1 a 9/9
b 9/9
c 9/9

2 a 9/9
b 9/9
c 9/9

3 7/9

4 a 9/9
b 9/9
c 9/9

5 10/10

Total 98/100

36-763: Homework 05

Vencislav Popov

2015-12-18

Excercise 1

Part (a)

There are some issues with the data that should be cleared before analyzing it further. Classical and popular ratings should vary from 0-10, but there's one response coded as 19, which later on shows strikingly on residual plots. Let's remove it from the start. There are also several responses that are not integers, which is strange given the likert-type scale. Given that there's only a handful of them they are probably coding mistakes so I'll also round all ratings beforehand to get rid of that.

good considerations

```
table(df1$classical)
```

```
##
##   0   1   2   3 3.5   4 4.2 4.6   5   6   7   8   9 9.5  10  19
##   8 111 231 254   1 239   1   1 297 266 334 305 203   1 240   1
```

First, I am fitting a conventional anova with no error structure that ignores that some observations all come from a specific subject.

```
ml1 <- aov(classical ~ instrument + harmony + voice, data=df1)
summary(ml1)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## instrument     2   4092  2045.8 391.691  < 2e-16 ***
## harmony        3    277    92.4  17.689 2.32e-11 ***
## voice          2     79    39.6   7.587 0.000519 ***
## Residuals   2484  12974     5.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 28 observations deleted due to missingness
```

We can see that all three factors are significantly related with the classical ratings. As evidenced by the η^2 change for each factor (3rd column bellow), we can see that the classical ratings are mostly influenced by the instrument, and very weakly by the harmony or the voice:

```
library(lmSupport)
modelEffectSizes(ml1)
```

```
## aov(formula = classical ~ instrument + harmony + voice, data = df1)
##
## Coefficients
##                    SSR df pEta-sqr dR-sqr
## (Intercept) 5876.9533  1   0.3118     NA
## instrument  4091.4847  2   0.2398 0.2348
## harmony      277.1294  3   0.0209 0.0159
```

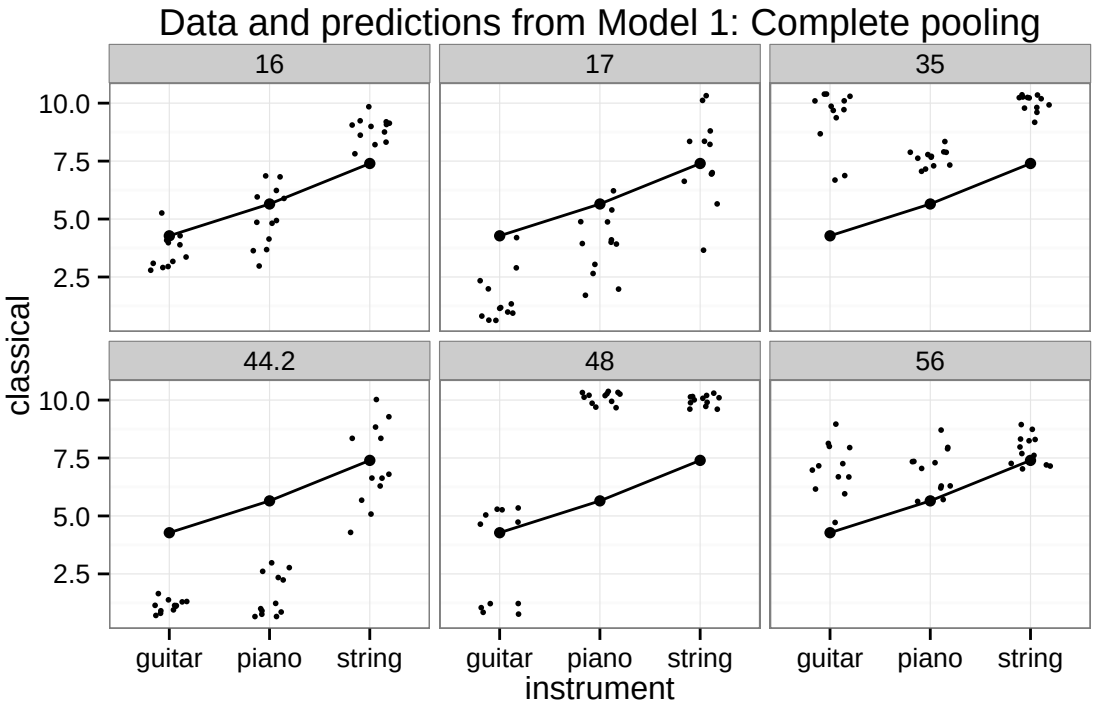
```
## voice         79.2538  2   0.0061 0.0045
##
## Sum of squared errors (SSE): 12973.8
## Sum of squared total  (SST): 17421.7
```

To see the specific effects of each condition we can perform pairwise comparisons with Tukey's HSD test. We see that classical ratings increase from guitar through piano to string performances of the pieces. We can also see that parallel 3rds and 5ths are rated as less classical than contrary voice leadings, but they do not differ from one another. Finally, for the harmonies, I-V-VI are rated as more classical than each of the other 3 sequences, while they do not differ from one another:

```
options(digits=3)
TukeyHSD(ml1)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = classical ~ instrument + harmony + voice, data = df1)
##
## $instrument
##               diff  lwr  upr p adj
## piano-guitar  1.37 1.11 1.64     0
## string-guitar 3.12 2.86 3.38     0
## string-piano  1.75 1.48 2.01     0
##
## $harmony
##                   diff    lwr    upr p adj
## I-V-IV-I-IV-V -0.0322 -0.365  0.301 0.995
## I-V-VI-I-IV-V  0.7685  0.436  1.101 0.000
## IV-I-V-I-IV-V  0.0310 -0.302  0.364 0.995
## I-V-VI-I-V-IV  0.8006  0.467  1.134 0.000
## IV-I-V-I-V-IV  0.0631 -0.270  0.396 0.962
## IV-I-V-I-V-VI -0.7375 -1.071 -0.405 0.000
##
## $voice
##                     diff    lwr     upr p adj
## par3rd-contrary -0.3978 -0.661 -0.1347 0.001
## par5th-contrary -0.3554 -0.618 -0.0925 0.004
## par5th-par3rd    0.0424 -0.221  0.3053 0.924
```

One major drawback of the preceding analysis is that measurements for the different conditions come from the same subjects, and this violates the assumption of independence for the analysis of variance and the Tukey test. To alleviate this, we will fit a hierarchical model that accounts for the data structure in the next part. The issue is evident when we plot the model's predictions for a few subjects (plotting only instrument for simplicity; data points are jittered for readability):



9

Part (b)

The full model we are going to fit is written below. Since our categorical independent variables have more than two levels, the formula contains dummy variables for all factors, and the intercept reflects the value for the combination of non-mentioned levels (guitar + I-IV-V + contrary).

$$y_i = \alpha_{0j[i]} + \alpha_1 * \text{piano} + \alpha_2 * \text{string} + \alpha_3 * \text{I-V-IV} + \alpha_4 * \text{I-V-VI} + \alpha_5 * \text{IV-I-V} + \alpha_6 * \text{3rd} + \alpha_7 * \text{5th} + \varepsilon_i, \varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$$

$$\alpha_{0j} = \beta_0 + \eta_j, \eta_j \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$$
$$j = 1, 2, ..., 70$$

We can fit this model with `lmer`:

```
ml2 <-  lmer(classical ~ instrument + harmony + voice + (1|subject), data=df1, REML=F)
```

To test whether the random intercept is needed in the model we compare the AIC estimates of the model without and with the random intercept. There's a huge improvement in the model (δ AIC = -761):

```
round(c(AIC(ml1),AIC(ml2)))
```

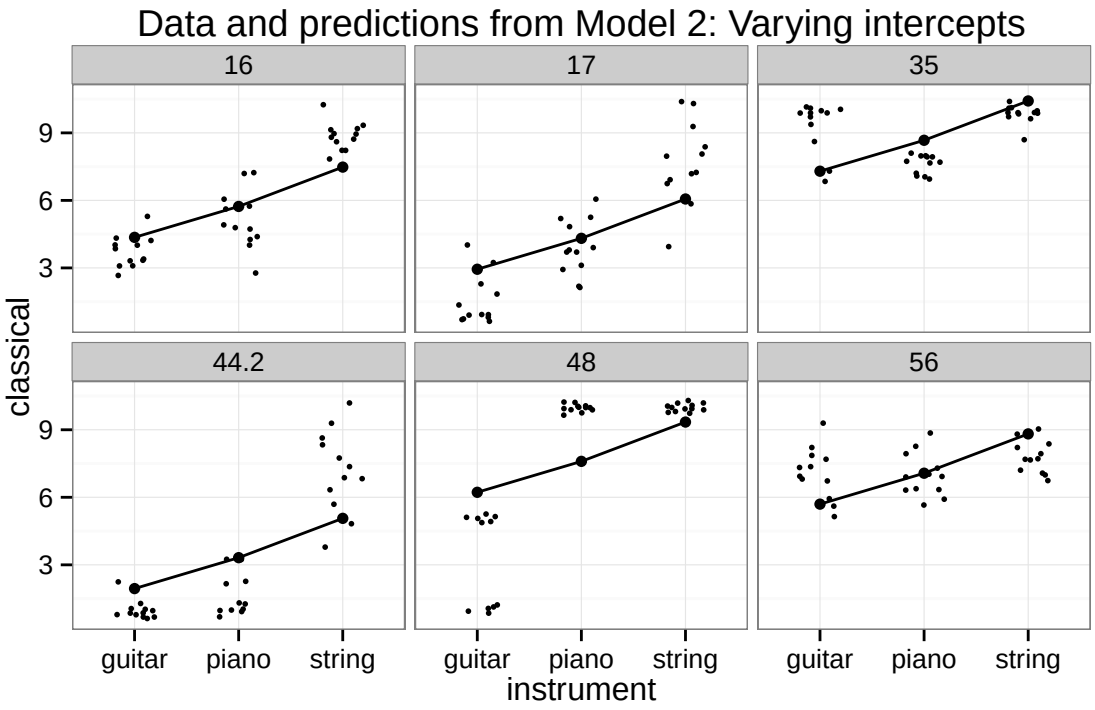
```
## [1] 11201 10434
```

We can also perform a restricted likelihood ratio test that tells us that the variance of the random effect is reliably different from 0:

```
library(RLRsim)
exactRLRT(ml2)
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 800, p-value <2e-16
```

And from graphing the model prediction we can see that the fit is much better at the individual level:



Let's compare the fixed effects. All model fit parameters and the logLikelihood ratio tests suggest that all three factors are still significant even after accounting for the clustering structure within subjects. Again, the greatest improvement of the model is due to instrument, folowed by harmony, and least for voice, where the BIC decrease is only 1 (all other fit indices still suggest a significan influence of voice):

```
do.call(anova,list(update(ml2, .~.-instrument-voice-harmony),
                   update(ml2, .~.-instrument-voice), update(ml2, .~.-instrument),ml2))
```

```
## Data: df1
## Models:
## MODEL1: classical ~ (1 | subject)
## MODEL2: classical ~ harmony + (1 | subject)
## MODEL3: classical ~ harmony + voice + (1 | subject)
## MODEL4: classical ~ instrument + harmony + voice + (1 | subject)
##        Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
```

```
## MODEL1  3 11435 11453  -5715    11429
## MODEL2  6 11388 11423  -5688    11376  52.8      3      2e-11 ***
## MODEL3  8 11377 11424  -5681    11361  15.2      2    0.00051 ***
## MODEL4 10 10434 10492  -5207    10414 946.9      2    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

When we look at the pair-wise comparisons from this model we get the same results as the model without the random intercept.

```
library(multcomp)
summary(glht(ml2, linfct=mcp(instrument="Tukey", voice="Tukey", harmony="Tukey")))
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lmer(formula = classical ~ instrument + harmony + voice + (1 |
##     subject), data = df1, REML = F)
##
## Linear Hypotheses:
##                                  Estimate Std. Error z value Pr(>|z|)
## instrument: piano - guitar == 0    1.3737     0.0925   14.85  < 1e-04 ***
## instrument: string - guitar == 0   3.1202     0.0919   33.96  < 1e-04 ***
## instrument: string - piano == 0    1.7465     0.0923   18.92  < 1e-04 ***
## voice: par3rd - contrary == 0     -0.4013     0.0922   -4.35  0.00017 ***
## voice: par5th - contrary == 0     -0.3597     0.0921   -3.90  0.00111 **
## voice: par5th - par3rd == 0        0.0416     0.0921    0.45  0.99965
## harmony: I-V-IV - I-IV-V == 0     -0.0324     0.1064   -0.30  0.99998
## harmony: I-V-VI - I-IV-V == 0      0.7708     0.1064    7.25  < 1e-04 ***
## harmony: IV-I-V - I-IV-V == 0      0.0320     0.1063    0.30  0.99998
## harmony: I-V-VI - I-V-IV == 0      0.8032     0.1065    7.54  < 1e-04 ***
## harmony: IV-I-V - I-V-IV == 0      0.0644     0.1064    0.60  0.99769
## harmony: IV-I-V - I-V-VI == 0     -0.7388     0.1064   -6.94  < 1e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Part (c)

The model with separete random intercepts for each factor within subjects is:

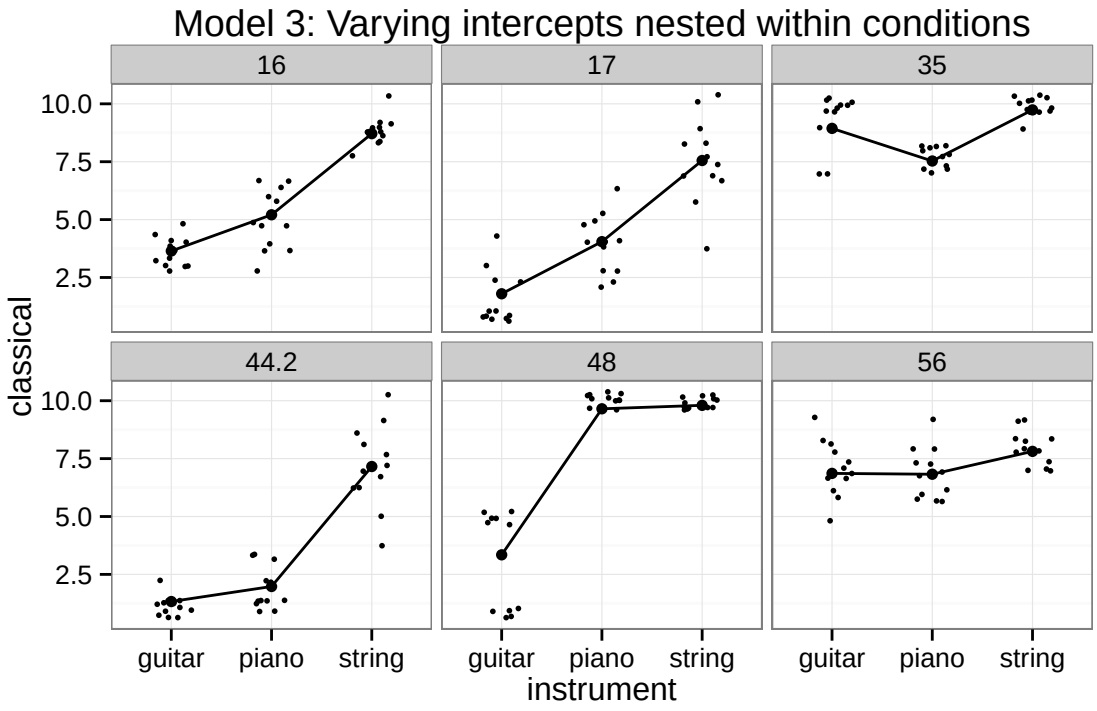
```
ml3 <-  lmer(classical ~ instrument + harmony + voice + (1|subject:instrument) +
              (1|subject:harmony)+ (1|subject:voice), data=df1, REML=F)
```

We can Use AIC to compare this model to the previous two models, and it looks like there is a large improvement over the overall random intercept for each subject model:

```
round(c(AIC(ml1),AIC(ml2),AIC(ml3)))
```

```
## [1] 11201 10434 10015
```

Again, we can see that in the improved model fit from graphing predictions for each subject:



When we look at the fixed effects, we see that they still significantly predict the classical ratings. The difference with the previous model is that their relative influence has changed as evidenced by the $\Delta AICs, BICs$ and $Chisq's$. While the effects of instrument and harmony seem to be smaller after accounting for the nested random effects, the influence of voice has increased.

```
do.call(anova,list(update(ml3, .~.-instrument-voice-harmony),
                   update(ml3, .~.-instrument-voice), update(ml3, .~.-instrument),ml3))
```

```
## Data: df1
## Models:
## MODEL1: classical ~ (1 | subject:instrument) + (1 | subject:harmony) +
## MODEL1:     (1 | subject:voice)
## MODEL2: classical ~ harmony + (1 | subject:instrument) + (1 | subject:harmony) +
## MODEL2:     (1 | subject:voice)
## MODEL3: classical ~ harmony + voice + (1 | subject:instrument) + (1 |
## MODEL3:     subject:harmony) + (1 | subject:voice)
## MODEL4: classical ~ instrument + harmony + voice + (1 | subject:instrument) +
## MODEL4:     (1 | subject:harmony) + (1 | subject:voice)
##        Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## MODEL1  5 10177 10207  -5084    10167
## MODEL2  8 10142 10188  -5063    10126  41.8      3    4.5e-09 ***
## MODEL3 10 10118 10176  -5049    10098  27.8      2    9.4e-07 ***
```

```
## MODEL4 12 10015 10085  -4995     9991 107.0      2    < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As for the variance of the random effects, we see that variance for instrument within subjects is greater that the variance of the other two factors, suggesting that while instrument is the most influential variable on the subjects' ratings, there's also much more variation in the way subjects weight that variable, compared to the others. It is comparable to the residuals variance, suggesting that there's as much variation within subject:instrument combinations as there is between them. The variation in subject:voice intercepts is negligible compared to the residual variance.

The model's mathematical form is:

$$y_i = \alpha_{0j[i]p[i]q[i]m[i]} + \alpha_1 * \text{piano} + \alpha_2 * \text{string} + \alpha_3 * \text{I-V-IV} + \alpha_4 * \text{I-V-VI} + \alpha_5 * \text{IV-I-V} + \alpha_6 * \text{3rd} + \alpha_7 * \text{5th} + \varepsilon_i$$

$$\alpha_{jpqm} = \beta_0 + \eta_{jp} + \eta_{jq} + \eta_{jm}$$

$$\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \eta_{jp} \overset{iid}{\sim} \mathcal{N}(0, \tau_1^2), \eta_{jq} \overset{iid}{\sim} \mathcal{N}(0, \tau_2^2), \eta_{jm} \overset{iid}{\sim} \mathcal{N}(0, \tau_3^2)$$

$$j = 1, ..., 70; p = 1, 2, 3, 4; q = 1, 2, 3; m = 1, 2, 3;$$

This nested random intercepts model models the variance to be the same for each condition within a factor - people vary as much in their evaluation of pieces as classical when they are played by a guitar, piano or a string quartet (as evidenced by the single variance parameters for the random intercept for subject:instrument). However, we can go even further in positing that not only the type of instrument would affect how classical a piece sounds, but that for some instruments people's estimates would vary more than others (the same with the other factors). To do this we can try to fit the following models, by changing one random effect at a time from a nested to a slope random effect:

```
ml4 <- lmer(classical ~ instrument + harmony + voice + (instrument-1|subject) +
              (1|subject:voice) + (1|subject:harmony),
              data=df1, REML=F, control=lmerControl(optCtrl=list(maxfun=100000)))
ml5 <- lmer(classical ~ instrument + harmony + voice + (instrument-1|subject) +
              (1|subject:voice) + (harmony-1|subject),
              data=df1, REML=F, control=lmerControl(optCtrl=list(maxfun=100000)))
ml6 <- lmer(classical ~ instrument + harmony + voice + (instrument-1|subject) +
              (voice-1|subject) + (harmony-1|subject), data=df1, REML=F,
              control=lmerControl(optCtrl=list(maxfun=100000)))
```

I don't think random effects of this type make sense for this problem.

And then we compare the models with the original model and with one another. Fitting a separate random effect for each instrument and for each harmony improved the model significantly, but the random effects for voice complicate the model unnecessarily, while decreasing the fit (There's no straight forward way to do this with `exactRLRT`, so I'm using AIC to compare them):

```
anova(ml3,ml4,ml5,ml6)
```

```
## Data: df1
## Models:
## ml3: classical ~ instrument + harmony + voice + (1 | subject:instrument) +
## ml3:     (1 | subject:harmony) + (1 | subject:voice)
## ml4: classical ~ instrument + harmony + voice + (instrument - 1 |
## ml4:     subject) + (1 | subject:voice) + (1 | subject:harmony)
## ml5: classical ~ instrument + harmony + voice + (instrument - 1 |
```

```
## ml5:     subject) + (1 | subject:voice) + (harmony - 1 | subject)
## ml6: classical ~ instrument + harmony + voice + (instrument - 1 |
## ml6:     subject) + (voice - 1 | subject) + (harmony - 1 | subject)
##     Df   AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## ml3 12 10015 10085  -4995     9991
## ml4 17  9972 10071  -4969     9938 53.42      5    2.8e-10 ***
## ml5 26  9918 10070  -4933     9866 71.27      9    8.6e-12 ***
## ml6 31  9927 10107  -4932     9865  1.59      5        0.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally, given how low the variance is for the random effect of voice nested within subjects, we can test whether we need it at all with the exact RLRT test. This random effect is unnecessary:

```
(ran.test1 <- exactRLRT(update(ml5, .~.-(instrument-1|subject)-(harmony-1|subject)),
                        ml5,update(ml5, .~.-(1|subject:voice))))
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 0.3, p-value = 0.3
```

Thus, our final model is:

```
ml7 <- lmer(classical ~ instrument + harmony + voice + (instrument-1|subject) +
  (harmony-1|subject), data=df1, REML=F, control=lmerControl(optCtrl=list(maxfun=100000)))
```

Below, we can see that people's ratings are more variable for pieces played with a guitar, than with a piano and a string quartet, and the ratings for those played with piano are also more variable than with a string quartet. That is, pieces played with a string quartet are uniformly rated as highly classical, while those played with a guitar are on average rated as not very classical (~4.34/10), but people are also much more variable in how classical they perceive guitar pieces to be. We also see a strong negative correlation between the random effects for string and guitar (r = -1), suggesting that people who tend to rate guitar pieces as less classical that other consistently tend to rate string pieces as more classical than others. The correlations with the random effect of piano are much lower, possibly suggesting that people evaluate piano pieces by different/additional criteria, or that piano pieces are more ambiguous as to how classical they sound. The variance estimates for the random effect of voice are very similar to one another, as are those for the harmony. Judging by the correlation estimates for the random effects of harmony, the I-V-VI seems to be judged rather differently than the other three, which correlated strongly with one another, while the I-V-VI harmony correlations are in the range [0.51,0.57].

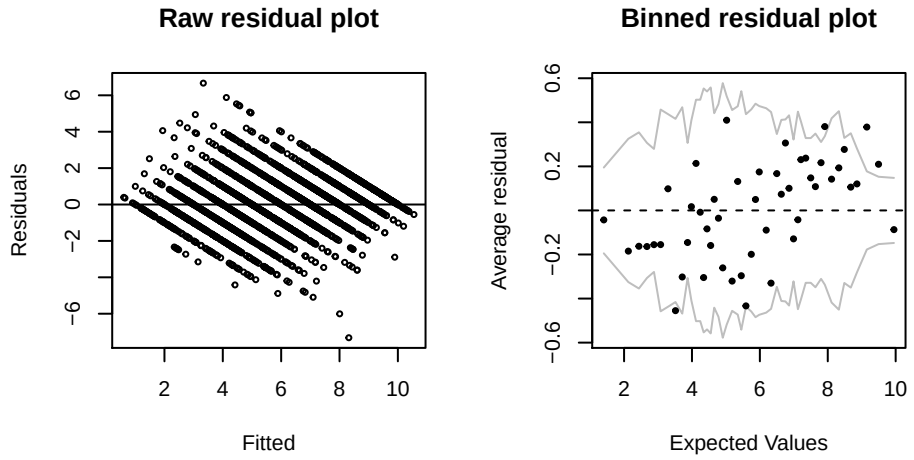
```
summary(ml7)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: classical ~ instrument + harmony + voice + (instrument - 1 |
##     subject) + (harmony - 1 | subject)
##    Data: df1
## Control: lmerControl(optCtrl = list(maxfun = 1e+05))
##
```

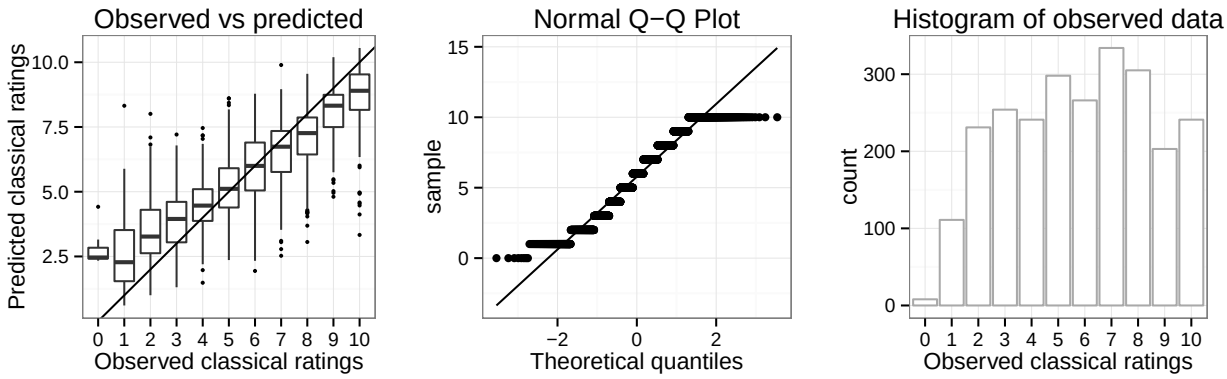
```
##      AIC      BIC   logLik deviance df.resid
##     9916    10062    -4933     9866     2467
##
## Scaled residuals:
##    Min     1Q Median     3Q    Max
## -4.714 -0.579  0.029  0.552  4.294
##
## Random effects:
##  Groups    Name             Variance Std.Dev. Corr
##  subject   instrumentguitar 1.283    1.133
##            instrumentpiano  0.983    0.992     0.28
##            instrumentstring 0.534    0.731    -1.00 -0.29
##  subject.1 harmonyI-IV-V    1.748    1.322
##            harmonyI-V-IV    2.023    1.422    0.99
##            harmonyI-V-VI    1.970    1.404    0.57  0.62
##            harmonyIV-I-V    1.802    1.342    1.00  0.98  0.58
##  Residual                   2.411    1.553
## Number of obs: 2492, groups:  subject, 70
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)         4.3381     0.2259   19.21
## instrumentpiano     1.3673     0.1715    7.97
## instrumentstring    3.1168     0.2354   13.24
## harmonyI-V-IV      -0.0307     0.0913   -0.34
## harmonyI-V-VI       0.7711     0.1746    4.42
## harmonyIV-I-V       0.0348     0.0885    0.39
## voicepar3rd        -0.3952     0.0762   -5.18
## voicepar5th        -0.3554     0.0762   -4.66
##
## Correlation of Fixed Effects:
##             (Intr) instrmntp instrmnts hI-V-I hI-V-V hIV-I- vcpr3r
## instrumntpn -0.434
## instrmntstr -0.622  0.634
## hrmnyI-V-IV -0.106  0.000     0.000
## hrmnyI-V-VI -0.346  0.000     0.000     0.291
## hrmnyIV-I-V -0.176 -0.001     0.000     0.459  0.253
## voicepar3rd -0.168 -0.001    -0.001    -0.002  0.000  0.001
## voicepar5th -0.168 -0.001     0.000    -0.002 -0.002 -0.002  0.500
```

Examining the fixed effects tells the same story as before, so I will not display them for the sake of brevity.

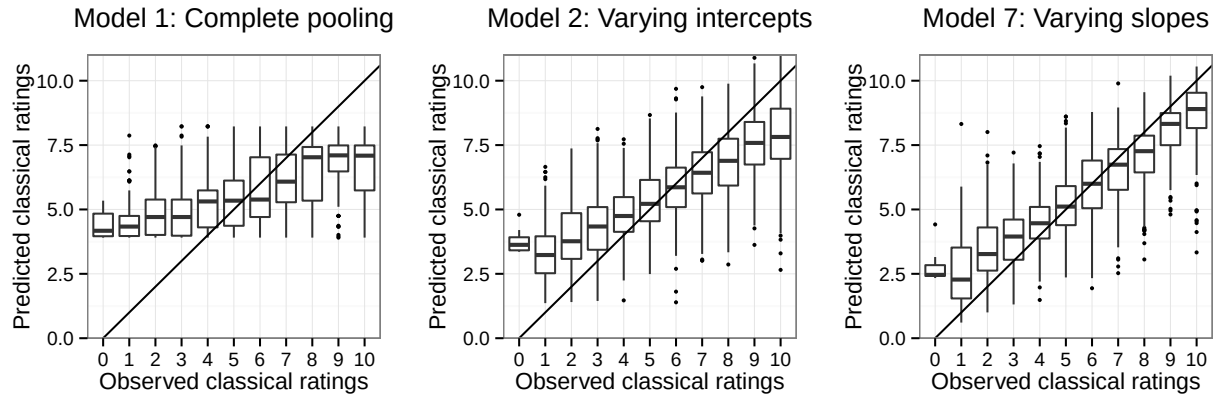
Before we continue we should check the model fit by looking at the residuals (raw on the right, binned on the left):



Since the response is multinomial and bounded between 0 and 10 wee see clustering in the residuals, which makes it difficult to interpret. Judging by the binned residuals on the right, the model is overpredicting low ratings and underpredicting high ratings. This can also be more clearly seen on the graph below on the left which plots the observed vs predicted classical ratings. This might be occuring because the model is assuming a normal ditribution of the errors, while the observed responses actually deviate from normality by having many more responses at the extreme values han predicted by a normal distribution (the middle and the right graph below demonstrate this). By using a normal distribution for the errors the model is predicted there to be less extreme values than there are in the observed data.



In fact, we can compare three of the key models tested so far using the same observed vs predicted plot - the completely pooled model (model 1, no random effects), the model with varying intercepts overall per subject (model 2, varying overall intercepts), and the last fitted model with varying slopes for instrument and harmony. We can see that including the random intercept and then the random slopes drastically reduces the overprediction of low ratings and the underprediction of high ratings. This is probably due to the fact that in the completely pooled model predicted values are the same for all subjects, without regard for their bias



After some additional exploration I found that there's a significant interaction between voice and harmony.

```
ml8 <- lmer(classical ~ instrument+voice*harmony + (instrument-1|subject) +
  (harmony-1|subject), data=df1, REML=F, control=lmerControl(optCtrl=list(maxfun=100000)))
anova(ml7, ml8)
```

```
## Data: df1
## Models:
## ml7: classical ~ instrument + harmony + voice + (instrument - 1 |
## ml7:     subject) + (harmony - 1 | subject)
## ml8: classical ~ instrument + voice * harmony + (instrument - 1 |
## ml8:     subject) + (harmony - 1 | subject)
##     Df  AIC   BIC logLik deviance Chisq Chi Df Pr(>Chisq)
## ml7 25 9916 10062  -4933     9866
## ml8 31 9892 10073  -4915     9830  36.1      6    2.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

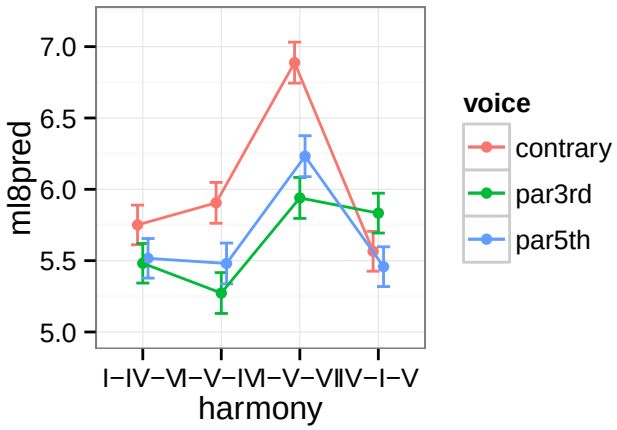
The main effects for voice are no longer significant, though they have rather similar estimates as before. From the interaction we can see that the contrary voice (the implicit basline) is higher than the 3rd and 5th voices only for the I-V-VI harmony condition, while we have a reversal for the IV-I-V condition, where the 3rd voice condition is actually rated as more classical than the contrary condition and the voice 5th condition (the latter not shown below).

```
options(digits=2)
summary(ml8)$coefficients
```

```
##                             Estimate Std. Error t value
## (Intercept)                     4.26       0.24   17.89
## instrumentpiano                 1.37       0.17    7.99
## instrumentstring                3.12       0.24   13.24
## voicepar3rd                    -0.27       0.15   -1.78
## voicepar5th                    -0.23       0.15   -1.55
## harmonyI-V-IV                   0.15       0.15    1.01
## harmonyI-V-VI                   1.14       0.21    5.32
## harmonyIV-I-V                  -0.18       0.15   -1.22
## voicepar3rd:harmonyI-V-IV      -0.36       0.21   -1.70
## voicepar5th:harmonyI-V-IV      -0.19       0.21   -0.89
## voicepar3rd:harmonyI-V-VI      -0.68       0.21   -3.18
```

```
## voicepar5th:harmonyI-V-VI    -0.42       0.21   -1.97
## voicepar3rd:harmonyIV-I-V     0.54       0.21    2.51
## voicepar5th:harmonyIV-I-V     0.13       0.21    0.59
```

For more clarity you can see the behavioral data below (Error bars represent 1 between-subject SE; I've chosen line graphs instead of barplots because the pattern is easier to see. It is not meant to imply that there are harmony levels in-between the 4 stated on the x axis):



9

Excercise 2

Part (a)

The general approach I took was to use the `bfFixefLMER_F.fnc` function from the `LMERConvenienceFunctions` package to back-fit the fixed effects of a maximal model based on log-likelihood ratio tests. I started with the best fitting model from excercise 1, and I added almost all covariates from the data, and their interactions with the three within-subject factors (instrument, voice and harmony).

One problem with this approach was that a lot of the covariates had missing values for a few of the subjects and those subjects differed between the covariates. Including all covariates in the model (with the exception of `X1stInstr` and `X2ndInstr` because both had missing values for more than half of the subjects), resulted in a model that used only 60% of the data (42 out of 70 subjects). Fitting the model on this subset might be biased, and I would be very cautious in interpreting any results, since they might depend on the pattern of missing data.

```
# number of NAs per predictor:
```

```
##                                [,1]
## consnotes                       360
## pachlisten                       72
## clslisten                        36
## knowrob                         180
## knowaxis                        288
## x1990s2000s                     144
## x1990s2000s.minus.1960s1970s    180
## collegemusic                    108
## noclass                         288
## aptheory                        216
## composing                        72
## x1stinstr                      1512
## x2ndinstr                      2196
## classical                        28
## popular                          28
## clslisten.c                      36
```

To circumvent this problem, I did the following. I fit the maximal model described above on the complete cases (42/70 subjects) and I used the back-fitting procedure to get a simpler model. Since some of the parameters that were removed had missing data themselves, I was able to refit the resulting model on a slightly bigger dataset (50/70 subjects). I then repeated the back-fit procedure on this larger dataset. This procedure continued itteratively for a total of 4 times until there was no change in the resulting model. This results in a more conservative model incroporating less parameters than those fit on the complete subset of the data, or on the data with imputed missing values, but also incorporates the generalization of predictions to values on which the initial model was fit, and avoids biased results that might be dependent on the pattern of missing values.

The final model was fit on 66/70 subjects and was of the following form (all the covariates were mean centered; omsi and x16.minus.17 were also divided by their SDs):

```
classical ~ instrument + voice + harmony + voice:harmony +          # within-subject
  selfdeclare + omsi + x16.minus.17 + clslisten + collegemusic +    # covariates
  voice:x16.minus.17 + voice:collegemusic + harmony:selfdeclare +   # interactions
  (instrument-1|subject) + (harmony-1|subject)                      # random-effects
```

9

Part (b)

Since all of the covariates in the model are at the subject level, there are no random slopes for subject that can be added for them, and it makes little sense to add random intercepts for them since they are included as fixed effects. I can check whether the random effect for the voice, which was droped in the initial model specification, is now needed, and it is not:

```
(rand.test2 <- exactRLRT(update(ml17, .~. +(1|subject:voice)-(instrument-1|subject)-
                        (harmony-1|subject)), update(ml17, .~. +(1|subject:voice)), ml17))
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 0.0059559, p-value = 0.4523
```

Let's also see if the two random slopes for instrument and harmony are also needed, or the nested random intercepts would suffice now that we've accounted for much of the variation by the covariates:

```
ml19 <- update(ml18, .~. -(instrument-1|subject)+(1|subject:instrument))
anova(ml17,ml19)
```

```
## Data: ml17@frame
## Models:
## ml19: classical ~ instrument + voice + harmony + selfdeclare + x16.minus.17 +
## ml19:     clslisten + collegemusic + (harmony - 1 | subject) + (1 |
## ml19:     subject:instrument) + voice:harmony + voice:x16.minus.17 +
## ml19:     voice:collegemusic + harmony:selfdeclare
## ml17: classical ~ instrument + voice + harmony + selfdeclare + omsi +
## ml17:     x16.minus.17 + clslisten + collegemusic + (instrument - 1 |
## ml17:     subject) + (harmony - 1 | subject) + voice:harmony + voice:x16.minus.17 +
## ml17:     voice:collegemusic + harmony:selfdeclare
##      Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
## ml19 37 9493.9 9707.8  -4710   9419.9
## ml17 43 9486.1 9734.7  -4700   9400.1 19.854      6    0.00294 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Judging by the AIC, the random slopes argument for instrument is still necessary. And so is the random slopes argument for harmony as well:

```
ml20 <- update(ml18, .~. -(harmony-1|subject)+(1|subject:harmony))
anova(ml17,ml20)
```

```
## Data: ml17@frame
## Models:
## ml20: classical ~ instrument + voice + harmony + selfdeclare + x16.minus.17 +
## ml20:     clslisten + collegemusic + (instrument - 1 | subject) + (1 |
## ml20:     subject:harmony) + voice:harmony + voice:x16.minus.17 + voice:collegemusic +
## ml20:     harmony:selfdeclare
```

```
## ml17: classical ~ instrument + voice + harmony + selfdeclare + omsi +
## ml17:     x16.minus.17 + clslisten + collegemusic + (instrument - 1 |
## ml17:     subject) + (harmony - 1 | subject) + voice:harmony + voice:x16.minus.17 +
## ml17:     voice:collegemusic + harmony:selfdeclare
##      Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## ml20 33 9522.1 9712.9 -4728.1   9456.1
## ml17 43 9486.1 9734.7 -4700.0   9400.1 56.068     10  1.994e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore the final model remains as above.

Finally, let's compare our final model with the covariates we've chosen, with the best model without the covariates (fitted on the same subset of the data):

```
ml7.new <- update(ml7, data=ml17@frame)
anova(ml7.new, ml17)
```

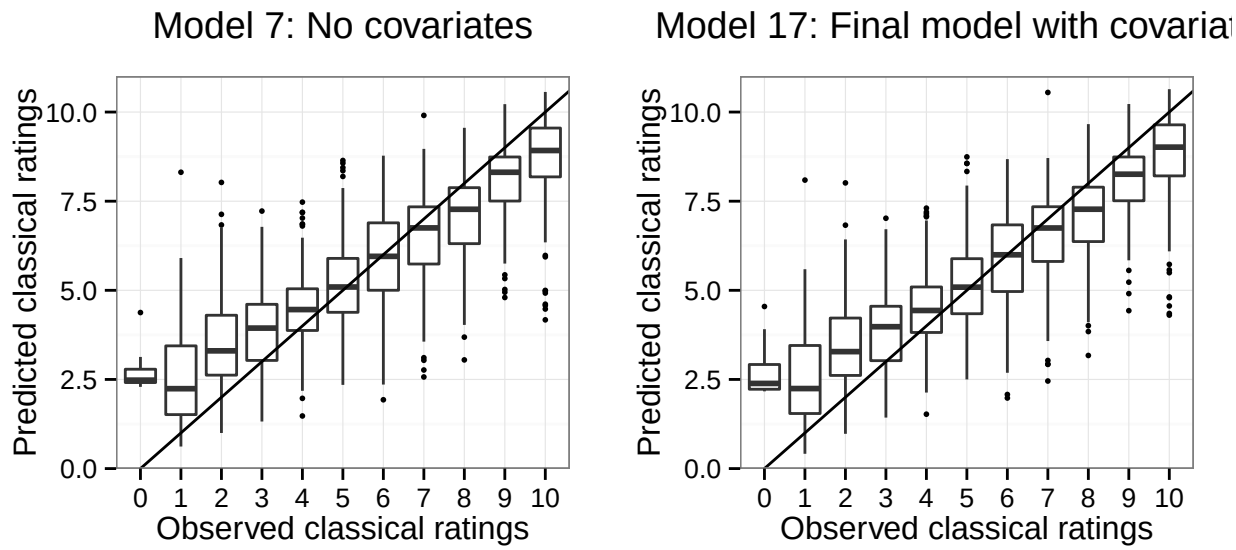
```
## Data: ml17@frame
## Models:
## ml7.new: classical ~ instrument + harmony + voice + (instrument - 1 |
## ml7.new:     subject) + (harmony - 1 | subject)
## ml17: classical ~ instrument + voice + harmony + selfdeclare + omsi +
## ml17:     x16.minus.17 + clslisten + collegemusic + (instrument - 1 |
## ml17:     subject) + (harmony - 1 | subject) + voice:harmony + voice:x16.minus.17 +
## ml17:     voice:collegemusic + harmony:selfdeclare
##         Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## ml7.new 25 9531.5 9676.1 -4740.8   9481.5
## ml17    43 9486.1 9734.7 -4700.0   9400.1 81.454     18  4.763e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There's overall improvement in the model by adding the covariates chose by the automated method, though BIC penalizes the model quite a bit for the increased number of parameters. Also, even though the covariates are significant they account only for a small proportion of the remaining residuals of the model:

```
## Residual variance of full model with covariates: 2.354
```

```
## Residual variance of model without covariates:   2.398
```

This can also be seen from the observed vs predicted plot which we had for the previous models as well, and there's no striking change:



9

Part (c)

Here's a summary of the model parameters:

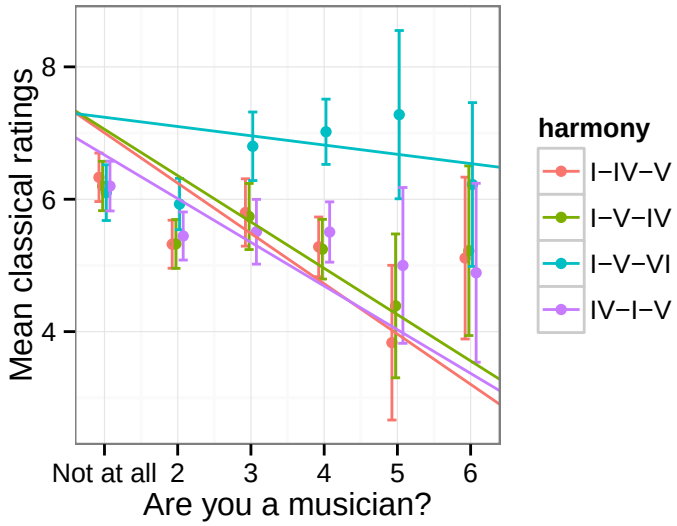
```
round(summary(ml17)$coefficients,3)
```

```
##                                Estimate Std. Error t value
## (Intercept)                       4.394      0.369  11.900
## instrumentpiano                   1.393      0.178   7.837
## instrumentstring                  3.164      0.243  13.009
## voicepar3rd                      -0.516      0.213  -2.417
## voicepar5th                      -0.575      0.214  -2.694
## harmonyI-V-IV                     0.149      0.156   0.957
## harmonyI-V-VI                     1.138      0.201   5.653
## harmonyIV-I-V                    -0.191      0.154  -1.242
## selfdeclare.c                    -0.764      0.225  -3.388
## omsi.c                            0.473      0.222   2.124
## x16.minus.17.c                   -0.059      0.159  -0.374
## clslisten.c                       0.335      0.103   3.260
## collegemusic1                    -0.198      0.363  -0.546
## voicepar3rd:harmonyI-V-IV        -0.345      0.216  -1.594
## voicepar5th:harmonyI-V-IV        -0.206      0.217  -0.951
## voicepar3rd:harmonyI-V-VI        -0.658      0.217  -3.040
## voicepar5th:harmonyI-V-VI        -0.434      0.217  -2.003
## voicepar3rd:harmonyIV-I-V         0.548      0.216   2.534
## voicepar5th:harmonyIV-I-V         0.107      0.216   0.496
## voicepar3rd:x16.minus.17.c       -0.202      0.080  -2.526
## voicepar5th:x16.minus.17.c       -0.246      0.080  -3.083
## voicepar3rd:collegemusic1         0.307      0.188   1.628
## voicepar5th:collegemusic1         0.452      0.188   2.400
## harmonyI-V-IV:selfdeclare.c       0.056      0.079   0.706
## harmonyI-V-VI:selfdeclare.c       0.616      0.135   4.555
## harmonyIV-I-V:selfdeclare.c       0.102      0.077   1.333
```

The interpretations of all covariates given below are with respect to controling for the effects of the other covariates. Thus graphs presented from here on are more difficult to interpret, since the raw data doesn't account for this.

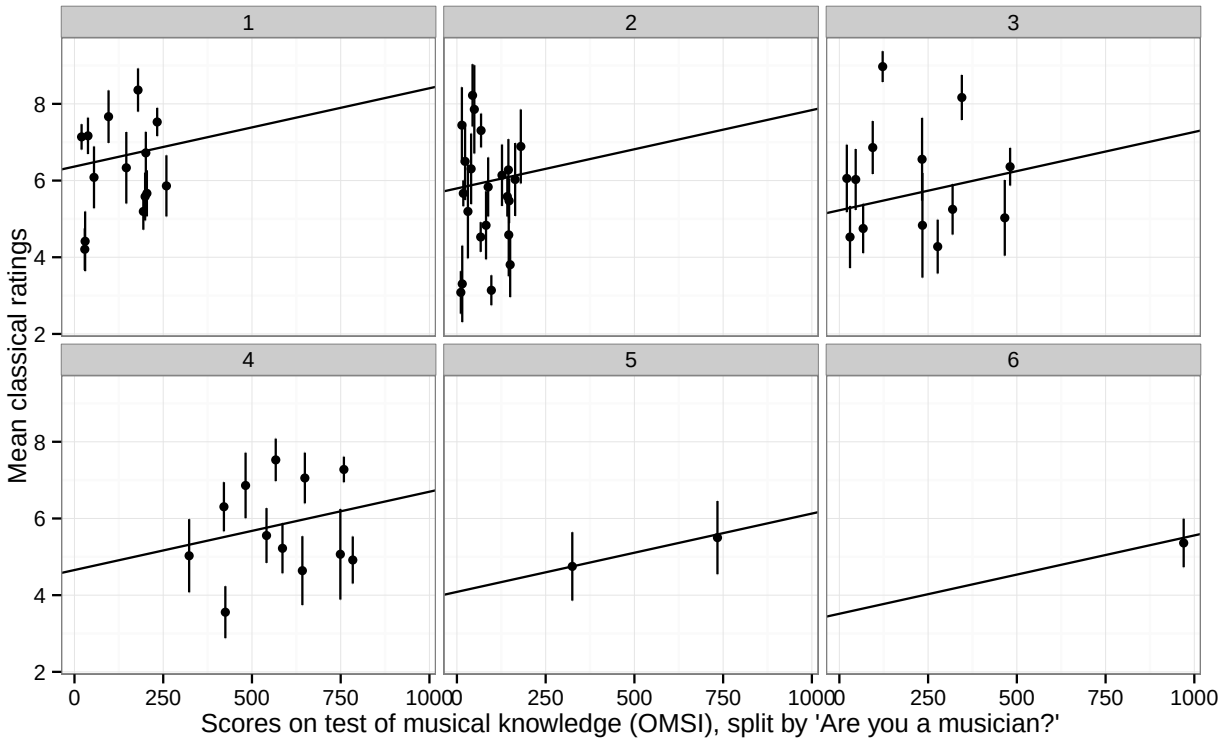
There's no change in the significance of main within-subject effects. The interpretation of their estimates however is at the mean values of all covariates. As for the covariates, there was a general negative effect of `selfdeclare` - the more people cosidered themselves to be musicians, the less classical they rated the pieces to be. However, there was a significant interaction between harmony and self-ratings of being a musician (`selfdeclare`). The decrease in classical ratings with higher self-ratings of being a musician wasn't true for the I-V-VI harmonies, where the slopes was notably closer to 0 compared to the other three harmonies. That is, people's estimation of them being a musician affected their ratings in general, but not for I-V-VI which was similar for all people

The main effect of `selfdeclare` and it's interaction with harmony can be seen on the figure below (I'm plotting the mean classical ratings at each `selfdeclare` point along with errorbars instead of each data point; keep in mind this is the observed data, and the regression lines are conditional on the other covariates in the model).



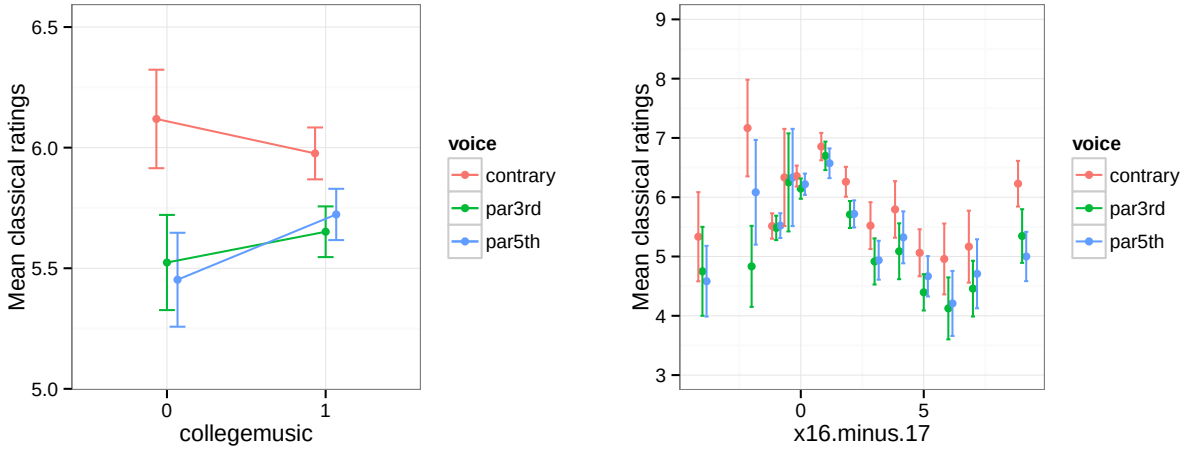
Interestingly, while self-ratings of being a music are negatively related with ratings of classicicity, the opposite is true for scores on an actual test of musical knowledge, which are positively related with ratings of classicicity (variable `omsi`) after controlling for the self ratings. 1 standard deviation increase in musical knowledge scores leads to rating pieces as 0.5 points more classical. This relationship is not obvious in the raw data, and the coefficient for `omsi` is near 0 and nonsignificant if `selfdeclare` isn't included in the model. That is, while people who consider themselves to be musicians rate the pieces as less classical in general, among people with the same self ratings musical knowledge actually leads to higher ratings of classicicity. This relationship is show on the figure below.

9



Next, we can see that people who tend to listen to classical music a lot are likely to rate the pieces as more classical than those who do not (1 point change on the scale 0-5 of How much do you listen to classical music leads to 0.33 increase in classical ratings). The next two factors don't have significant main effects, but they interact with the within-subject factors.

There's a significant interaction between whether people have taken music classes in college and how classical they rate the each of the different voice leadings. While overall contrary voice leadings are rated the highest (~0.54 higher than 3rd and 5th voice), this difference is reduced for people who have taken music classes in college, where contrary voice leadings are rated only 0.41 and 0.32 points higher than 3rd and 5th voice respectively. Thus, people who have taken music classes in college don't weight the voice factor in their judgements as much as people who haven't taken such classes. The results are presented on the figure below.



There's an interaction between the 'Auxiliary measure of listener's ability to distinguish classical vs popular music' (`x16.minus.17`, whatever thath is!), and how classical the different voice leadings are. Since I don't

know what scale this is in nor what is the direction of the scale, I can only say that 1 SD increase on this scale leads to an even lower classical ratings for voice leadings 3rd and 5th compared to the contrary leading. The behavioral results are shown above. I haven't plotted the regression lines since the relation between this measure and overall classical ratings doesn't seem to be linear. There seems to be some kind of a sinusoidal pattern, but since I have no idea of what this measure is, I don't know if this makes any sense or if it's just an artifact of the data. Anyway, it looks like as the scores on this measure increase, the contrary voice becomes more distinct than the other two and is rated more as being more classical than them.

Excercise 3

In the previous excercise I showed that there's an interaction between the continuous variable 'self-identification of being a musician' (`selfdeclare`) and the classical ratings for the factor `harmony`. To explore further interactions, I refit the last model with a median split on that variable and explore the various interactions with other factors.

```
df1$musician <- factor(as.numeric(df1$selfdeclare > median(df1$selfdeclare)),
                       labels=c('Non-musician','Musician'))
table(df1$musician)
```

```
##
## Non-musician     Musician
##         1512         1008
```

This is as equal as we can get for the two groups. After examining a multitude of models, there were no other significant interactions than the one identified above. I'm not showing the results here for the sake of brevity.

7

I know less about the particular subset of the data you are using. In most other analyses of this data, there were interactions between Musician and both Harmony and PianoPlayer

showing me something would be useful. can't tell what you did or didn't do here...

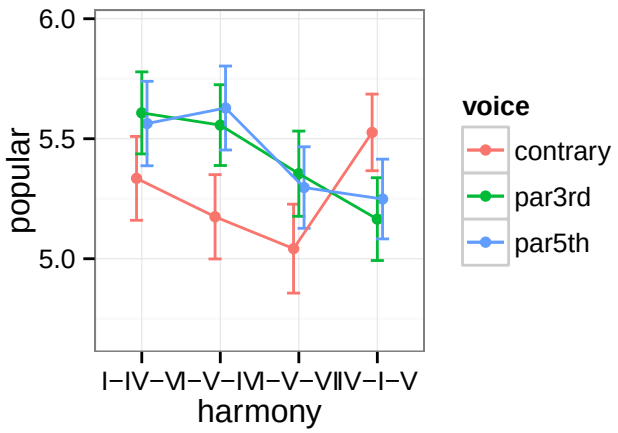
Excercise 4

Part (a)

Similarly to the effects of the three factors on the classical ratings, the main effects of instrument and harmony are significant, but the effect of voice is not. There's again a significant interaction between harmony and voice, which also removed the main effect harmony.

```
## Data: df1
## Models:
## pml1: popular ~ 1 + (1 | subject)
## pml2: popular ~ 1 + instrument + (1 | subject)
## pml3: popular ~ 1 + instrument + harmony + (1 | subject)
## pml4: popular ~ 1 + instrument + harmony + voice + (1 | subject)
## pml5: popular ~ 1 + instrument + harmony * voice + (1 | subject)
##      Df   AIC   BIC  logLik deviance    Chisq Chi Df Pr(>Chisq)
## pml1  3 11108 11125 -5551.0    11102
## pml2  5 10397 10426 -5193.6    10387 714.7984      2    < 2e-16 ***
## pml3  8 10395 10441 -5189.3    10379   8.5483      3    0.03594 *
## pml4 10 10394 10453 -5187.2    10374   4.1487      2    0.12563
## pml5 16 10393 10486 -5180.6    10361  13.2655      6    0.03901 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The behavior data for the interaction is presented on the following figure:



When we look at the estimats for each level of a factor, we can see that the tendencies are similar to those identified for the classical ratings. Piano and string pieces are rated as less popular sounding than guitar pieces, and string pieces are respectively less popular sounding than piano pieces. Voice leadings 3rd and 5th are rated as more popular sounding than the contrary voice leadings, but this relationship is reversed for the IV-I-V harmonies. For contrary voice leadings we see that IV-I-V is rated as most popular, while I-V-VI is least popular. For the 3rd and 5th voice leadings, IV-I-V is rated as least popular sounding.

```
pml6 <- lmer(popular ~ 1 + instrument + harmony*voice +
             (harmony-1|subject) + (instrument-1|subject),
             data=df1, control=lmerControl(optCtrl=list(maxfun=100000)))
summary(pml6, correlation=FALSE)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: popular ~ 1 + instrument + harmony * voice + (harmony - 1 | subject) +
##     (instrument - 1 | subject)
##    Data: df1
## Control: lmerControl(optCtrl = list(maxfun = 1e+05))
##
## REML criterion at convergence: 9923.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0304 -0.5712  0.0112  0.5791  3.2567
##
## Random effects:
##  Groups    Name              Variance Std.Dev. Corr
##  subject   harmonyI-IV-V     1.5014   1.2253
##            harmonyI-V-IV     1.8813   1.3716   0.98
##            harmonyI-V-VI     1.7611   1.3271   0.72  0.65
##            harmonyIV-I-V     1.3444   1.1595   0.91  0.83  0.57
##  subject.1 instrumentguitar 0.2697   0.5193
##            instrumentpiano   0.7810   0.8837   -0.37
##            instrumentstring  1.7629   1.3278   -1.00  0.38
##  Residual                    2.4615   1.5689
## Number of obs: 2492, groups:  subject, 70
##
## Fixed effects:
##                             Estimate Std. Error t value
## (Intercept)                  6.52449    0.19762   33.02
## instrumentpiano             -0.95858    0.16085   -5.96
## instrumentstring            -2.60412    0.23374  -11.14
## harmonyI-V-IV               -0.15834    0.15831   -1.00
## harmonyI-V-VI               -0.29192    0.19199   -1.52
## harmonyIV-I-V                0.19354    0.16483    1.17
## voicepar3rd                  0.27088    0.15353    1.76
## voicepar5th                  0.22995    0.15389    1.49
## harmonyI-V-IV:voicepar3rd    0.12832    0.21776    0.59
## harmonyI-V-VI:voicepar3rd    0.03654    0.21782    0.17
## harmonyIV-I-V:voicepar3rd   -0.63011    0.21752   -2.90
## harmonyI-V-IV:voicepar5th    0.22171    0.21802    1.02
## harmonyI-V-VI:voicepar5th    0.02010    0.21775    0.09
## harmonyIV-I-V:voicepar5th   -0.50990    0.21738   -2.35
```

Interestingly, when we look at the variance of the random effects, while classical ratings varied more for guitar than for string pieces, here's the opposite.

Part (b)

I repeated the itterative back-fit procedure I did for the classical ratings on the popular ratings. The end model used data from 63/70 subjects. Four covariates improved the model, though only through interactions with the within-subject factors. Those are:

`selfdeclare`, `x16.minus.17`, `x1990s2000s`, `collegemusic`

Below is presented a summary of the model coefficients. The interpretation of the main within-subject effects hasn't changed, though the inclusion of the covariates has made the effect of voice leadings significant. None

of the main effects of the covariates are significant. Similarly to classical ratings, the auxiliary measure of musical knowledge, `x16.minus.17` interacts with voice leadings in that as this measure increases so does the difference between popularity ratings in 3rd and 5th voice leadings compared to contrary leadings. Also similarly to the classical ratings, there's an interaction between voice leadings and whether the subject has taken music classes in college - if they have, they don't differentiate between the three voice leadings in terms of how popular they sound, which was the same conclusion as in the classical ratings.

A novel relationship compared to the classical ratings is an interaction between how much people have listen to pop and rock music from the 90s (`x1990s2000s`) and how they rate how popular the three voice leadings sound. The more pop and rock music from the 90s people have listened to, more popular sounding they rate the 3rd and 5th voice leadings to be compared to the contrary voice leadings. Finally, we observe a similar relationship between harmony and selfdeclare ratings as for the classical ratings. While the main effect of harmony is not significant, overall I-V-VI harmonies are rated as less popular sounding than the other three harmonies. As evidenced by the significant interaction, this effect increases for people who consider themselves to be musicians, which is the same result as we got from the classical ratings. Consider this to be the answer for part (c) of this excercise as well.

```
round(summary(pml13)$coefficients,3)
```

```
##                                Estimate Std. Error t value
## (Intercept)                       6.167      0.361  17.068
## instrumentpiano                  -1.032      0.171  -6.033
## instrumentstring                 -2.757      0.240 -11.498
## voicepar3rd                       0.698      0.234   2.986
## voicepar5th                       0.780      0.234   3.333
## harmonyI-V-IV                    -0.156      0.164  -0.949
## harmonyI-V-VI                    -0.344      0.193  -1.785
## harmonyIV-I-V                     0.175      0.172   1.017
## selfdeclare.c                     0.211      0.138   1.528
## x16.minus.17.c                    0.119      0.155   0.769
## x1990s2000s.c                    -0.014      0.100  -0.139
## collegemusic1                     0.482      0.369   1.306
## voicepar3rd:harmonyI-V-IV         0.180      0.226   0.793
## voicepar5th:harmonyI-V-IV         0.158      0.227   0.699
## voicepar3rd:harmonyI-V-VI         0.128      0.226   0.564
## voicepar5th:harmonyI-V-VI         0.017      0.227   0.076
## voicepar3rd:harmonyIV-I-V        -0.635      0.226  -2.808
## voicepar5th:harmonyIV-I-V        -0.563      0.226  -2.488
## voicepar3rd:x16.minus.17.c        0.273      0.087   3.129
## voicepar5th:x16.minus.17.c        0.296      0.087   3.387
## voicepar3rd:x1990s2000s.c         0.178      0.057   3.102
## voicepar5th:x1990s2000s.c         0.161      0.057   2.809
## voicepar3rd:collegemusic1        -0.521      0.210  -2.479
## voicepar5th:collegemusic1        -0.610      0.210  -2.901
## harmonyI-V-IV:selfdeclare.c       0.114      0.086   1.334
## harmonyI-V-VI:selfdeclare.c      -0.363      0.122  -2.987
## harmonyIV-I-V:selfdeclare.c      -0.003      0.097  -0.034
```

5. Brief writeup[1]

Results were analyzed via linear mixed effects regressions. Continuous covariates were centered around their mean and, where desirable, also scaled by their standard deviation. The three within-subject effects were entered step-wise in the model in the order presented below, and likelihood ratio tests were performed after each step. Significant covariates were identified by back-fitting a maximal model with all covariates and their interactions with the within-subject effects; non-significant estimates were droped based on likelihood ratio tests at the $\alpha = 0.05$ level. The final model included varying effects for the estimates of each instrument and harmony by subject.

There was a significant main effect of the type of harmony on classical ratings, $\Delta AIC = -24, LLR\chi^2(3) = 19.404, p < .001$. Pairwise comparisons revealed that the I-V-VI harmonies were rated on average as 0.77 points more classical than the other three harmonies (pairwise $t's \sim 4.4$), while they did not differ from each other (pairwise $t's \sim$ from -0.336 to 0.700). Voice leading had a significant effect on classical ratings, $\Delta AIC = -28, LLR\chi^2(2) = 32.262, p < .001$. Parallel 3rds and parallel 5ths were rated as less classical sounding than contrary motion ($\beta = -0.40, SE_\beta = 0.08, t = -5.18, p < .001$, and $\beta = -0.36, SE_\beta = 0.08, t = -4.66, p < .001$, respectively), while the two did not differ from each other. Classical ratings were also influenced by the type of instrument, $\Delta AIC = -84, LLR\chi^2(2) = 87.847, p < .001$. Pieces played by a piano were rated as $\beta = 1.37$ points more classical than those played by a guitar ($SE_\beta = 0.17, t = 7.97, p < .001$), and string pieces were rated $\beta = 1.75$ points more classical than those played by a piano ($SE_\beta = 0.18, t = 9.54, p < .001$).

There was also a significant interaction between harmony and voice leadings, $\Delta AIC = -24, \chi^2(6) = 36.078, p < .001$, that can be seen on the figure on page 12. While overall Parallel 3rds and 5ths were rated as less classical than contrary motion, this effect was largest for the I-V-VI harmony (additional decrease of $\beta = -0.67$ and $\beta = -0.42$ points for the Parallel 3rds and 5ths respectively, $t = -3.18, p < .01$ and $t = -1.98, p < .05$). For the IV-I-V harmony there was a reversal - Parallel 3rds were rated as more classical than contrary motion and Parallel 5ths ($\beta = 0.27$ and $\beta = 0.39$ respectively), which did not differ from each other.

10 Several covariates affected classical ratings. People's classical ratings decreased the more they considered themselves to be musicians ($\beta = -0.76, SE_\beta = 0.225, t = -3.38, p < .001$). This decrease was modulated by the type of harmony - it was present for harmonies I-IV-V, I-V-IV abd IV-I-V, but not for the I-IV-V harmony. Thus I-V-VI harmonies were rated as sounding more classical than the other three harmonies only by people who considered themselves to be musicians (see the figure on p. 17). Interestingly, the opposite relationship was present betweem the objective measure of musical knowledge OMSI and the classical ratings - 1 SD increase in the scores on this test led to $\beta = 0.47$ points higher classical ratings ($SE_\beta = 0.22, t = 2.13, p < .05$). The final measure of musical knowedge, whether people had taken music classes in college interacted wtih type of voice. They classical ratings were less influenced by voice leading type than those of people who haven't taken such classes, in that there was a smaller difference between their ratings of pieces contatin contrary motion and parallel 3rds/5ths (effect reduced respectively by $\beta = 0.31, t = 1.69, p > .05$ and $\beta = 0.45, t = 2.40, p < .05$, see figure on the bottom left of p. 18).

Futher, people who listen to classical music more often tended to rate the pieces as more classical sounding ($\beta = 0.34, SE_\beta = 0.10, t = 3.26$). There was a significant interaction between the auxiliary measure of people's ability to distinguish between classical and popular music and how they rated the different voice leadings. With increasing discriminatory ability the parallel 3rds and 5ths pieces were rated more and more less classical sounding than the contrary motion ($\beta = -0.20, t = -2.52$ and $\beta = -0.25, t = -3.083$ respectively). That is, the more people were able to distinguish classical from popular music, the more they discriminated between contrary and the parallel motions in their ratings.

[1] I am writing this up as I would for an empirical article in psychology.