structured data and many covariates, in fitting.
2) A potential disadvantage of LMM is that people are not quite familar with LMM, in particular when there are non-linear trends appearring in the data. Also approximations (no closed form) usually have to be used in estimating parameters of models. Nevertheless, LMM offer a powerful and flexible tool for analysis of longitudinal data.
3) We perform how to fit LLM in R, how to visualize the data and compare models using ANOVA test in R. Due the limited time, further resisual analysis and assumptions validation are not included in this report. The results support the model well.

Let's end here...

Thank you