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

```
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.

Answer:

Group-level: age, gender, instrument, years_study, mpqsr, mpqab, mpqpem, mpqnem, mpqcon
Unit-level: diary, previous, perform_type, memory, audience, pa, na

Problem 2

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

```
music %>% count(id) %>% dplyr::select(n) %>% summary()
```

```
##        n
##  Min.   : 2.00
##  1st Qu.:14.00
##  Median :15.00
##  Mean   :13.43
##  3rd Qu.:15.00
##  Max.   :15.00
```

Answer: The minimum number of entries is 2 and maximum and median are both 15.

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.

Answer:

$$y_i = \alpha_{j[i]} + \epsilon_i$$
$$\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$$
$$\epsilon_i \sim N(0, \sigma^2_y)$$

where y_i is the positive affect of diary entry i which was made by musician j. $\alpha_{j[i]}$ represents the random effect of musician j. σ^2_α is the variance between musicians and σ^2_y is the variance within a musician. μ_α is the mean value of positive affect for all musicians.

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

```
## 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
```

```
# Get standard deviations
sigmas <- arm::sigma.hat(model.a)$sigma
# Calculate ICC
icc <- sigmas$id^2/(sigmas$id^2 + sigmas$data^2)
icc
```

```
## (Intercept)
##   0.3626141
```

$\mu_\alpha = 32.56$ is the estimated mean value of positive affect for all musicians and performances. 36.3 % of the total variablility in performance happiness scores is attributable to the difference among subjects.

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.

Answer:

$$y_i = \alpha_{j[i]} + \beta_1 per(juried)_i + \beta_2 per(public)_i + \beta_3 per(students)_i + \beta_4 solo_i + \beta_5 years_{j[i]} + \epsilon_i$$

$$\alpha_j \sim N(\mu_\alpha, \sigma_\alpha^2)$$
$$\epsilon_i \sim N(0, \sigma_y^2)$$

β_1 is the difference between a public performance and Instructor performance, β_2 is the difference between a juried recital and a instructor performance, β_3 is the difference between a performance to students and a instructor performance, β_4 is the effect of performing solo, and β_5 is the change in positive affect with a one year change in years study.

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: pa ~ 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
```

```r
# Get standard deviations
sigmas <- arm::sigma.hat(model.a)$sigma
# Calculate ICC
icc <- sigmas$id^2/(sigmas$id^2 + sigmas$data^2)
icc
```

```
## (Intercept)
##   0.3626141
```

Given the performance type and years of study the within musician variance ($\sigma_y = 6.2$) is greater than the between musician variance ($\sigma_\alpha = 4.5$). Happiness when performing in front of other students is not significantly different than performing for an instructor, (the standard error is greater than the estimated effect). There is evidence that performing in front of a juried audience or public audience increases happiness. Years of study does not seem to have a linear effect on happieness. Performing solo also doesn't seem to have an effect.

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.

$$y_i = \alpha_{j[i]} + \gamma_{j[i]} solo_i + \beta_1 per(juried)_i + \beta_2 per(public)_i + \beta_3 per(students)_i + \beta_5 years_{j[i]} + \epsilon_i$$

$$\alpha_j \sim N(\mu_\alpha, \sigma^2_\alpha)$$
$$\gamma_j \sim N(\mu_\gamma, \sigma^2_\gamma)$$
$$\epsilon_i \sim N(0, \sigma^2_y)$$

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: pa ~ years_study + audience + solo + (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
## years_study                  -0.1843     0.2009  -0.917
## 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
##
## Correlation of Fixed Effects:
##             (Intr) yrs_st adncJR adncPP adnS()
## years_study -0.804
## adncJrdRctl -0.163  0.029
## adncPblcPrf -0.421  0.057  0.301
## adncStdnt() -0.230 -0.013  0.300  0.507
## solo        -0.450  0.080  0.029  0.461  0.182
```

The mean effect of playing solo is $\mu_{gamma} = -0.19$ the variance of the effect of playing solo is $\sigma^2_\gamma = 22.23$.

Problem 6

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

```
anova(model.b, model.c)
```

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

```
## Data: music
## Models:
## model.b: pa ~ audience + solo + years_study + (1 | id)
## model.c: pa ~ years_study + audience + solo + (1 + solo | id)
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## model.b  8 3310.2 3343.9 -1647.1   3294.2
## model.c 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
```

Including solo performance as a random slope significantly improves the model.

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.

```
x <- model.matrix(pa~years_study + audience + solo+id, music)
beta <- fixef(model.c)
x[1,names(beta)]%*%beta
```

```
##          [,1]
## [1,] 31.32012
```

```
predict(model.c, newdata = music[1,], allow.new.levels = TRUE)
```

```
##        1
## 36.26188
```

```
new_musician <- music[1,]
new_musician$id <- 100
predict(model.c, newdata = new_musician, allow.new.levels = TRUE)
```

```
##        1
## 31.32012
```

```
predictInterval(model.c, newdata = new_musician)
```

```
## 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.
```

```
##        fit      upr      lwr
## 1 31.35635 39.19532 23.89344
```

```
predictInterval(model.c, newdata=music[1,])
```

```
##        fit      upr      lwr
## 1 36.87039 44.75776 27.84054
```