practical_exercise_2, Methods 3, 2021, autumn semester

[Sara Krejberg]

[29/09/2021]

Assignment 1: Using mixed effects modelling to model hierarchical data

In this assignment we will be investigating the *politeness* dataset of Winter and Grawunder (2012) and apply basic methods of multilevel modelling.

Dataset

The dataset has been shared on GitHub, so make sure that the csv-file is on your current path. Otherwise you can supply the full path.

```
politeness <- read.csv('politeness.csv') ## read in data
```

Exercises and objectives

The objectives of the exercises of this assignment are:
1) Learning to recognize hierarchical structures within datasets and describing them
2) Creating simple multilevel models and assessing their fitness
3) Write up a report about the findings of the study

REMEMBER: In your report, make sure to include code that can reproduce the answers requested in the exercises below
REMEMBER: This assignment will be part of your final portfolio

Exercise 1 - describing the dataset and making some initial plots

1) Describe the dataset, such that someone who happened upon this dataset could understand the variables and what they contain
   subject: Participants participating in the experiment - f = female, m = male Gender: f = female , m = male Scenario: Describing which scenario the participant had to ask the question in Attitude: inf = informal, pol = polite, if they had to say the sentence polite or informal. Total_duration: is the time measured in seconds F0mn: frequency measured in hertz, also called pitch
   hiss_count: Number of loud breath taking of the participant.

   i. Also consider whether any of the variables in *politeness* should be encoded as factors or have the factor encoding removed. Hint: `?factor`

```
#changing to factors
politeness$gender <- as.factor(politeness$gender)
politeness$attitude <- as.factor(politeness$attitude)
politeness <- politeness %>%
  mutate(scenarioF = scenario)
politeness$scenarioF <- as.factor(politeness$scenarioF)
```

2) Create a new data frame that just contains the subject *F1* and run two linear models; one that expresses *f0mn* as dependent on *scenario* as an integer; and one that expresses *f0mn* as dependent on *scenario* encoded as a factor

   i. Include the model matrices, X from the General Linear Model, for these two models in your report and describe the different interpretations of *scenario* that these entail
   ii. Which coding of *scenario*, as a factor or not, is more fitting?

```
#filtering so we only use F1
polF1 <- politeness %>% filter(subject == 'F1')

#Model with scenario as factor
modelF <- lm(f0mn ~ scenarioF, data = polF1)
summary(modelF)
```

```
##
## Call:
## lm(formula = f0mn ~ scenarioF, data = polF1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -37.50 -13.86   0.00  13.86  37.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   212.75      20.35  10.453  1.6e-05 ***
## scenarioF2     62.40      28.78   2.168   0.0668 .
## scenarioF3     35.35      28.78   1.228   0.2591
## scenarioF4     53.75      28.78   1.867   0.1041
## scenarioF5     27.30      28.78   0.948   0.3745
## scenarioF6     -7.55      28.78  -0.262   0.8006
## scenarioF7    -14.95      28.78  -0.519   0.6195
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.78 on 7 degrees of freedom
## Multiple R-squared:  0.6576, Adjusted R-squared:  0.364
## F-statistic:  2.24 on 6 and 7 DF,  p-value: 0.1576
```

```
#Model with scenario as integer - X
modelI <- lm(f0mn ~ scenario, data = polF1)
summary(modelI)
```

```
##
## Call:
```

```
## lm(formula = f0mn ~ scenario, data = polF1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -44.836 -36.807   6.686  20.918  46.421
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  262.621     20.616  12.738 2.48e-08 ***
## scenario      -6.886      4.610  -1.494    0.161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.5 on 12 degrees of freedom
## Multiple R-squared:  0.1568, Adjusted R-squared:  0.0865
## F-statistic: 2.231 on 1 and 12 DF,  p-value: 0.1611
```

```
#making the model matrices
XF <- model.matrix(modelF)
#making the model matrices
XI <- model.matrix(modelI)
```

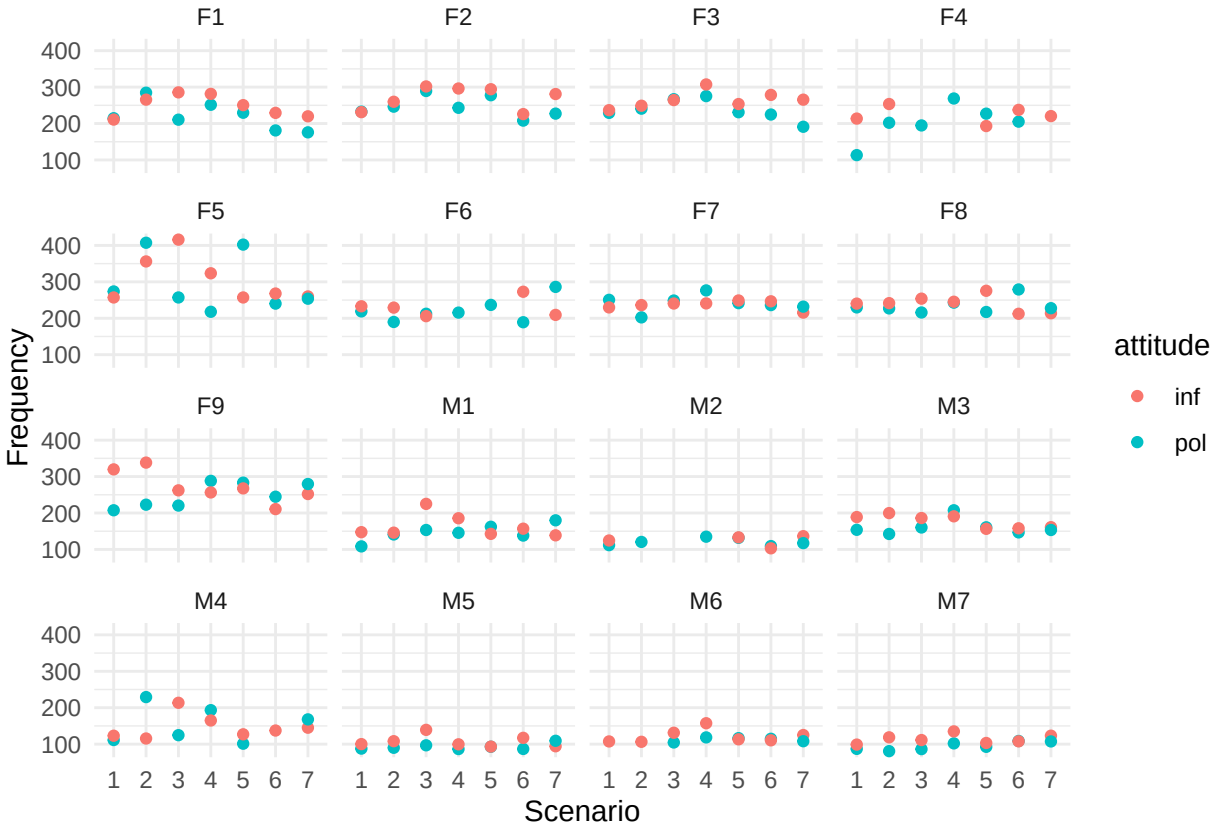
When using scenario as an integer its following the order in the scenarios, so it assume that theres a development from scenario to scenario. Where with factor it takes account for individual differences between the scenarios not thinking they depend on each other. Therefor the model where scenario is a factor is the best, because it dosent make sense to use the one where

3) Make a plot that includes a subplot for each subject that has *scenario* on the x-axis and *f0mn* on the y-axis and where points are colour coded according to *attitude*

    i. Describe the differences between subjects

```
politeness %>%
  ggplot(aes(scenarioF, f0mn, color = attitude))+
  geom_point()+
  facet_wrap(~subject)+
  theme_minimal()+
  xlab("Scenario")+
  ylab('Frequency')
```

```
## Warning: Removed 12 rows containing missing values (geom_point).
```



i.

Describe the differences between subjects¨

Exercise 2 - comparison of models

For this part, make sure to have `lme4` installed.
You can install it using `install.packages("lme4")` and load it using `library(lme4)`
`lmer` is used for multilevel modelling

1) Build four models and do some comparisons

   i. a single level model that models *f0mn* as dependent on *gender*

```
m1 <- lm(f0mn ~ gender, data = politeness)
summary(m1)
```

```
##
## Call:
## lm(formula = f0mn ~ gender, data = politeness)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -134.283  -24.928   -6.783   20.517  168.217
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  247.583      3.588   69.01   <2e-16 ***
## genderM     -115.821      5.476  -21.15   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 39.46 on 210 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.6806, Adjusted R-squared:  0.679
## F-statistic: 447.4 on 1 and 210 DF,  p-value: < 2.2e-16
```

ii. a two-level model that adds a second level on top of i. where unique intercepts are modelled for ea

```
m2 <- lmer(f0mn ~ gender + (1 | scenarioF), data = politeness, REML = FALSE)
summary(m2)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: f0mn ~ gender + (1 | scenarioF)
##    Data: politeness
##
##      AIC      BIC   logLik deviance df.resid
##   2162.3   2175.7  -1077.1   2154.3      208
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2617 -0.6192 -0.1537  0.4899  4.2318
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  scenarioF (Intercept)   71.82   8.475
##  Residual              1471.08  38.355
## Number of obs: 212, groups:  scenarioF, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  247.768      4.735   52.32
## genderM     -115.870      5.324  -21.76
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.483
```

iii. a two-level model that only has _subject_ as an intercept

```
m3 <- lmer(f0mn ~ gender + (1 | subject), data = politeness, REML = FALSE)
summary(m3)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: f0mn ~ gender + (1 | subject)
##    Data: politeness
##
##      AIC      BIC   logLik deviance df.resid
##   2112.0   2125.5  -1052.0   2104.0      208
##
## Scaled residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.2405 -0.5471 -0.1431  0.4360  3.8443
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept)  511.2   22.61
##  Residual             1026.7   32.04
## Number of obs: 212, groups:  subject, 16
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.547      8.083  30.501
## genderM     -115.193     12.239  -9.412
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.660
```

iv. a two-level model that models intercepts for both _scenario_ and _subject_

```
m4 <- lmer(f0mn ~ gender + (1 | scenarioF) + (1 | subject), data = politeness, REML = FALSE)
summary(m4)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: f0mn ~ gender + (1 | scenarioF) + (1 | subject)
##    Data: politeness
##
##      AIC      BIC   logLik deviance df.resid
##   2105.2   2122.0  -1047.6   2095.2      207
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0357 -0.5384 -0.1177  0.4346  3.7808
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  subject   (Intercept) 516.19   22.720
##  scenarioF (Intercept)  89.36    9.453
##  Residual              940.25   30.664
## Number of obs: 212, groups:  subject, 16; scenarioF, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.778      8.829  27.952
## genderM     -115.186     12.223  -9.424
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.604
```

v. which of the models has the lowest residual standard deviation, also compare the Akaike Information C

```
#Comparing the residual standard deviation
sigma(m1)
```

```
## [1] 39.46268
```

```
sigma(m2)
```

```
## [1] 38.3546
```

```
sigma(m3)
```

```
## [1] 32.04227
```

```
sigma(m4)
```

```
## [1] 30.66355
```

```
#comparing the Akaike information criterion
AIC(m1,m2,m3,m4)
```

```
##    df      AIC
## m1  3 2163.971
## m2  4 2162.257
## m3  4 2112.048
## m4  5 2105.176
```

```
#model 4 is best because it both has the lowest residual standard deviation and the lowest AIC
```

vi. which of the second-level effects explains the most variance?

```
r.squaredGLMM(m2)
```

```
## Warning: 'r.squaredGLMM' now calculates a revised statistic. See the help page.
```

```
##            R2m       R2c
## [1,] 0.6817304 0.6965456
```

```
#68% of the variance is explained by the model
r.squaredGLMM(m3)
```

```
##            R2m       R2c
## [1,] 0.6798832 0.7862932
```

```
#67% of the variance is explained by the model
```

Looking at the R2m for model 2 and 3 we see that model2 has the explained 68% of the variance. Therefor

2) Why is our single-level model bad?

   i. create a new data frame that has three variables, *subject*, *gender* and *f0mn*, where *f0mn* is the average of all responses of each subject, i.e. averaging across *attitude* and_scenario_

```
new.politeness <- politeness %>%
  filter(!is.na(f0mn)) %>%
  select(f0mn, attitude, subject) %>%
  group_by(subject) %>%
  summarise(mean = mean(f0mn))

new.politeness <- new.politeness %>%
  mutate(gender = if_else(grepl("F", new.politeness$subject, ignore.case = T), "F", "M")) %>%
  mutate(gender = as.factor(gender))
```

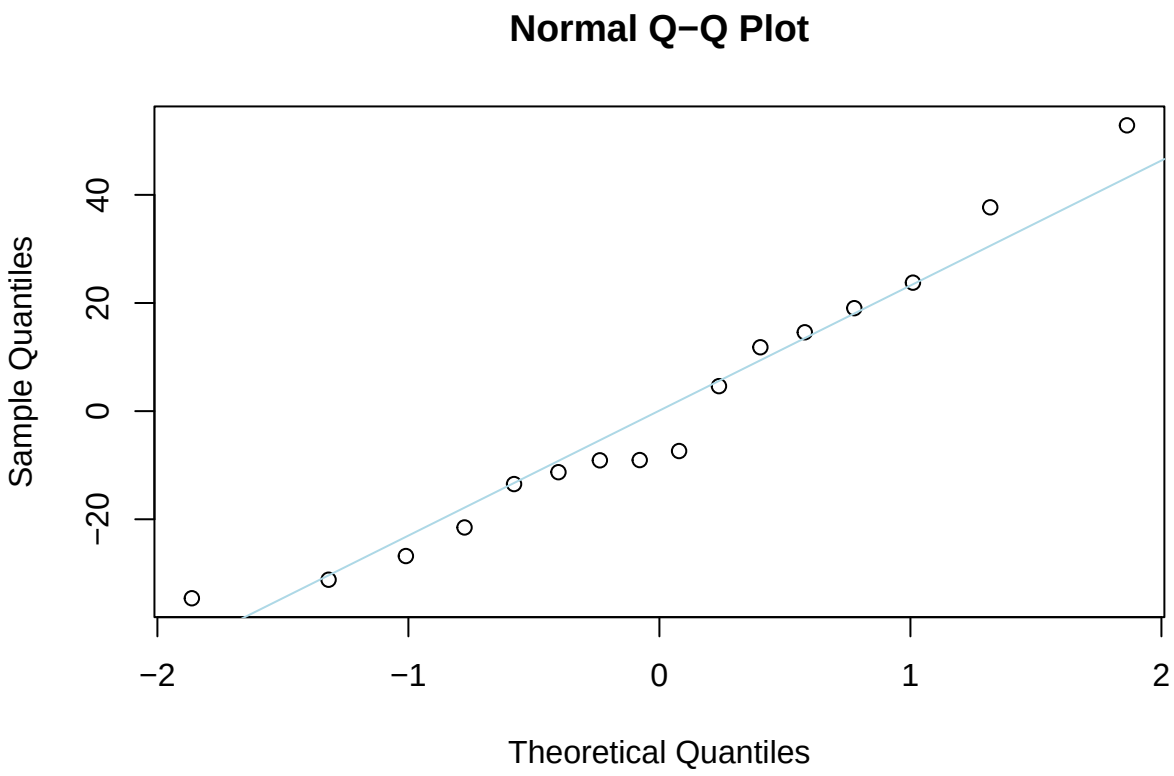
ii. build a single-level model that models _f0mn_ as dependent on _gender_ using this new dataset

```
ms <- lm(mean  ~ gender, data = new.politeness)
summary(ms)
```

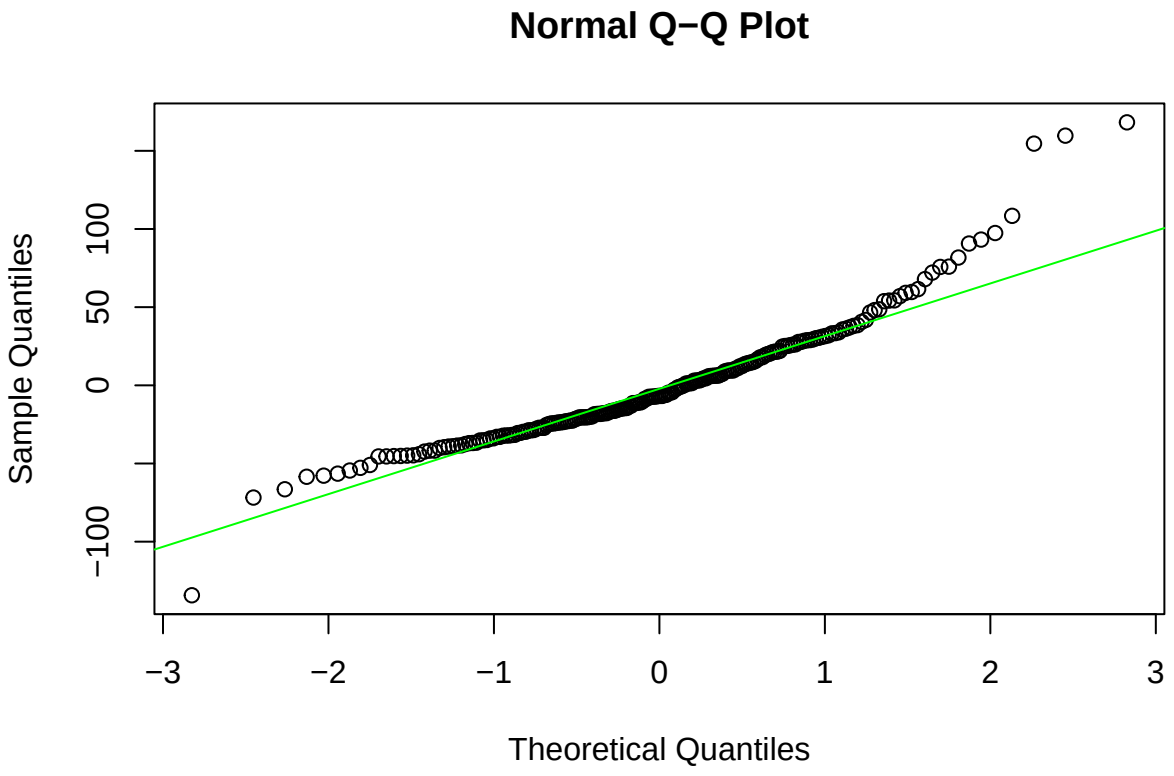
```
##
## Call:
## lm(formula = mean ~ gender, data = new.politeness)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.606 -15.493  -8.212  15.702  52.859
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  246.370      8.635  28.530 8.34e-14 ***
## genderM     -115.092     13.055  -8.816 4.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.91 on 14 degrees of freedom
## Multiple R-squared:  0.8474, Adjusted R-squared:  0.8365
## F-statistic: 77.72 on 1 and 14 DF,  p-value: 4.346e-07
```

iii. make Quantile-Quantile plots, comparing theoretical quantiles to the sample quantiles) using 'qqno

```
#the new single model
qqnorm(resid(ms))
qqline(resid(ms), col = 'lightblue')
```




```
#The old single model
qqnorm(resid(m1))
qqline(resid(m1), col = 'green')
```

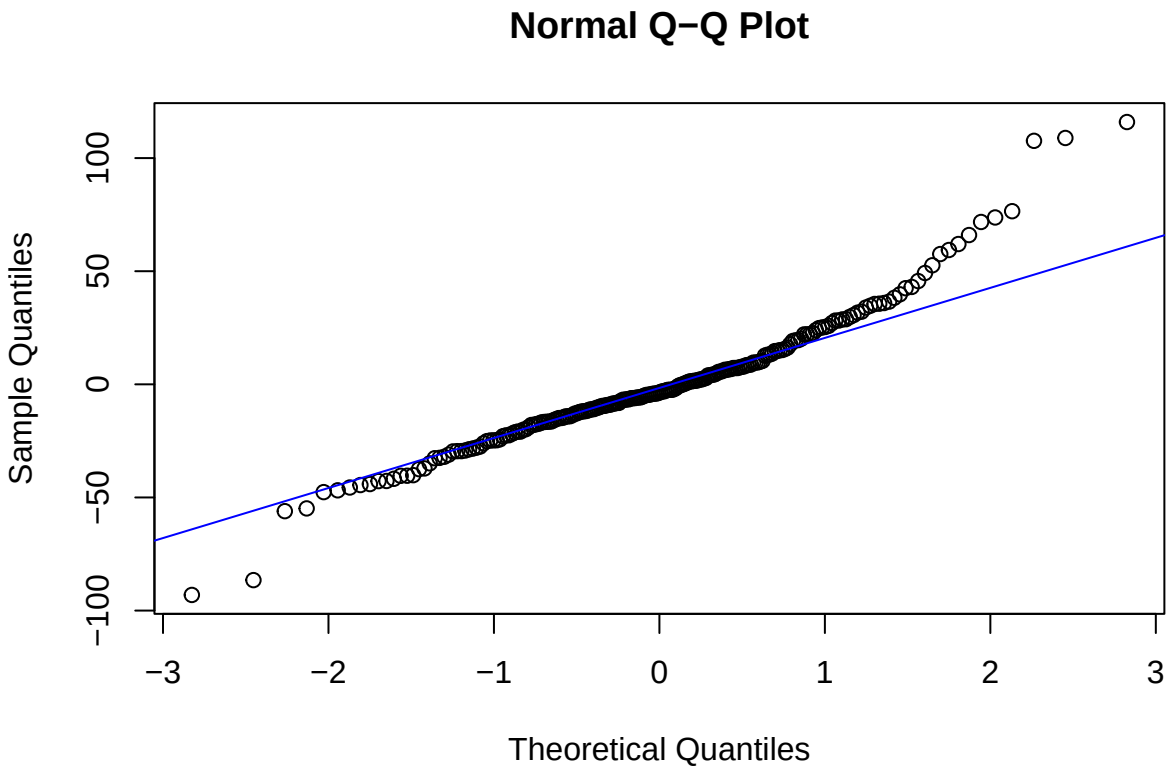


Looking at the qqplots its shown that the old model is pretty skewed and therefore the new model ms fulfil

the assumptions of general linear model.

iv. Also make a quantile-quantile plot for the residuals of the multilevel model with two intercepts. I

```
qqnorm(resid(m4))
qqline(resid(m4), col = 'blue')
```




```
#it looks okay, but its pretty right skewed. but most of the points are gathered in the middle around t
```

3) Plotting the two-intercepts model

   i. Create a plot for each subject, (similar to part 3 in Exercise 1), this time also indicating the fitted value for each of the subjects for each for the scenarios (hint use `fixef` to get the "grand effects" for each gender and `ranef` to get the subject- and scenario-specific effects)

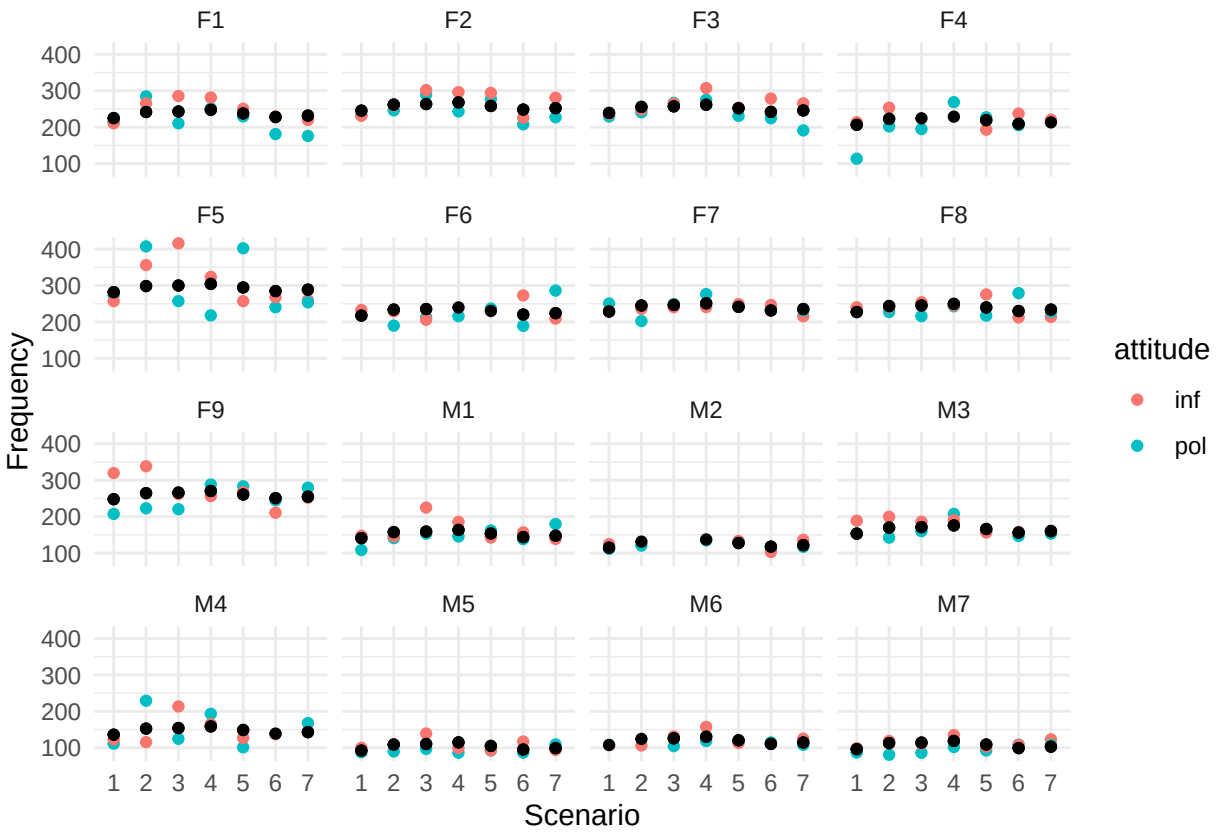
```
politeness.na <- politeness %>%
   filter(!is.na(f0mn))

fitted <- fitted(m4)

politeness.na <- cbind(politeness.na, fitted)

politeness.na %>%
  ggplot(aes(scenarioF, f0mn, color = attitude))+
  geom_point()+
  geom_point(aes(y = fitted), colour = 'black')+
  facet_wrap(~subject)+
  theme_minimal()+
```

```
  xlab("Scenario")+
  ylab('Frequency')
```



Exercise 3 - now with attitude

1) Carry on with the model with the two unique intercepts fitted (*scenario* and *subject*).
    i. now build a model that has *attitude* as a main effect besides *gender*

```
m5 <- lmer(f0mn ~ gender + attitude + (1 | scenarioF) + (1 | subject), data = politeness, REML = FALSE)
summary(m5)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: f0mn ~ gender + attitude + (1 | scenarioF) + (1 | subject)
##    Data: politeness
##
##      AIC      BIC   logLik deviance df.resid
##   2094.5   2114.6  -1041.2   2082.5      206
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8791 -0.5968 -0.0569  0.4260  3.9068
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
```

```
##  subject   (Intercept) 514.92   22.692
##  scenarioF (Intercept)  99.22    9.961
##  Residual              878.39   29.638
## Number of obs: 212, groups:  subject, 16; scenarioF, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  254.408      9.117  27.904
## genderM     -115.447     12.161  -9.494
## attitudepol  -14.817      4.086  -3.626
##
## Correlation of Fixed Effects:
##             (Intr) gendrM
## genderM     -0.583
## attitudepol -0.231  0.006
```

ii. make a separate model that besides the main effects of _attitude_ and _gender_ also include their i

```
m6 <- lmer(f0mn ~ gender * attitude + (1 | scenarioF) + (1 | subject), data = politeness, REML = FALSE)
summary(m6)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: f0mn ~ gender * attitude + (1 | scenarioF) + (1 | subject)
##    Data: politeness
##
##      AIC      BIC   logLik deviance df.resid
##   2096.0   2119.5  -1041.0   2082.0      205
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8460 -0.5893 -0.0685  0.3946  3.9518
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  subject   (Intercept) 514.09   22.674
##  scenarioF (Intercept)  99.08    9.954
##  Residual              876.46   29.605
## Number of obs: 212, groups:  subject, 16; scenarioF, 7
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)          255.632      9.289  27.521
## genderM             -118.251     12.841  -9.209
## attitudepol          -17.198      5.395  -3.188
## genderM:attitudepol    5.563      8.241   0.675
##
## Correlation of Fixed Effects:
##             (Intr) gendrM atttdp
## genderM     -0.605
## attitudepol -0.299  0.216
## gndrM:tttdp  0.195 -0.323 -0.654
```

iii. describe what the interaction term in the model says about Korean men's pitch when they are polite

Looking at the model we can see that the pitch of males are 118 lower than females. Also when being polite ones pitch also becomes lower. Looking at the interaction between pitch and attitude its shown that when the males are being polite their differences in their frequency are less than females.

2) Compare the three models (1. gender as a main effect; 2. gender and attitude as main effects;3. gender and attitude as main effects and the interaction between them. For all three models model unique intercepts for *subject* and *scenario*) using residual variance, residual standard deviation and AIC.

```
#Cheking AIC
AIC(m4,m5,m6)
```

```
##    df      AIC
## m4  5 2105.176
## m5  6 2094.489
## m6  7 2096.034
```

```
#Looking at the Residual standard deviation
sigma(m4)
```

```
## [1] 30.66355
```

```
sigma(m5)
```

```
## [1] 29.63771
```

```
sigma(m6)
```

```
## [1] 29.60505
```

```
#residual variance - what is not explained by the model
sum(residuals(m4)^2)
```

```
## [1] 181913
```

```
sum(residuals(m5)^2)
```

```
## [1] 169681.1
```

```
sum(residuals(m6)^2)
```

```
## [1] 169305.6
```

```
#R2 what is explained by the model
r.squaredGLMM(m4)
```

```
##            R2m       R2c
## [1,] 0.6787423 0.8045921
```

```
r.squaredGLMM(m5)
```

```
##            R2m       R2c
## [1,] 0.6899193 0.8175096
```

```
r.squaredGLMM(m6)
```

```
##            R2m       R2c
## [1,] 0.6904904 0.8178935
```

```
#model 5 is the best it has the lowest AIC, and when measuring the other parameters its very close to m
```

When comparing we see that model 6 is the best because it has the lowest residual variance and the lowest residual standard deviation.

3) Choose the model that you think describe the data the best - and write a short report on the main findings based on this model. At least include the following:

i. describe what the dataset consists of
The dataset consist of different parameters. Gender telling if the participant is a female or a male. Scenario is showing in which scenario the participants were placed and thereby had to have a attitude either informal or polite. It was measured in seconds under total duration. The pitch were measured in hertz, and the number of loud histing breath were also gathered.

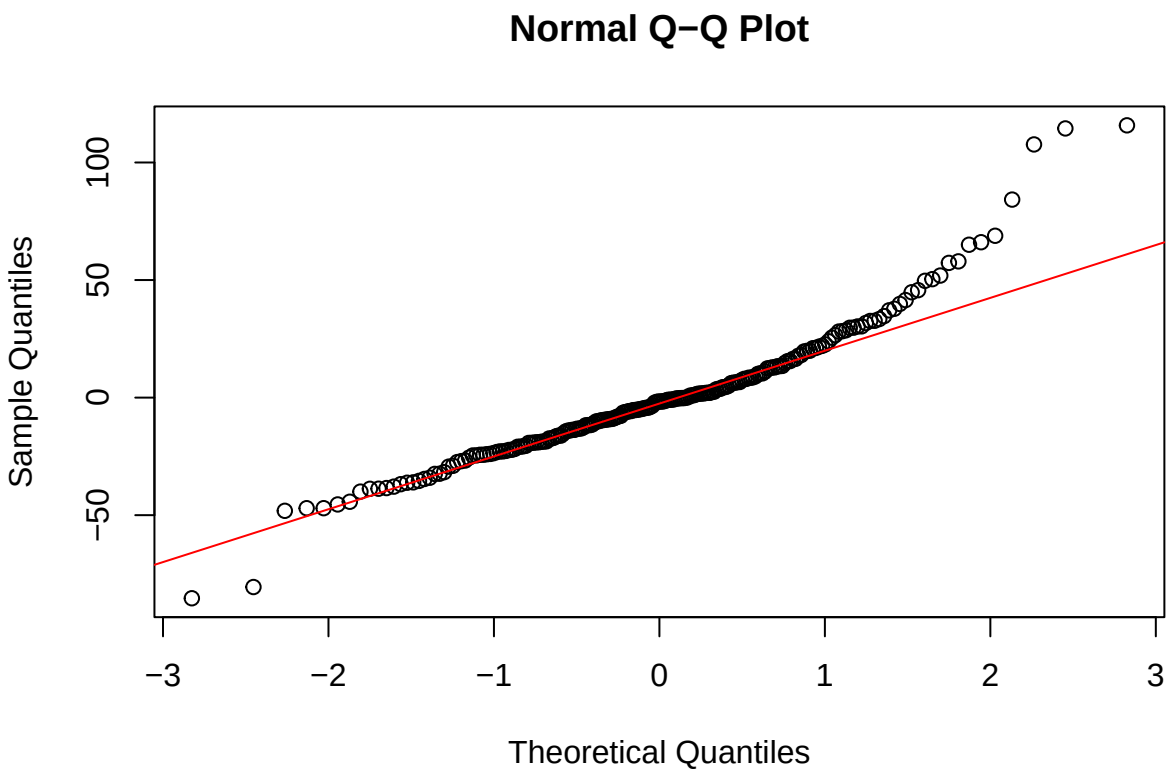
ii. what can you conclude about the effect of gender and attitude on pitch (if anything)?
It can be concluded that Korean male has lower pitch than females. Its also shown that when being polite the pitch is also lower.

iii. motivate why you would include separate intercepts for subjects and scenarios (if you think they should be included)
I included separate intercepts for subjects because the different participants has different pitch. Also the scenarios differs from each other and are not dependent ond eachother, therefore its good to make separate intercepts.

iv. describe the variance components of the second level (if any)

v. include a Quantile-Quantile plot of your chosen model

```
qqnorm(resid(m5))
qqline(resid(m5), col = 'red')
```


Normal Q−Q Plot
Sample Quantiles
Theoretical Quantiles
100
50
0
−50
−3
−2
−1
0
1
2
3