Linear Mixed Models

Alexander McLain

Contents

1 Example one: HIV data

The Six Cities Study of Air Pollution and Health example (see the first R notes for details).

```
Six_cities <- read.csv("Six_cities.csv", header = TRUE)
tail(Six_cities,8)
```

```
##       ID Height     Age INI_Height INI_Age Log_FEV1
## 1987 299   1.64 17.9904       1.57 12.9555  1.09527
## 1988 300   1.44 11.9617       1.44 11.9617  0.68310
## 1989 300   1.50 12.9993       1.44 11.9617  0.85015
## 1990 300   1.57 13.9055       1.44 11.9617  0.81536
## 1991 300   1.61 14.9596       1.44 11.9617  1.11841
## 1992 300   1.62 15.9398       1.44 11.9617  1.08181
## 1993 300   1.62 17.0075       1.44 11.9617  1.12817
## 1994 300   1.63 17.8645       1.44 11.9617  1.16938
```

Recall that we're looking atthe data by Age and by Height. Where height has a clear strong linear relationship.

Recall that this is unbalenced data, which is perfect for fitting with linear mixed effect models. We're going to consider log transformed age again. However, here we're going to focus on the initial impact of height and age along with the impact of time-varying height and age. As a result, I'm going to subtract the initial height and ages from the time-varying height and age to improve interpretation.

```
library(tidyverse)
Six_cities <- Six_cities %>% mutate( ID = as.factor(ID),
                                     log_Age = log( Age - INI_Age + 1 ),
                                     INI_log_Age = log( INI_Age + 1 ),
                                     H_minus_base = Height - INI_Height,
                                     A_minus_base = Age - INI_Age)

library(lme4)
library(lmerTest) # Added to get p-values
```

```
# First model with initial height, height, initial age and age with a random intercept
LMM_formula <- Log_FEV1 ~ INI_Height + H_minus_base + INI_Age + A_minus_base + (1|ID)

LMM_int_all <- lmer( formula = LMM_formula , data = Six_cities)
summary(LMM_int_all)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: LMM_formula
##    Data: Six_cities
##
## REML criterion at convergence: -4477.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.8804 -0.5239  0.0712  0.5954  2.8286
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  ID       (Intercept) 0.011014 0.10495
##  Residual             0.004054 0.06367
## Number of obs: 1994, groups:  ID, 300
##
## Fixed effects:
##                 Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  -2.044e+00  1.046e-01  2.828e+02 -19.546   <2e-16 ***
## INI_Height    1.892e+00  1.184e-01  2.784e+02  15.971   <2e-16 ***
## H_minus_base  1.603e+00  3.099e-02  1.739e+03  51.711   <2e-16 ***
## INI_Age      -2.238e-04  8.207e-03  2.813e+02  -0.027    0.978
## A_minus_base  2.044e-02  1.344e-03  1.728e+03  15.211   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) INI_Hg H_mns_ INI_Ag
## INI_Height  -0.926
## H_minus_bas -0.053  0.022
## INI_Age      0.547 -0.821  0.020
## A_minus_bas  0.039 -0.023 -0.938 -0.005
```

```r
VarCorr(LMM_int_all)
```

```
##  Groups   Name        Std.Dev.
##  ID       (Intercept) 0.104947
##  Residual             0.063667
```

```r
print( VarCorr(LMM_int_all), comp = c("Variance", "Std.Dev."))
```

```
##  Groups   Name        Variance  Std.Dev.
##  ID       (Intercept) 0.0110138 0.104947
##  Residual             0.0040535 0.063667
```

```r
# First model  + random effect of height
LMM_formula <- Log_FEV1 ~ INI_Height +  H_minus_base + INI_Age + A_minus_base +
  (1 + H_minus_base |ID)

LMM_int_heig_all <- lmer( formula = LMM_formula , data = Six_cities)
VarCorr(LMM_int_heig_all)
```

```
##  Groups   Name         Std.Dev. Corr
##  ID       (Intercept)  0.105603
##           H_minus_base 0.191178 -0.189
```

```
##  Residual              0.058237
```

```r
data.frame( VarCorr(LMM_int_heig_all) )
```

```
##        grp        var1       var2         vcov       sdcor
## 1       ID  (Intercept)       <NA>  0.011152030  0.10560317
## 2       ID H_minus_base       <NA>  0.036549125  0.19117826
## 3       ID  (Intercept) H_minus_base -0.003812102 -0.18882047
## 4 Residual        <NA>       <NA>  0.003391518  0.05823674
```

```r
anova(LMM_int_all, LMM_int_heig_all)
```

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

```
## Data: Six_cities
## Models:
## LMM_int_all: LMM_formula
## LMM_int_heig_all: LMM_formula
##                  npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## LMM_int_all         7 -4501.8 -4462.6 2257.9  -4515.8
## LMM_int_heig_all    9 -4602.9 -4552.5 2310.4  -4620.9 105.14  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# First model  + random effect of height (no correlation between RE's)
LMM_formula <- Log_FEV1 ~ INI_Height +  H_minus_base + INI_Age + A_minus_base +
  (1 + H_minus_base ||ID)

LMM_int_heig_all2 <- lmer( formula = LMM_formula , data = Six_cities)
VarCorr(LMM_int_heig_all2)
```

```
##  Groups   Name         Std.Dev.
##  ID       (Intercept)  0.103223
##  ID.1     H_minus_base 0.182315
##  Residual              0.058424
```

```r
anova(LMM_int_heig_all, LMM_int_heig_all2)
```

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

```
## Data: Six_cities
## Models:
## LMM_int_heig_all2: LMM_formula
## LMM_int_heig_all: LMM_formula
##                   npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## LMM_int_heig_all2    8 -4601.6 -4556.9 2308.8  -4617.6
## LMM_int_heig_all     9 -4602.9 -4552.5 2310.4  -4620.9 3.2558  1    0.07117 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# First model  + random effect of Age
LMM_formula <- Log_FEV1 ~ INI_Height + H_minus_base + INI_Age + A_minus_base +
  (1 + A_minus_base|ID)

LMM_int_age_all <- lmer( formula = LMM_formula , data = Six_cities)
VarCorr(LMM_int_age_all)
```

```
##  Groups   Name         Std.Dev.  Corr
##  ID       (Intercept)  0.1044420
```

```
##             A_minus_base 0.0073505 -0.112
##  Residual                0.0591377
```

```r
anova(LMM_int_heig_all, LMM_int_age_all)
```

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

## Data: Six_cities
## Models:
## LMM_int_heig_all: LMM_formula
## LMM_int_age_all: LMM_formula
##                  npar     AIC     BIC logLik deviance Chisq Df Pr(>Chisq)
## LMM_int_heig_all    9 -4602.9 -4552.5 2310.4  -4620.9
## LMM_int_age_all     9 -4580.9 -4530.5 2299.4  -4598.9     0  0
```

```r
# Model with log age effects + random intercept and height
LMM_formula <- Log_FEV1 ~ INI_Height + H_minus_base + INI_log_Age + log_Age +
  (1 + H_minus_base|ID)

LMM_int_heig_logall <- lmer( formula = LMM_formula , data = Six_cities)
VarCorr(LMM_int_heig_logall)
```

```
##  Groups   Name         Std.Dev. Corr
##  ID       (Intercept)  0.10484
##           H_minus_base 0.18370  -0.158
##  Residual              0.06105
```

```r
anova(LMM_int_heig_all, LMM_int_heig_logall)
```

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

## Data: Six_cities
## Models:
## LMM_int_heig_all: LMM_formula
## LMM_int_heig_logall: LMM_formula
##                     npar     AIC     BIC logLik deviance Chisq Df Pr(>Chisq)
## LMM_int_heig_all       9 -4602.9 -4552.5 2310.4  -4620.9
## LMM_int_heig_logall    9 -4462.4 -4412.0 2240.2  -4480.4     0  0
```

```r
# Model with log age effects + random intercept and log age
LMM_formula <- Log_FEV1 ~ INI_Height + H_minus_base + INI_log_Age + log_Age +
  (1 + log_Age|ID)

LMM_int_log_age_all <- lmer( formula = LMM_formula , data = Six_cities)
VarCorr(LMM_int_log_age_all)
```

```
##  Groups   Name        Std.Dev. Corr
##  ID       (Intercept) 0.106802
##           log_Age     0.034594 -0.275
##  Residual             0.060496
```

```r
anova(LMM_int_heig_all, LMM_int_log_age_all)
```

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

## Data: Six_cities
## Models:
## LMM_int_heig_all: LMM_formula
## LMM_int_log_age_all: LMM_formula
```

```
##                       npar     AIC     BIC logLik deviance Chisq Df Pr(>Chisq)
## LMM_int_heig_all       9 -4602.9 -4552.5 2310.4  -4620.9
## LMM_int_log_age_all    9 -4478.1 -4427.7 2248.1  -4496.1     0  0
```

The best model appears to be the following linear mixed model.

$$
\begin{aligned}
\log(FEV_{ij}) &= \beta_0 + \beta_1 H_{i1} + \beta_2 (H_{ij} - H_{i1}) + \beta_3 \log(Age_{i1} + 1) + \beta_4 \log(Age_{ij} - Age_{i1} + 1) \\
&+ b_{i0} + b_{i1}(H_{ij} + H_{i1} + \epsilon_{ij}
\end{aligned}
$$

Here are the final estimates:

```r
summary(LMM_int_heig_all)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: LMM_formula
##    Data: Six_cities
##
## REML criterion at convergence: -4583.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -6.4943 -0.4938  0.0765  0.5696  2.8632
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  ID       (Intercept) 0.011152 0.10560
##           H_minus_base 0.036549 0.19118  -0.19
##  Residual             0.003392 0.05824
## Number of obs: 1994, groups:  ID, 300
##
## Fixed effects:
##                Estimate Std. Error         df t value Pr(>|t|)
## (Intercept)  -2.067e+00  1.035e-01  2.807e+02 -19.965   <2e-16 ***
## INI_Height    1.915e+00  1.173e-01  2.766e+02  16.323   <2e-16 ***
## H_minus_base  1.643e+00  3.279e-02  1.174e+03  50.114   <2e-16 ***
## INI_Age      -1.408e-03  8.121e-03  2.784e+02  -0.173    0.862
## A_minus_base  1.936e-02  1.280e-03  1.629e+03  15.125   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) INI_Hg H_mns_ INI_Ag
## INI_Height  -0.926
## H_minus_bas -0.044  0.012
## INI_Age      0.550 -0.822  0.021
## A_minus_bas  0.031 -0.015 -0.858 -0.009
```

```r
round( confint(LMM_int_heig_all), 3)
```

```
## Computing profile confidence intervals ...
```

```
##                2.5 % 97.5 %
## .sig01         0.095  0.116
## .sig02        -0.370  0.017
```

```
## .sig03        0.159  0.226
## .sigma        0.056  0.060
## (Intercept)  -2.270 -1.865
## INI_Height    1.685  2.145
## H_minus_base  1.578  1.708
## INI_Age      -0.017  0.015
## A_minus_base  0.017  0.022
```

We can get predictions of the random effects and their corresponding coefficients.

```
pred_rand_eff <- ranef(LMM_int_heig_all)$ID
head( pred_rand_eff )
```

```
##   (Intercept) H_minus_base
## 1  0.01857832   0.08783002
## 2  0.14024794  -0.20971401
## 3  0.18653410  -0.29085921
## 4 -0.05998491   0.10260790
## 5 -0.01928736  -0.05736296
## 6  0.02556707  -0.10913229
```

```
invid_coef <- coef(LMM_int_heig_all)$ID
head( invid_coef )
```

```
##   (Intercept) INI_Height H_minus_base      INI_Age A_minus_base
## 1   -2.048598   1.915311     1.731053 -0.001407903    0.0193639
## 2   -1.926929   1.915311     1.433509 -0.001407903    0.0193639
## 3   -1.880643   1.915311     1.352364 -0.001407903    0.0193639
## 4   -2.127162   1.915311     1.745831 -0.001407903    0.0193639
## 5   -2.086464   1.915311     1.585860 -0.001407903    0.0193639
## 6   -2.041610   1.915311     1.534091 -0.001407903    0.0193639
```

2 Example Two: Dental data

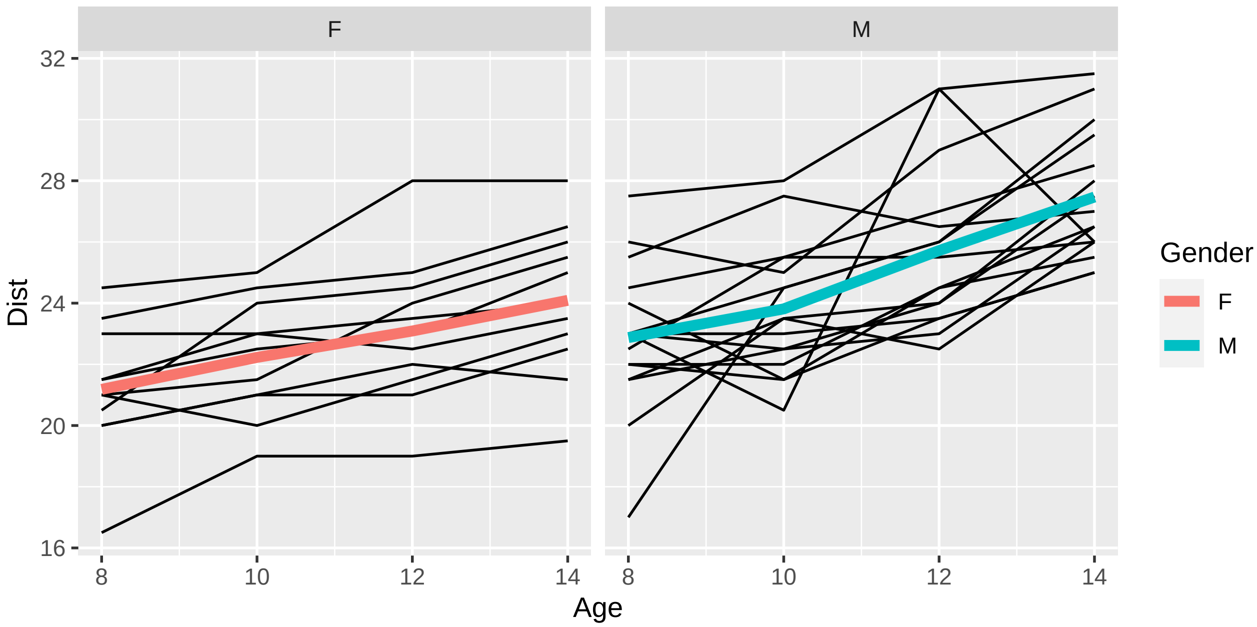
Here, we're going to look at the dental data that you used in your homework.

```
wide_dental <- read.csv("dental.csv",header = TRUE, na.strings = "",
                        stringsAsFactors = FALSE)
long_dental <- pivot_longer(wide_dental, cols = starts_with("Dist"), names_to = "Age",
                            names_prefix = "Dist", values_to = "Dist",
                            values_drop_na = TRUE)
long_dental <- long_dental %>% mutate( Age = as.numeric(Age) )
str(long_dental)
```

```
## tibble [108 x 4] (S3: tbl_df/tbl/data.frame)
##  $ ID    : int [1:108] 1 1 1 1 2 2 2 2 3 3 ...
##  $ Gender: chr [1:108] "F" "F" "F" "F" ...
##  $ Age   : num [1:108] 8 10 12 14 8 10 12 14 8 10 ...
##  $ Dist  : num [1:108] 21 20 21.5 23 21 21.5 24 25.5 20.5 24 ...
```

Now were going to look at the data as a function of age and gender.

```
p <- ggplot(data = long_dental, aes(x = Age, y = Dist, group = ID))
p + geom_line() +
  stat_summary(aes(group = 1, color = Gender), geom = "line",
               fun = mean, size = 2) +
  facet_grid(. ~ Gender)
```



Now we'll look at various models.

```
LMM_formula <- Dist ~ Age + Gender + Age:Gender + (1|ID)
LMM_int <- lmer( formula = LMM_formula , data = long_dental)
LMM_formula <- Dist ~ Age + Gender + Age:Gender + (1 + Age|ID)
LMM_int_slope <- lmer( formula = LMM_formula , data = long_dental)
anova(LMM_int,LMM_int_slope)
```

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

## Data: long_dental
## Models:
## LMM_int: LMM_formula
## LMM_int_slope: LMM_formula
##               npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## LMM_int          6 440.64 456.73 -214.32   428.64
## LMM_int_slope    8 443.81 465.26 -213.90   427.81 0.8331  2     0.6593
```

Now, let's test if there is different random effect variances by group. Do do this we'll add a separate random effect for only females.

```
long_dental <- long_dental %>% mutate( Female = 1*I(Gender == "F") ,
                                       Male = 1*I(Gender == "M") )
## Notice the || between Female and ID.
LMM_formula <- Dist ~ Age + Gender + Age:Gender + (1 + Female || ID)
LMM_int_by_gen <- lmer( formula = LMM_formula , data = long_dental)
anova(LMM_int, LMM_int_by_gen)
```

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

## Data: long_dental
## Models:
## LMM_int: LMM_formula
## LMM_int_by_gen: LMM_formula
##                npar    AIC    BIC  logLik deviance Chisq Df Pr(>Chisq)
## LMM_int           6 440.64 456.73 -214.32   428.64
## LMM_int_by_gen    7 442.43 461.21 -214.22   428.43 0.207  1     0.6491
```

```r
VarCorr(LMM_int_by_gen)
```

```
##  Groups   Name        Std.Dev.
##  ID       (Intercept) 1.6924
##  ID.1     Female      1.0421
##  Residual             1.3864
```

```r
ranef(LMM_int_by_gen)
```

```
## $ID
##    (Intercept)       Female
## 1  -0.82277256 -0.311925762
## 2   0.22773169  0.086336595
## 3   0.71257980  0.270149990
## 4   1.43985198  0.545870084
## 5  -0.01469237 -0.005570103
## 6  -0.98438860 -0.373196894
## 7   0.22773169  0.086336595
## 8   0.47015575  0.178243293
## 9  -0.98438860 -0.373196894
## 10 -2.68135700 -1.016543778
## 11  2.40954821  0.913496875
## 12  2.38169502  0.000000000
## 13 -1.36479153  0.000000000
## 14 -0.61549422  0.000000000
## 15  1.41831277  0.000000000
## 16 -1.68591895  0.000000000
## 17  1.20422782  0.000000000
## 18 -1.04366411  0.000000000
## 19 -0.93662164  0.000000000
## 20  0.13380309  0.000000000
## 21  3.88028965  0.000000000
## 22 -1.15070658  0.000000000
## 23 -0.61549422  0.000000000
## 24 -0.61549422  0.000000000
## 25 -0.08028185  0.000000000
## 26  0.77605793  0.000000000
## 27 -1.68591895  0.000000000
##
## with conditional variances for "ID"
```

The fact that we did an extra **females** group was important. When checking for different random effect variances using the above method the extra group can only **add** variance.

Notice what happens when we do it with males:

```r
LMM_formula <- Dist ~ Age + Gender + Age:Gender + (1 + Male || ID)
LMM_int_by_gen_M <- lmer( formula = LMM_formula , data = long_dental)
```

```
## boundary (singular) fit: see ?isSingular
```

```r
VarCorr(LMM_int_by_gen_M)
```

```
##  Groups   Name        Std.Dev.
##  ID       (Intercept) 1.8162e+00
##  ID.1     Male        6.2703e-05
##  Residual             1.3864e+00
```

Here's another way:

```
LMM_formula <- Dist ~ Age + Gender + Age:Gender + (0 + Gender || ID)
LMM_int_by_gen2 <- lmer( formula = LMM_formula , data = long_dental)
VarCorr(LMM_int_by_gen2)
```

```
##  Groups   Name    Std.Dev. Corr
##  ID       GenderF 1.9874
##           GenderM 1.6924   0.065
##  Residual         1.3864
```

This was really doesn't make sense. There can be any correlation between these random effects!

Now, we'll go deeper down the rabbit hole and look to see if there's a difference between the variance of the random effect of age by gender

```
LMM_formula <- Dist ~ Age + Gender + Age:Gender + (1 + Female + Age + Age*Female || ID)
LMM_int_age_by_gen <- lmer( formula = LMM_formula , data = long_dental)
anova(LMM_int, LMM_int_age_by_gen)
```

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

```
## Data: long_dental
## Models:
## LMM_int: LMM_formula
## LMM_int_age_by_gen: LMM_formula
##                    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## LMM_int               6 440.64 456.73 -214.32   428.64
## LMM_int_age_by_gen    9 445.96 470.10 -213.98   427.96 0.6798  3     0.8779
```

```
VarCorr(LMM_int_age_by_gen)
```

```
##  Groups   Name        Std.Dev.
##  ID       (Intercept) 1.5124582
##  ID.1     Female      0.0038062
##  ID.2     Age         0.0758460
##  ID.3     Female:Age  0.0809161
##  Residual             1.3679902
```

Still nothing there.