Multilevel Modeling

Answer Key: Homework 4

Andy Stone

October 15, 2016

13.2

a.

Let $i = 1, 2, ..., 300$ index each rating (since we have a committee of 10 people who review 30 applications each, we have a total of 300 ratings). Let $k = 1, 2, ..., 100$ index the applicants, and $j = 1, 2, ..., 10$ index the raters. Then, our model is the following:

$$\begin{aligned} rating_i &\sim \mathcal{N}(\mu + \gamma_{j[i]} + \delta_{k[i]}, \sigma_y^2) \text{ for } i = 1, ..., 300 \\ \gamma_j &\sim \mathcal{N}(0, \sigma_\gamma^2) \text{ for } j = 1, ..., 10 \\ \delta_k &\sim \mathcal{N}(0, \sigma_\delta^2) \text{ for } k = 1, ..., 100 \end{aligned}$$

b.

Now, we specify the same model, but allow σ_y^2 to vary by rater. We can draw this from a scaled inverse χ^2 distribution (or any other appropriate distribution for drawing a variance, i.e., those distributions constrained to be positive). Our model is:

$$\begin{aligned} rating_i &\sim \mathcal{N}(\mu + \gamma_{j[i]} + \delta_{k[i]}, \sigma_{j[i]}^2) \text{ for } i = 1, ..., 300 \\ \gamma_j &\sim \mathcal{N}(0, \sigma_\gamma^2) \text{ for } j = 1, ..., 10 \\ \delta_k &\sim \mathcal{N}(0, \sigma_\delta^2) \text{ for } k = 1, ..., 100 \\ \sigma_j^2 &\sim \text{Scale-inverse-}\chi^2(\nu, \tau^2) \end{aligned}$$

13.4

In this problem, we have 100 observations y_i of how far off group j's guess was of individual k's true age, that is, the absolute value of the difference between a guess and the true age. We have 10 individuals k and 10 groups j. We can first write out a non-nested model to this data. In doing so, we would like to have separate coefficients for each individual k and for each group j, and to allow for a separate error variance for each group. We can write such a model as follows:

$$\begin{aligned} y_i &\sim \mathcal{N}(\alpha + \gamma_{j[i]} + \delta_{k[i]}, \sigma_{j[i]}^2) \text{ for } i = 1, ..., 100 \\ \gamma_j &\sim \mathcal{N}(0, \sigma_\gamma^2) \text{ for } j = 1, ..., 10 \\ \delta_k &\sim \mathcal{N}(0, \sigma_\delta^2) \text{ for } k = 1, ..., 10 \\ \sigma_j^2 &\sim \text{Scale-inverse-}\chi^2(\nu, \tau^2) \end{aligned}$$

We can't quite fit this model in `lmer()`, as we can't specify the different variance for each group. But, assuming constant group-level variance, we can fit the model:

```
# Loading the data from a csv
age.data <- read.csv("age.guessing.csv")

# Empty matrix to put observations into
analysis.matrix <- matrix(NA, nrow=100, ncol=3)

ages <- c()
group <- c()
person <- rep(c(1:10), times=10) # Creating individual ID variable

# For loop creates the (non-abs value) dependent variable and the group ID variable
for(i in 1:10){
  ages <- c(ages, as.integer(age.data[i,3:12]))
  group <- c(group, rep(age.data[i,1], times=10))
}

# Adding the variables to the matrix
analysis.matrix[,1] <- ages
analysis.matrix[,2] <- group
analysis.matrix[,3] <- person
# Turning the matrix into a data frame
model.data <- data.frame(analysis.matrix)
# Giving the variables in the data frame names
colnames(model.data) <- c("error", "group.id", "person.id")

# Turning the group and individual ID variables into factors
model.data$group.id <- as.factor(model.data$group.id)
model.data$person.id <- as.factor(model.data$person.id)
# Making the true DV by taking the absolute value
model.data$error <- abs(model.data$error)

# Multilevel model with separate coefficinets for each group and individual
age.model <- lmer(error ~ 1 + (1 | group.id) + (1 | person.id), data=model.data)
summary(age.model)
Linear mixed model fit by REML ['lmerMod']
Formula: error ~ 1 + (1 | group.id) + (1 | person.id)
   Data: model.data

REML criterion at convergence: 545.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8256 -0.5603 -0.1148  0.6566  3.8882

Random effects:
 Groups    Name        Variance Std.Dev.
 group.id  (Intercept)  0.2002  0.4475
 person.id (Intercept) 10.9625  3.3110
 Residual              10.9320  3.3064
Number of obs: 100, groups:  group.id, 10; person.id, 10

Fixed effects:
            Estimate Std. Error t value
```

```
(Intercept)    5.470      1.107   4.941
```

We see that the estimated residual standard deviations are $\hat{\sigma}_y = 3.306$ across the dependent variable, $\hat{\sigma}_\delta = 3.311$ across individuals, and $\hat{\sigma}_\gamma = 0.448$ across rating groups. So, the variation across individuals in how accurately they are rated is much higher than the variation among groups. This makes sense, as it is hard to tell how old some people are, but is reasonable to expect groups of students to exhibit generally similar behavior in how well they can guess ages.

13.5

a.

Let us return to the `CD4` data, which allows us to model the trend of childrens' square root CD4 level, which captures the intensity of a child's HIV infection, over time. We wish to extend our model from 12.2b – which modeled the square root CD4 level as a function of the observation-level variable `time` and the group (that is, child) level variables `treatmnt` and `baseage` – to allow for varying slopes for the `time` predictor. In this model, we already had allowed for a varying intercept for each child, and now we want allow for a varying slope for `time` for each child. We can specify such a model as follows:

$$rootCD4_i \sim \mathcal{N}(\alpha_{j[i]} + \beta_{j[i]} time_i, \sigma_y^2)$$

$$\begin{pmatrix} \alpha_j \\ \beta_j \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \gamma_0^\alpha + \gamma_1^\alpha treatmnt_j + \gamma_2^\alpha baseage_j \\ \gamma_0^\beta + \gamma_1^\beta treatmnt_j + \gamma_2^\beta baseage_j \end{pmatrix}, \begin{pmatrix} \sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta \\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2 \end{pmatrix} \right) \text{ for } j = 1, ..., J$$

This model has the intercept and slope on time varying by group (thus the subscript $j[i]$); models these group-level effects as a function of the group-level predictors `treatmnt` and `baseage`; and allows there to be a correlation between the time and intercept parameters. We can also write this model in traditional linear equation form:

$$rootCD4_i = \alpha_{j[i]} + \beta_{j[i]} time_i + \epsilon_i$$
$$\alpha_j = \gamma_0^\alpha + \gamma_1^\alpha treatmnt_j + \gamma_2^\alpha baseage_j + \eta_j^\alpha$$
$$\beta_j = \gamma_0^\beta + \gamma_1^\beta treatmnt_j + \gamma_2^\beta baseage_j + \eta_j^\beta$$

This can also be re-expressed as a single model by plugging in the formulas for α_j and β_j into the equation for $rootCD4_i$:

$$rootCD4_i = [\gamma_0^\alpha + \gamma_1^\alpha treatmnt_{j[i]} + \gamma_2^\alpha baseage_{j[i]} + \eta_{j[i]}^\alpha] + [\gamma_0^\beta + \gamma_1^\beta treatmnt_{j[i]} + \gamma_2^\beta baseage_{j[i]} + \eta_{j[i]}^\beta] time_i + \epsilon_i$$

$$rootCD4_i = \gamma_0^\alpha + \gamma_1^\alpha treatmnt_{j[i]} + \gamma_2^\alpha baseage_{j[i]} + \eta_{j[i]}^\alpha + \gamma_0^\beta * time_i + \gamma_1^\beta treatmnt_{j[i]} * time_i + \gamma_2^\beta baseage_{j[i]} * time_i + \eta_{j[i]}^\beta * time_i + \epsilon_i$$

Note the interactions between treatment and time and baseage and time. These must be specified in our `lmer()` call. Our model is run in `R` as follows:

```r
# Loading HIV data
hiv.data <- read.csv("allvar.csv", header=TRUE)
# Square root transformation of the CD4PCT
hiv.data$rootCD4 <- sqrt(hiv.data$CD4PCT)
# Creation of time variable
hiv.data$time <- hiv.data$visage - hiv.data$baseage

# Removing those cases that have NAs for the DV or the time variable
data.noNA.CD4 <- hiv.data[complete.cases(hiv.data[,4]),]
data.noNA.CD4 <- data.noNA.CD4[complete.cases(data.noNA.CD4[,11]),]

# Creating indicators for each of the children
inddummies <- as.factor(data.noNA.CD4$newpid)

# Extended model, varying intercept and slope on time
extend.mod <- lmer(rootCD4 ~ time + treatmnt + baseage + time:treatmnt + time:baseage +
                     (1 + time | inddummies), data=data.noNA.CD4)
summary(extend.mod)
Linear mixed model fit by REML ['lmerMod']
Formula:
rootCD4 ~ time + treatmnt + baseage + time:treatmnt + time:baseage +
    (1 + time | inddummies)
   Data: data.noNA.CD4

REML criterion at convergence: 3113.6

Scaled residuals:
    Min      1Q  Median      3Q     Max
-5.0973 -0.4003  0.0224  0.4025  5.0121

Random effects:
 Groups     Name        Variance Std.Dev. Corr
 inddummies (Intercept) 1.8474   1.3592
            time        0.3412   0.5841   -0.05
 Residual               0.5151   0.7177
Number of obs: 1072, groups:  inddummies, 250

Fixed effects:
              Estimate Std. Error t value
(Intercept)    5.00817    0.32297  15.506
time          -0.53777    0.24083  -2.233
treatmnt       0.13061    0.18676   0.699
baseage       -0.12894    0.04095  -3.149
time:treatmnt  0.09039    0.13600   0.665
time:baseage   0.01489    0.03071   0.485

Correlation of Fixed Effects:
            (Intr) time   trtmnt baseag tm:trt
time        -0.245
treatmnt    -0.853  0.204
baseage     -0.431  0.113 -0.004
time:trtmnt  0.209 -0.843 -0.241 -0.004
time:baseag  0.113 -0.487 -0.004 -0.247  0.035
```

b.

Now, we can move to fitting a model that does not allow for varying slopes but does allow there to be a different coefficient for each time point. This can be done by converting the `time` variable into a set of dummy variables. I round each time observation to one decimal place, so there are only 19 different time observations. Then, we can include 18 dummy variables (omitting one as the baseline group, meaning its coefficient will be equal to zero) in the regression to estimate the coefficient for each rounded time point. In linear form, such a model would appear as follows:

$$rootCD4_i = \alpha_{j[i]} + \beta_1 time1_i + \beta_2 time2_i + \cdots + \beta_{18} time18_i + \epsilon_i$$

$$\alpha_j = \gamma_0^\alpha + \gamma_1^\alpha treatmnt_j + \gamma_2^\alpha baseage_j + \eta_j^\alpha$$

This model still allows the intercept to vary by group, as a function of group-level predictors `treatmnt` and `baseage`, while allowing us to estimate a different coefficient for each of the rounded time variables (with, in this specification, the 19th time variable group as the omitted category, with a coefficient of zero).

I reproduce output from this model in the summary below. We see that our coefficient estimates for the time variable are all generally negative (as expected as the baseline omitted group is `timeround = 0`), but do exhibit substantial variation.

```
# Rounding the time variable to 1 decimal place
data.noNA.CD4$timeround <- round(data.noNA.CD4$time, 1)
data.noNA.CD4$timeround <- as.factor(data.noNA.CD4$timeround)

# Varying intercept, allowing for different coefficients for each rounded time point
difcoef.mod <- lmer(rootCD4 ~ timeround + treatmnt + baseage + (1 | inddummies), data=data.noNA.CD4)
summary(difcoef.mod)
Linear mixed model fit by REML ['lmerMod']
Formula: rootCD4 ~ timeround + treatmnt + baseage + (1 | inddummies)
   Data: data.noNA.CD4

REML criterion at convergence: 3152.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.6940 -0.4470  0.0128  0.4458  4.8441

Random effects:
 Groups     Name        Variance Std.Dev.
 inddummies (Intercept) 1.8876   1.374
 Residual               0.6037   0.777
Number of obs: 1072, groups:  inddummies, 250

Fixed effects:
             Estimate Std. Error t value
(Intercept)   4.92482    0.31883  15.447
timeround0.2 -0.13066    0.08699  -1.502
timeround0.3  0.02489    0.16091   0.155
timeround0.4 -0.09644    0.29238  -0.330
timeround0.5 -0.19846    0.08532  -2.326
timeround0.6 -0.14598    0.26634  -0.548
timeround0.7 -0.29768    0.09174  -3.245
timeround0.8 -0.34632    0.17790  -1.947
timeround0.9 -0.24371    0.10891  -2.238
timeround1   -0.45368    0.12850  -3.531
timeround1.1 -0.38336    0.17306  -2.215
timeround1.2 -0.49534    0.10807  -4.583
timeround1.3 -0.54660    0.25078  -2.180
timeround1.4 -0.55772    0.12260  -4.549
timeround1.5 -0.47650    0.16320  -2.920
timeround1.6  0.19355    0.43212   0.448
timeround1.7 -0.47188    0.61123  -0.772
```

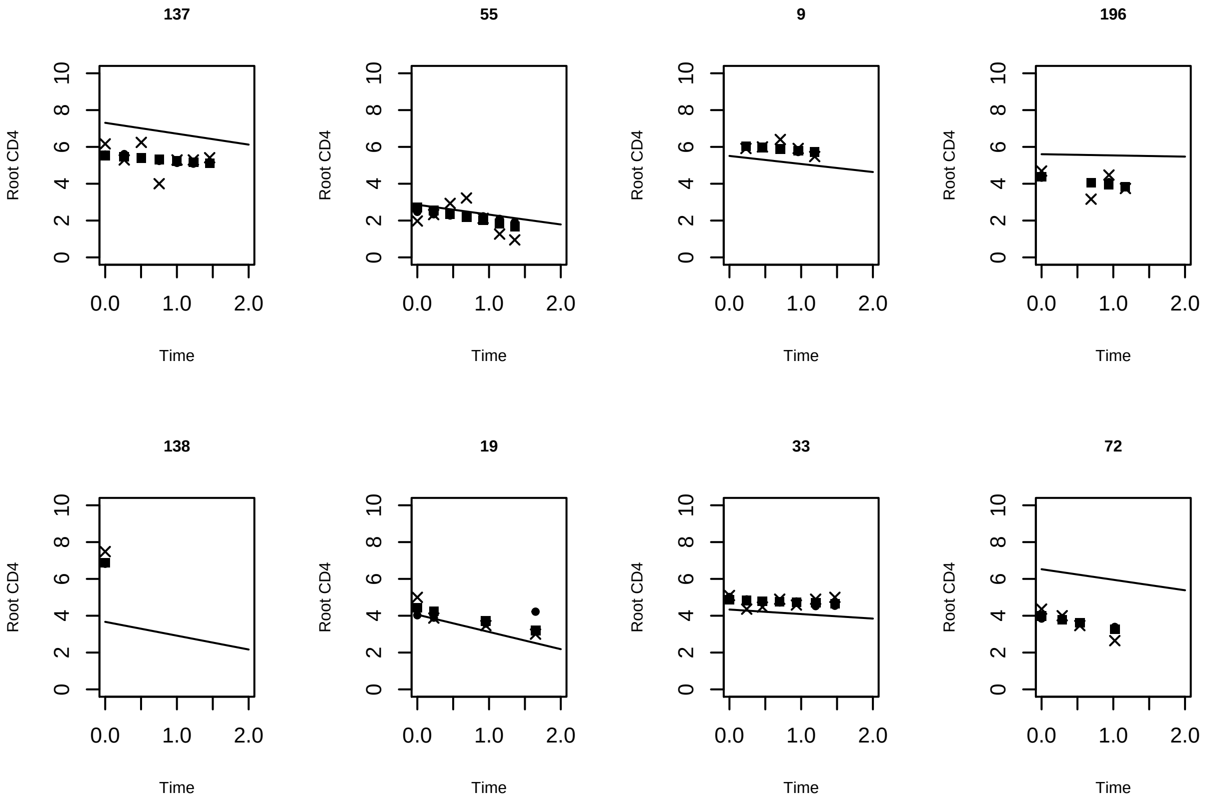
```
timeround1.8  0.07672    0.84920   0.090
timeround1.9 -0.85451    0.50270  -1.700
treatmnt      0.17112    0.18272   0.937
baseage      -0.11835    0.04003  -2.957
```

c.

To compare the results from the varying-intercept, varying-slope model in part (a) and the varying-intercept, factored time variable model in part (b), we can turn to both numerical and graphical methods. Firstly, we can compare the variances of the individual-level (that is, each hospital visit) data ($\hat{\sigma_y}^2$) and of the group-level (that is, each child) data ($\hat{\sigma_\alpha}^2$) for each regression. In the model from part (a), which allows the slope on time to vary by group, we have $\hat{\sigma_y}^2 = 0.515$ and $\hat{\sigma_\alpha}^2 = 1.847$. In the model from part (b), which estimates separate coefficients for each (rounded) time variable, we have $\hat{\sigma_y}^2 = 0.604$ and $\hat{\sigma_\alpha}^2 = 1.888$. We see, then, that by allowing the slope on time to vary by group, we are able to reduce the unexplained variation across observations as well as slightly reduce the amount of unexplaned variation between groups. This suggests that there are slightly different time trends in the reduction of rootCD4 across groups, which is captured by our model in (a) and helps us to explain more variance than just allowing the intercept to vary by group (the model in (b) does this with the dummies for the group).

Graphically, we can turn to plotting the linear fit of the time trend of `rootCD4` predicted by each model. We can then attempt to assess which model fits the data better. I decide to take a random sample of 8 of the cases from the dataset to do so. Then, I plot the actual observations (X marks); the fitted values from the model in part (a) (boxes); the fitted values from the model in part (b) (circles); and the fitted line from the model in part (a). An analysis of these plots suggests that, substantively, there are not too great of differences between the fitted values predicted by either model. There are no real clear graphical trends that one model does a better job of predicting the true `rootCD4` values. Nonetheless, the numerical methods above suggest that the model in part (a) does a better job of modeling the temporal patterns of `rootCD4` levels.

```r
# Code for figures displaying two models' fit
set.seed(55)
display <- as.numeric(sample(unique(inddummies), size=8)) # Sample 8 observations
par(mfrow=c(2,4))
for (i in display){
    age <- unique(data.noNA.CD4$baseage[which(data.noNA.CD4$newpid == i)])
    treat <- unique(data.noNA.CD4$treatmnt[which(data.noNA.CD4$newpid == i)])
    plot(rootCD4[which(data.noNA.CD4$newpid == i)] ~ time[which(data.noNA.CD4$newpid == i)],
        data=data.noNA.CD4, ylim=c(0,10), xlim=c(0, 2), xlab="Time", ylab="Root CD4", main=i,
        cex.lab=0.7, cex.main=0.7, pch=4)
    curve(coef(extend.mod)$inddummies[i,1] + coef(extend.mod)$inddummies[i,2]*x +
            coef(extend.mod)$inddummies[i,3]*data.noNA.CD4$treatmnt[i] +
            coef(extend.mod)$inddummies[i,4]*data.noNA.CD4$baseage[i], add=TRUE)
    points(data.noNA.CD4$time[which(data.noNA.CD4$newpid == i)],
           fitted(extend.mod)[which(data.noNA.CD4$newpid == i)], pch=15)
    points(data.noNA.CD4$time[which(data.noNA.CD4$newpid == i)],
           fitted(difcoef.mod)[which(data.noNA.CD4$newpid == i)], pch=20)
}
```


137
55
9
196
138
19
33
72
Root CD4
0
2
4
6
8
10
Time
0.0
1.0
2.0