Statistical Modeling Course

Multi-level Modeling Assignment

In this lab we will use the musicdata.csv dataset to develop a deeper understanding of multi-level (mixed effect) models.

Objective: To examine models for predicting the happiness of musicians prior to performances, as measured by the positive affect scale from the PANAS (Positive Affect Negative Affect Schedule) instrument, pa.

The dataset contains the following variable

Variables in original data set

- id: unique musician identification number
- diary: cumulative total of diaries filled out by musician
- previous: number of previous diary entries filled out
- perform_type: type of performance (solo, large or small ensemble)
- memory: performed from Memory, using Score, or Unspecified
- audience: who attended (Instructor, Public, Students, or Juried)
- pa: positive affect from PANAS
- na: negative affect from PANAS
- age: musician age
- gender: musician gender
- instrument: Voice, Orchestral, or Piano
- years_study: number of years studied the instrument
- mpqsr: stress reaction subscale from MPQ
- mpqab: absorption subscale from MPQ
- mpqpem: positive emotionality composite scale from MPQ
- mpqnem: negative emotionality composite scale from MPQ
- mpqcon: constraint composite scale from MPQ

```r
music <- read.csv("musicdata.csv")
music <- music %>% mutate(solo = ifelse(perform_type == "Solo", 1, 0))
```

Problem 1

In this dataset the group is the musician and the unit is the performance. Classify the predictors into unit-level and group-level.

The unit-level predictors (observed per performance) are:

- diary
- perform_type
- memory
- audience
- pa
- na

The group-level predictors are:

- gender
- mpqsr
- mpqab
- mpqpem
- mpqnem
- mpqcon
- instrument
- years_study
- previous

Problem 2

What is the max, min, and median number of diary entries for the musicians?

```
tibble("min"= min(music$diary),
       "max"=max(music$diary),
       "median"=median(music$diary)
)
```

```
## # A tibble: 1 x 3
##     min   max median
##   <int> <int>  <int>
## 1     1    15      8
```

Problem 3

Write the equations for the model that predicts positive affect, `pa`, with a random intercept term and no predictors. Clearly define all of the terms. Fit this model. What is the estimated mean positive affect across all diary entries and musicians? Use this model to calculate the intraclass correlation coefficient. Interpret this value.

The multilevel model for the varying intercept model is

$$y_i = \alpha_{j[i]} + \epsilon_i$$

$$\alpha_j \sim N(\mu_a, \sigma_a^2)$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

where

- y_i is the positive affect measurement for performance i
- $\alpha_{j[i]}$ is the random effect of individual j, this is also the mean positive affect in all performaces of individual j
- μ_a is the mean of the positive affect measurement of all performaces across all individuals
- σ_α^2 is the variance between individuals
- σ_y^2 is the variance within performances of an individual

```
int_only <- lmer(pa~1 + (1|id), data=music)
summary(int_only)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: pa ~ 1 + (1 | id)
##    Data: music
##
## REML criterion at convergence: 3340.2
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -3.12392 -0.64454  0.02559  0.64814  2.79434
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 23.72    4.871
##  Residual             41.70    6.457
## Number of obs: 497, groups:  id, 37
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  32.5622     0.8584   37.93
```

```
#getting the intraclass correlation coefficient
sigmas <- arm::sigma.hat(int_only)$sigma
icc <- sigmas$data/(sigmas$data + sigmas$id)
icc
```

```
## (Intercept)
##   0.5700409
```

The estimated mean positive affect across all diary entries and musicians is 32.5622.

Having an intraclass correlation coefficient of 0.57 means that 57% of the total variability in positive affect in performances are attributable to differences among performers.

Problem 4

Building on the model from the previous problem, include audience type (`audience`), performing solo (`solo`) and (`years_study`) in your model as fixed effects. Write the equation for this model. Fit the model and interpret the estimates.

The multilevel model for varying intercept but fixed predictors is:

$$\begin{aligned} y_i = \alpha_{j[i]} + \beta_1 (audience_Juried)_i + \beta_2 (audience\_Public)_i \\ + \beta_3 (audience\_Student)_i + \beta_4 (solo)_i + \beta_5 (years\_study)_i + \epsilon_i \end{aligned}$$

$$\alpha_j \sim N(\mu_a, \sigma_a^2)$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

where

- y_i is the positive affect measurement for performance i
- $\alpha_{j[i]}$ is the random effect of individual j, this is also the mean positive affect in all performaces with an Instructor audience, group performances, and 0 years of study in the instrument for individual j
- β_1 is the average difference in positive affect measurement for Juried Recitals
- β_2 is the average difference in positive affect measurement for Public Performances
- β_3 is the average difference in positive affect measurement for performances with student audience
- β_4 is the average difference in positive affect measurement for solo performances and those that are not solo
- β_5 is the average difference in positive affect measurement for each unit increase in years_study
- μ_a is the mean of the positive affect measurement of all performaces across all individuals
- σ^2_α is the variance between individuals
- σ^2_y is the variance within performances of an individual

```
mod2 <- lmer(pa~1 + audience + solo + years_study + (1|id), data=music)
summary(mod2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: pa ~ 1 + audience + solo + years_study + (1 | id)
##    Data: music
##
## REML criterion at convergence: 3288.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4259 -0.6207 -0.0235  0.6247  2.4753
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  id       (Intercept) 20.26    4.501
##  Residual             38.48    6.203
## Number of obs: 497, groups:  id, 37
##
## Fixed effects:
##                            Estimate Std. Error t value
## (Intercept)                 32.5354     2.0425  15.929
## audienceJuried Recital       6.3359     1.1205   5.655
## audiencePublic Performance   2.9928     0.9651   3.101
## audienceStudent(s)           0.1305     0.8836   0.148
## solo                        -0.6378     0.8376  -0.761
## years_study                 -0.1806     0.1980  -0.912
##
## Correlation of Fixed Effects:
##             (Intr) adncJR adncPP adnS() solo
## adncJrdRctl -0.169
## adncPblcPrf -0.436  0.314
## adncStdnt() -0.240  0.305  0.518
## solo        -0.446  0.040  0.627  0.248
## years_study -0.811  0.028  0.054 -0.018  0.092
```

The results of the model show that the mean positive affect in all performaces with an Instructor audience is 32.5354. There is a 6.33 increase to this for Juried Recitals, 2.3 increase for public performances, 0.13 increase for student audiences. The positive affect is lower by 0.64 for solo performances. Suprisingly, the positive affect decreases by 0.18 for a year increase in the study of an instrument.

Problem 5

Fit the model in the previous problem but now allow the effect of performing solo to vary by musician (random slopes). Write the equation for this model. What are the estimates for the mean effect of solo and the variance of the effect of solo.

The multilevel model for varying intercept and varying slope for the variable "solo" is:

$$y_i = \alpha_{j[i]} + \beta_1(audience\_Juried)_i + \beta_2(audience\_Public)_i +$$
$$\beta_3(audience\_Student)_i + \beta_{4[j]i}(solo)_i + \beta_5(years\_study)_i + \epsilon_i$$

$$\alpha_j \sim N(\mu_a, \sigma_a^2)$$
$$\epsilon_i \sim N(0, \sigma_y^2)$$
$$\beta_i \sim N(\mu_{\beta_4}, \sigma_{\beta_4}^2)$$

where

- y_i is the positive affect measurement for performance i
- $\alpha_{j[i]}$ is the random effect of individual j, this is also the mean positive affect in all performaces with an Instructor audience, group performances, and 0 years of study in the instrument for performer j
- β_1 is the average difference in positive affect measurement for Juried Recitals
- β_2 is the average difference in positive affect measurement for Public Performances
- β_3 is the average difference in positive affect measurement for performances with student audience
- $\beta_{4[j]i}$ is another random effect of individual j. This is the average difference in positive affect measurement for solo performances and those that are not solo of individual j.
- β_5 is the average difference in positive affect measurement for each unit increase in years_study
- μ_a is the mean of the positive affect measurement of all performaces across all individuals
- μ_{β_4} is the mean of the difference between the positive affect measurement of solo performances and those that are not across all individuals
- σ_α^2 and $\sigma_{\beta_4}^2$ is the variance between individuals
- σ_y^2 is the variance within performances of an individual

```r
mod3 <- lmer(pa~1 + audience + solo + years_study + (1 + solo|id), data=music)
summary(mod3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: pa ~ 1 + audience + solo + years_study + (1 + solo | id)
##    Data: music
##
## REML criterion at convergence: 3266.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.4430 -0.5258 -0.0065  0.6403  2.6931
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  id       (Intercept) 24.86    4.985
##           solo        22.02    4.692    -0.43
##  Residual             34.09    5.838
## Number of obs: 497, groups:  id, 37
##
## Fixed effects:
##                             Estimate Std. Error t value
## (Intercept)                  32.4502     2.1109  15.372
## audienceJuried Recital        6.6192     1.0844   6.104
## audiencePublic Performance    3.0943     0.9554   3.239
## audienceStudent(s)            0.1112     0.8570   0.130
## solo                         -0.5770     1.1534  -0.500
## years_study                  -0.1843     0.2009  -0.917
##
## Correlation of Fixed Effects:
##             (Intr) adncJR adncPP adnS() solo
## adncJrdRctl -0.163
## adncPblcPrf -0.421  0.301
## adncStdnt() -0.230  0.300  0.507
## solo        -0.450  0.029  0.461  0.182
## years_study -0.804  0.029  0.057 -0.013  0.080
```

Problem 6

Compare the models from the two previous problems using a likelihood ratio test. Which model is better?

```
anova(mod3,mod2)
```

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

```
## Data: music
## Models:
## mod2: pa ~ 1 + audience + solo + years_study + (1 | id)
## mod3: pa ~ 1 + audience + solo + years_study + (1 + solo | id)
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## mod2  8 3310.2 3343.9 -1647.1   3294.2
## mod3 10 3292.7 3334.8 -1636.3   3272.7 21.518      2  2.125e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the likelihood ratio test, the model with the additional varying slope for the “solo” variable is better.

Problem 7

Using the model chosen above, predict the happiness score for the first observation using just the fixed effects by (1) creating the model matrix, (2) obtaining the fixed effect coefficients using `fixedf` (3) multiplying them by the first row of the model matrix you created. Compare this result to the output of the `predict` fuction. Now create a new vector that is the same as the first row of music but with an id = 100. Make a prediction for this observation. Use `predictInterval` in the `merTools` package to get intervals for your two predictions.

```
fixed <- as.matrix(fixef(mod3))

music$Juried <- ifelse(music$audience=="Juried Recital",1,0)
music$Public <- ifelse(music$audience=="Public Performance",1,0)
music$Student <- ifelse(music$audience=="Student(s)",1,0)

model_matrix <- subset(music, select=c(Juried, Public, Student, solo, years_study))
model_matrix$constant <- rep(1,nrow(music))
model_matrix <- as.matrix(model_matrix)
model_matrix <- model_matrix[,c("constant","Juried", "Public", "Student", "solo", "years_study
music$pred_fixedef <- model_matrix%*%fixed
music[,c("fit","upr","lwr")]<-predictInterval(mod3, newdata = music)

head(music)
```

```
##   X id diary previous   perform_type      memory           audience pa na
## 1 1  1     1        0           Solo Unspecified         Instructor 40 11
## 2 2  1     2        1 Large Ensemble      Memory Public Performance 33 19
## 3 3  1     3        2 Large Ensemble      Memory Public Performance 49 14
## 4 4  1     4        3           Solo      Memory Public Performance 41 19
## 5 5  1     5        4           Solo      Memory         Student(s) 31 10
## 6 6  1     6        5           Solo      Memory         Student(s) 33 13
##   age gender instrument years_study mpqab mpqsr mpqpem mpqnem mpqcon solo
## 1  18 Female      voice           3    16     7     52     16     30    1
## 2  18 Female      voice           3    16     7     52     16     30    0
## 3  18 Female      voice           3    16     7     52     16     30    0
## 4  18 Female      voice           3    16     7     52     16     30    1
## 5  18 Female      voice           3    16     7     52     16     30    1
## 6  18 Female      voice           3    16     7     52     16     30    1
##   Juried Public Student pred_fixedef      fit      upr      lwr
## 1      0      0       0     31.32012 35.85677 44.05743 27.71695
## 2      0      1       0     34.99148 39.22960 48.04224 30.41895
## 3      0      1       0     34.99148 39.30678 47.64593 30.59676
## 4      0      1       0     34.41443 38.99217 47.52563 31.02065
## 5      0      0       1     31.43127 35.96234 44.00716 27.70057
## 6      0      0       1     31.43127 36.48120 44.36039 28.79481
```

```
new_data <- head(music,1)
new_data$id <- 100

pred_new <- predictInterval(mod3, newdata = new_data)
```

```
## Warning:      The following levels of id from newdata
##  -- 100 -- are not in the model data.
##      Currently, predictions for these values are based only on the
##  fixed coefficients and the observation-level error.
```

```
tibble(pred_orig_id=music$fit[1],pred_new_id = pred_new$fit)
```

```
## # A tibble: 1 x 2
##   pred_orig_id pred_new_id
##          <dbl>       <dbl>
## 1         35.9        31.5
```

The prediction value for the first row with changed id number is just the same value for predicting using the fixed effects coefficients.