ST 437/537: Applied Multivariate and Longitudinal Data Analysis
Longitudinal Data Analysis: Linear Mixed Effects Model (LMM)

Arnab Maity
NCSU Department of Statistics
SAS Hall 5240   919-515-1937   amaity[at]ncsu.edu

References:

- Modeling Longitudinal Data by Robert E. Weiss. New York: Springer.

- Linear Mixed Models for Longitudinal Data by Geert Verbeke and Geert Molenberghs. New York: Springer.

- Applied Longitudinal Analysis by Fitzmaurice by G.M., Laird, N.M., and Ware, J.H. New York: Wiley (on reserve at NCSU library)

Introduction

Observed data are repeated observations from n units/subjects/objects, indexed by i; $\{Y_{ij}, t_{ij}$, and other covariates $: j = 1, \ldots, m_i\}$. The **linear mixed effect model (LMM)** is the model obtained by aggregating the two levels below:

1st stage: $Y_i = X_i\beta + Z_i\boldsymbol{b}_i + \boldsymbol{e}_i$ with $\boldsymbol{e}_i$ zero-mean random error vector.

- β is p-dimensional vector of fixed parameters; X_i is a $m_i \times p$ dimensional fixed design matrix

- $\boldsymbol{b}_i$ is q-dimensional vector of random parameters; Z_i is a $m_i \times q$ dimensional random design matrix

- $\boldsymbol{e}_i$ is m_i-dimensional vector of residuals. It is assumed independent over i. We assume $\boldsymbol{e}_i \sim N(0, R_i)$, where R_i is the covariance matrix of $\boldsymbol{e}_i$

2nd level: Describe the model for the subject-parameters

$$b_i \sim N_q(0, D),$$

where D is a covariance matrix.

In our *Dental data* example, we had the model Putting the models from the two stages together, we obtain the following:

$$Y_{ij} = \beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij} + e_{ij}.$$

Note that we can write this model in a matrix form. Recall that $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_i})^T$ is the response vector of the i-th subject. So we can write

$$
\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{im_i} \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{im_i} \end{pmatrix}}_{X_i} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \underbrace{\begin{pmatrix} 1 & t_{i1} \\ \vdots & \vdots \\ 1 & t_{im_i} \end{pmatrix}}_{Z_i} \begin{pmatrix} b_{i0} \\ b_{i1} \end{pmatrix} + \begin{pmatrix} e_{i1} \\ \vdots \\ e_{im_i} \end{pmatrix}
$$

Interpretation of the model parameters

Conditional perspective: Write down the distribution of $Y_i | b_i$, that is calculate the conditional mean and the conditional variance.

- $E[Y_{ij} | b_i]$: The conditional mean describes the mean response of an individual/unit

- $Var(Y_i | b_i)$

Marginal perspective: Write down the distribution of Y_i, that is, calculate the marginal mean and the marginal variance

- $E[Y_{ij}]$: The marginal mean describes the mean response in the population. For this reason the fixed parameter β is also referred to by ``population-average '' parameter. It has population average interpretation.

- $Var(Y_i)$

Thus the model parameters are:

- Fixed:
    - mean regression parameters β

- ○ covariance regression parameters ω = vector of the elements of D and the parameters of R_i.
  - Random:
    - ○ subject specific random effects $\boldsymbol{b}_i$.

In the following section, we discuss the **estimation of the fixed effects** and the **prediction of the random effects**.

Estimation/inference of Fixed effects parameters

Intuition: Estimation and inference about the fixed effects is based on the marginal model for $\boldsymbol{Y}_i$, and hence it follows the same general ideas as the estimation/inference in the population average models (using `gls`). **Estimation of the fixed effects is based on the conditional model** $Y_i|b_i$.

Recall the model

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\beta + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i,$$

where $\boldsymbol{b}_i \sim N(0, D)$ and $\boldsymbol{e}_i \sim N(0, R_i)$ with $\boldsymbol{b}_i$ independent of $\boldsymbol{e}_i$. Denote by ω the vector of the covariance parameters (the parameters describing the covariances D and R_i).

Inference for the classical LMM focuses on the fixed effects parameters β. For these models estimation uses the methods based on the marginal log-likelihood. + ML/REML are used to estimate both β and ω.

- Consequently the inferential tools are very similar to those introduced for the marginal models for correlated data.

The same methods for carrying inference for β apply:

- GLS to estimate β, denoted by $\widehat{\beta}$

- $var[\widehat{\beta}] = V_{\widehat{\beta}}(\omega)$ as before (see the lecture on General Linear Models)

- Distribution of $\widehat{\beta} \sim N(\beta, V_{\widehat{\beta}}(\omega))$.

- For large samples substitute the true covariance parameter ω by its estimate $\widehat{\omega}$. For notational simplicity we use $\widehat{V_{\widehat{\beta}}} = V_{\widehat{\beta}}(\widehat{\omega})$.

- **Testing hypotheses** involving β is done by formulating the hypothesis as $L\beta = h$ and using **Wald/LRT procedures** studied earlier.

Prediction of Random effects

One main advantage of this methodology is the ability to predict individual trajectories, by incorporating individual specific information. The idea is to exploit the fact that the individual mean is described by $X_i\beta + Z_i b_i$. In particular if $\widehat{\beta}$ and $\widehat{b}_i$ are the estimates of β and b_i, then the ith individual trajectory can be "predicted" by

$$\widehat{E[Y_i|b_i]} = X_i\widehat{\beta} + Z_i\widehat{b}_i$$

> *Terminology: **Predictor** is an estimator of a random quantity. In contrast, **estimator** is a term correctly used only in regards to a fixed unknown parameter.*

Intuition: the subject specific random deviations $\boldsymbol{b}_i$ will be predicted using **best linear unbiased prediction (BLUP)**. The predictors are:

i. **linear functions** of the data (hence the name *linear*)

ii. **unbiased** - that is the average value of the predictor is equal to the average (random) quantity being predicted;

iii. **best** - in the sense that they achieve minimum mean squared error among all the linear + unbiased predictors.

BLUPs were introduced by Henderson (1950) - as the "joint maximum likelihood estimates"; although the actual name/acronym was used later by Goldberger (1962). He latter corrected this name on the basis that the function being maximized is not a proper likelihood function.

The predictor $\boldsymbol{b}_i$ is

$$\widehat{\boldsymbol{b}}_i = \widehat{D} \boldsymbol{Z}_i^T \widehat{\Sigma}_i^{-1} (\boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\beta}).$$

It is called the **empirical/approximate BLUP**, also the **Empirical Bayes estimator** for $\boldsymbol{b}_i$.

There are closed form expresions of standard errors of $\boldsymbol{b}_i$.

Let us look at the dental data again. We fit the model

$$y_{ij} = G_i \beta_{0G} + G_i t_{ij} \beta_{1G} + (1 - G_i) \beta_{0M} + (1 - G_i) t_{ij} \beta_{1M} + b_{0i} + t_{ij} b_{1i} + e_{ij},$$

where

- the errors have diagonal within-subject covariance, that is same for each gender
- the random effects have a unstructured 2×2 covariance matrix that is same for each gender.

```
library(nlme)
data(Orthodont)
# Create covariates
id <- Orthodont$Subject
age <- Orthodont$age
distance <- Orthodont$distance
sex <- Orthodont$Sex

# Dummy variables
G <- (sex == "Female") + 0 #Girls=1, Boys=0
M <- 1 - G #Defining (1-G) as a separate var for simplicity

# Formula for fixed and random effects
model.formula <- distance ~ -1 + G + G:age + M + M:age
random.formula <- ~ age | id

fit.reml <- lme(model.formula,
            random = random.formula,
            method="REML")
```

Notice that **we use `method="REML"`** ; this is to obtain the Empirical Bayes estimator for $\boldsymbol{b}_i$. The output of model fit is stored in `fit.reml`. We can obtain the predicted values as follows.

```
bhat <- random.effects(fit.reml)
head(bhat)
```

```
##       (Intercept)            age
## M16   -0.7832560 -0.088647061
## M05   -1.6453410 -0.008474015
## M02   -1.1523417 -0.023562635
## M11    0.3257711 -0.140519109
## M07   -0.9467040 -0.011926906
## M08    0.3670757 -0.123853840
```

We can use the predicted random effects in various ways.

- Make histograms and scatterplots of the random effects: this often shows individuals who can be regarded as unusual compared to others.

- We can evaluate the normality assumptions by creating normal QQ plots of predicted b_i
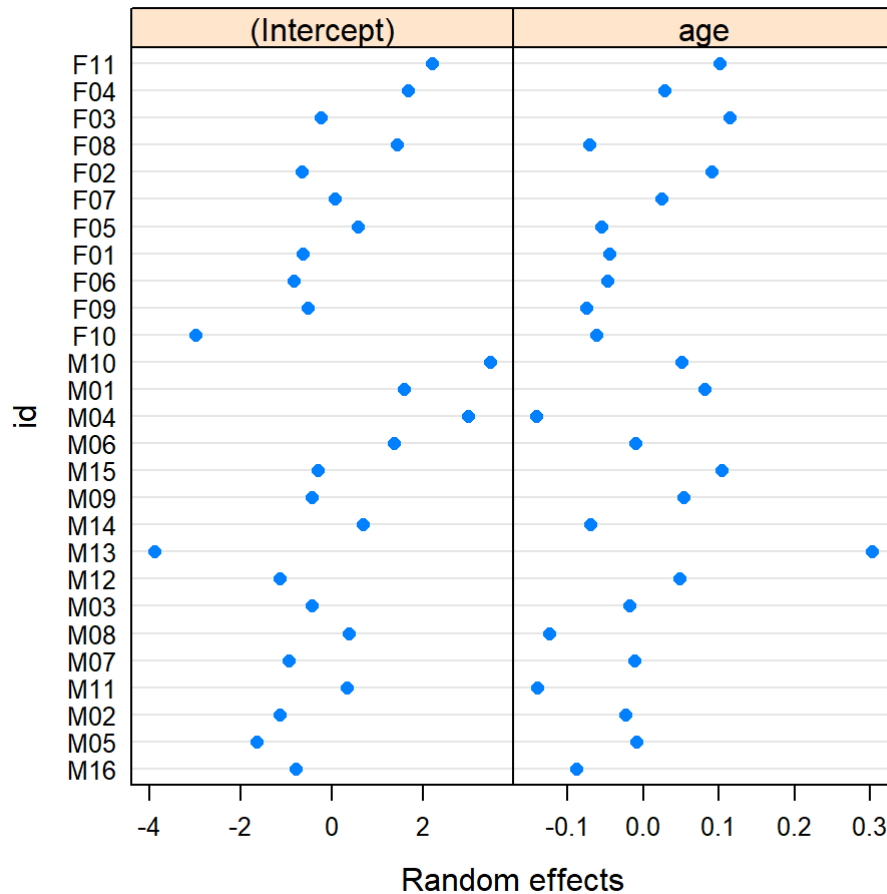
```
## Normal QQ plots
par(mfrow = c(1,2))
qqnorm(bhat[, 1], pch=19, main = "Intercept")
qqline(bhat[, 1])
qqnorm(bhat[, 2], pch=19, main = "Age")
qqline(bhat[, 2])
```
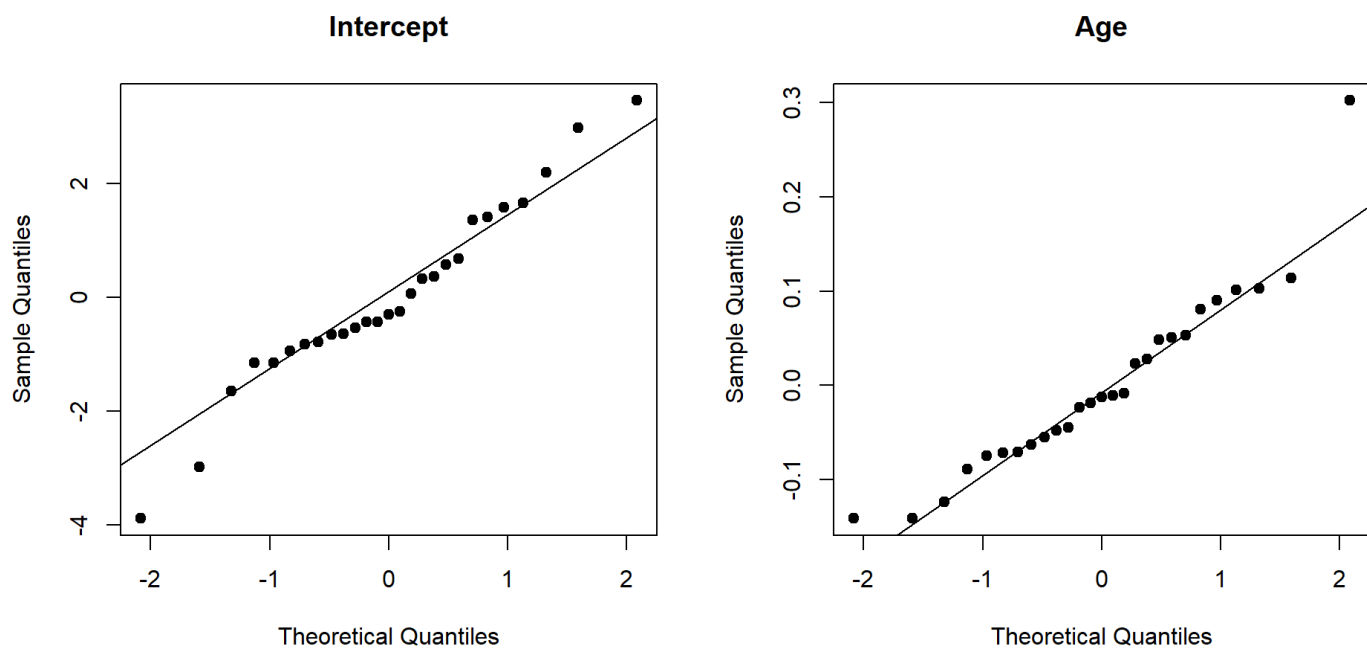


```
## Scatterplots
plot(bhat[,1], bhat[,2], pch=19, xlab = "Intercept", ylab = "Slope")
```
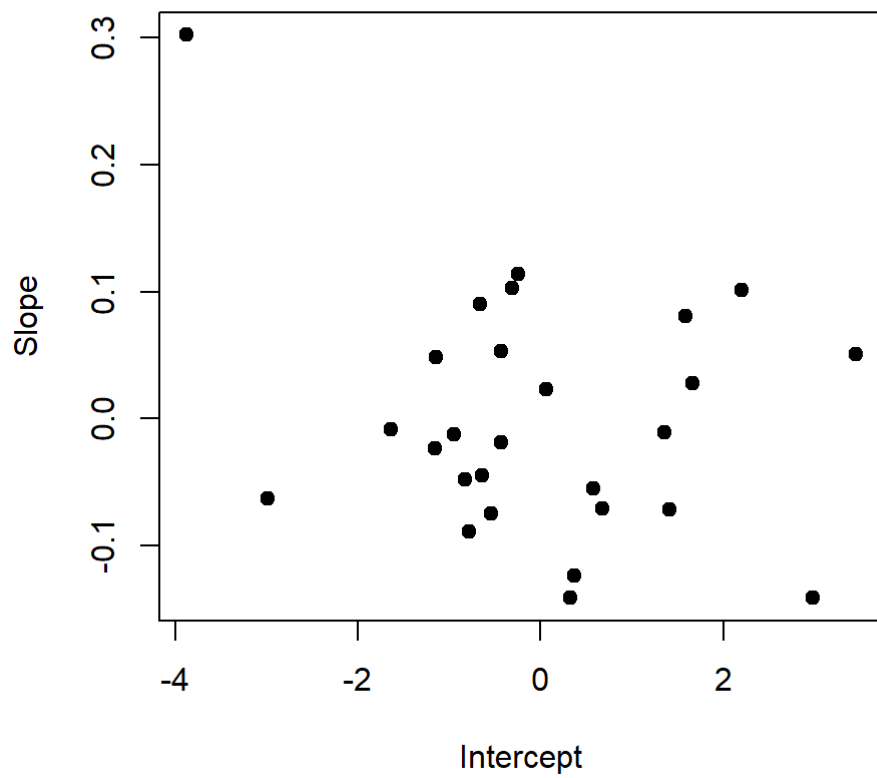
```
## Default plot
plot(bhat)
```

Random effects

Prediction of individual trajectories

Once the subject effects, b_i, are estimated, then the **individual mean can be estimated using the conditional expectation** $E[Y_i|b_i]$; the estimator for this is

$$\widehat{Y}_i = \widehat{E[Y_i|b_i]} = X_i\widehat{\beta} + Z_i\widehat{D}Z_i^T\widehat{\Sigma}_i^{-1}(Y_i - X_i\widehat{\beta})$$

> *With some math, we can show that the subject's predicted response can be represented as a weighted average of the population mean profile and the observed subject's mean profile*

Similarly, **estimated population mean trajectory** (at a specific X_i) is simply

$$\widehat{E[Y_i|b_i]} = X_i\widehat{\beta}.$$

We can obtain both population level and subject level prediction as follows.

```
pred <- fitted(fit.reml,level = 0:1)
```

The argument `level` specifies which type of prediction (population or individual) we want; `level = 0` refers to population level, and `level=1` individual predictions. Setting `level = 0:1` gives both.

```
head(pred)
```

```
##       fixed        id
## 1 22.61563 24.84572
## 2 24.18438 26.57649
## 3 25.75313 28.30725
## 4 27.32188 30.03802
## 5 22.61563 21.27478
## 6 24.18438 22.79641
```

The column `fixed` is the population levels prediction, and the column `id` contains individual level predictions. The predictions are ordered in the same way as in the original data set.
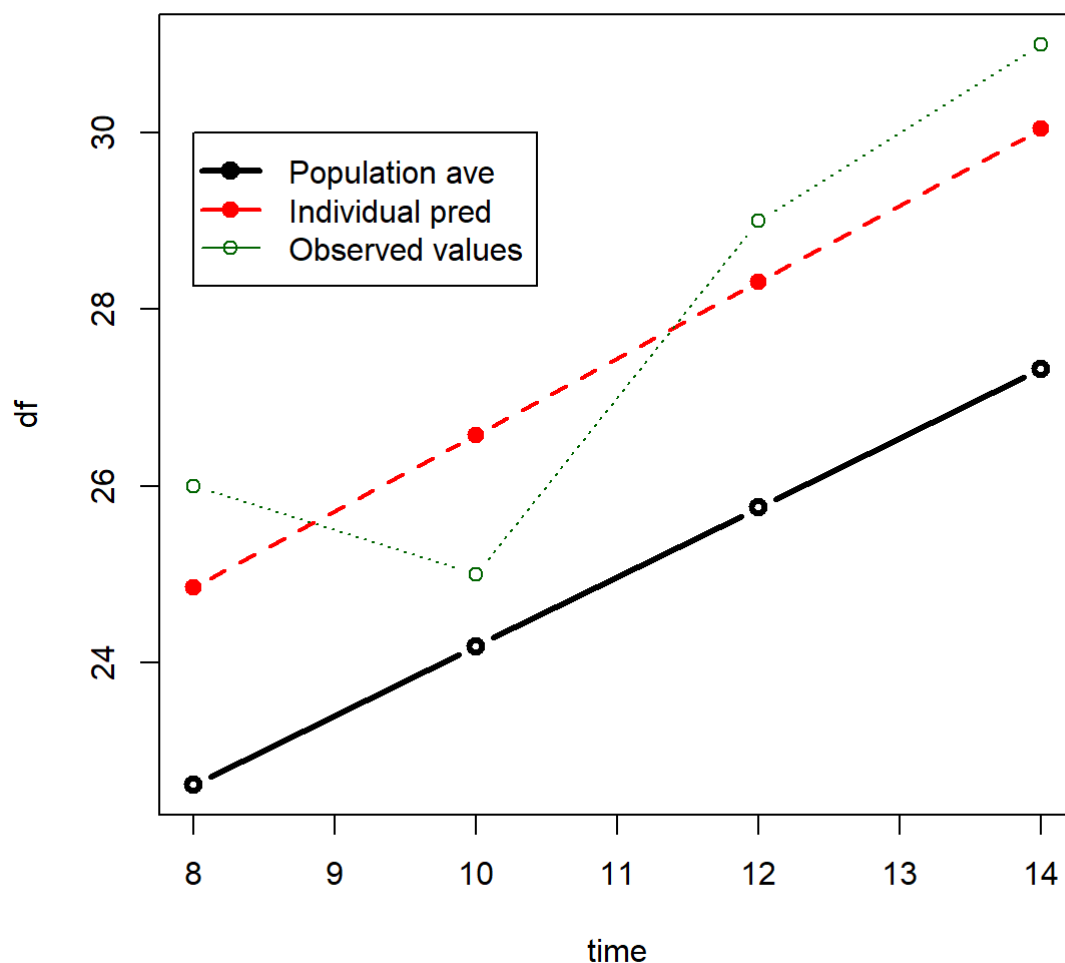
Let us look at prediction of a particular subject.

```
# Look at subject 1
pred.s1 <- pred[1:4, ]

# Observations
Y.s1 <- Orthodont$distance[1:4]

# Age
time <- c(8, 10, 12, 14)

# plot
df <- cbind(pred.s1, Y.s1)
matplot(time, df, type="b", lwd = c(3, 2, 1), pch = c(1, 19, 21),
        col = c("black", "red", "darkgreen"))
legend(8, 30, legend = c("Population ave", "Individual pred", "Observed values"),
       lwd = c(3, 2, 1), pch = c(1, 19, 21),
        col = c("black", "red", "darkgreen"))
```

We can obtain marginal/population average residuals

$$Y_i - X_i\widehat{\beta}$$

and the subject level residuals

$$Y_i - X_i\widehat{\beta} - Z_i\widehat{b}_i$$

as well.

```
res <- residuals(fit.reml,level=0:1)
head(res)
```

```
##        fixed          id
## 1   3.384375   1.1542823
## 2   0.815625  -1.5764860
## 3   3.246875   0.6927458
## 4   3.678125   0.9619775
## 5  -1.115625   0.2252178
## 6  -1.684375  -0.2964070
```

Final remarks

- The REML likelihood provides a measure of the goodness of fit of an assumed model for the covariance. It is possible to compare two nested models the LRT with the correct (asymptotic) null distribution.

- For non-standard covariance comparisons or when the models are non-nested, they can be compared in terms of information criteria (AIC, cAIC, BIC) that effectively penalize the complexity of the model assumed.

Main page: **ST 437/537: Applied Multivariate and Longitudinal Data Analysis (https://maityst537.wordpress.ncsu.edu/)**