practical_exercise_2, Methods 3, 2021, autumn semester

Rikke Uldbæk

22/9 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
attach(politeness)
```

```
## The following object is masked from package:datasets:
##
##     attitude
```

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

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

Describe the dataset

Subject variable The subject variable represents each participant's individual ID, which makes it easier for us to distinguish between the participants and account for possible individual differences. This variable is a character string containing 16 different subject ID's, each subject ID has 14 rows

```
#additional information about subject variable
ls.str(politeness) #general information
```

```
## attitude :  chr [1:224] "pol" "inf" "pol" "inf" "pol" "inf" "pol" "inf" "pol" "inf" ...
## f0mn :  num [1:224] 215 211 285 266 211 ...
## gender :  chr [1:224] "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" ...
## hiss_count :  int [1:224] 2 0 0 0 0 0 1 0 1 0 ...
## scenario :  int [1:224] 1 1 2 2 3 3 4 4 5 5 ...
## subject :  chr [1:224] "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" ...
## total_duration :  num [1:224] 18.39 13.55 5.22 4.25 6.79 ...
```

```
length(unique(subject)) #amount of different participants
```

```
## [1] 16
```

Gender variable The gender variable represents the gender each participant identifies with. "F" corresponds to "Female" and "M" corresponds to "Male". 56.25% of the participants are Female and 43.75% are Male, which is also visually illustrated by a barplot. The variable is binary as it only contains either F or M, and for analytic purposes it will be transformed into a factor variable.

```
#additional information
n <- nrow(politeness)   # Number of rows in total
(percent_gender <- table(politeness$gender)/n * 100) #generating percentages of gender
```

```
##
##     F     M
## 56.25 43.75
```

```
ls.str(politeness) #general information
```
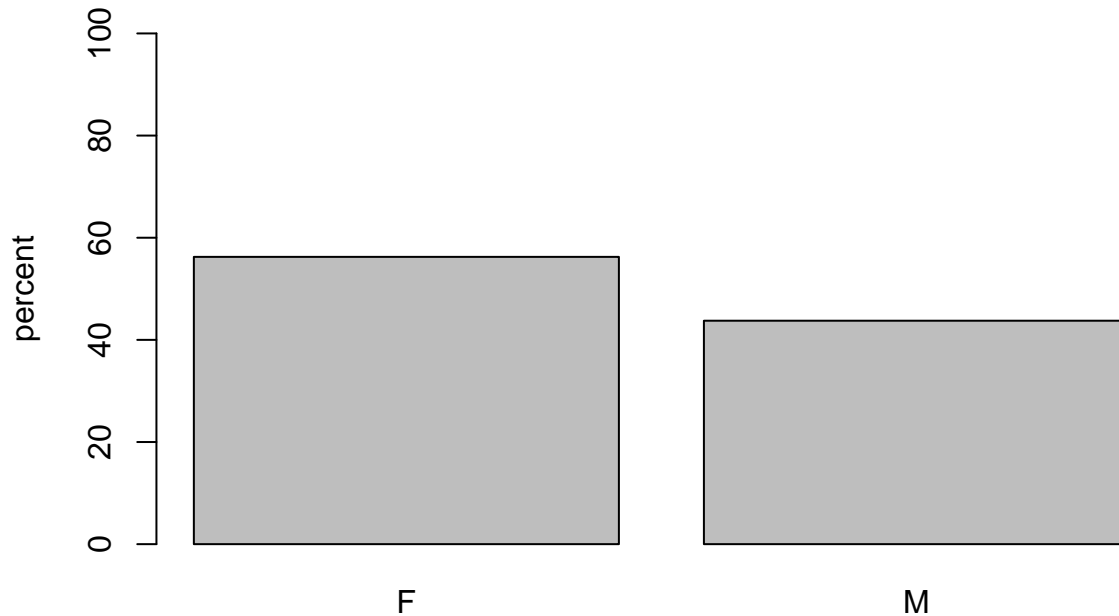
```
## attitude :  chr [1:224] "pol" "inf" "pol" "inf" "pol" "inf" "pol" "inf" "pol" "inf" ...
## f0mn :  num [1:224] 215 211 285 266 211 ...
## gender :  chr [1:224] "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" "F" ...
## hiss_count :  int [1:224] 2 0 0 0 0 0 1 0 1 0 ...
## scenario :  int [1:224] 1 1 2 2 3 3 4 4 5 5 ...
## subject :  chr [1:224] "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" ...
## total_duration :  num [1:224] 18.39 13.55 5.22 4.25 6.79 ...
```

```
politeness$gender <- as.factor(politeness$gender)
```

```
#gender variable plot
barplot(percent_gender,ylim=c(0,100), ylab="percent",main="Barplot of Gender")
```

Barplot of Gender

#####Scenario variable The scenario variable is an integer-variable and represents 7 different trials/scenarios, where each participant completes each trial/scenario twice, hence 14 rows per participant. Each participants completes these 7 different trials twice as each scenario is either "informal" or "polite". For analytic purposes it will be transformed into a factor variable.

```r
#additional information
ls.str(politeness) #general information
```

```
## attitude :  chr [1:224] "pol" "inf" "pol" "inf" "pol" "inf" "pol" "inf" "pol" "inf" ...
## f0mn :  num [1:224] 215 211 285 266 211 ...
## gender :  Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
## hiss_count :  int [1:224] 2 0 0 0 0 0 1 0 1 0 ...
## scenario :  int [1:224] 1 1 2 2 3 3 4 4 5 5 ...
## subject :  chr [1:224] "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" ...
## total_duration :  num [1:224] 18.39 13.55 5.22 4.25 6.79 ...
```

```r
politeness$scenario <- as.factor(politeness$scenario)
```

#####Attitude variable The attitude variable is a binary variable containing two options i.e "inf" which corresponds to "informal" and "pol" which corresponds to "formal". Each trial/scenario (n = 14) is either "informal" or "polite", making it 50% "informal" and 50% "polite".For analytic purposes it will be transformed into a factor variable.

```
#additional information
n <- nrow(politeness)   # Number of rows in total
(percent_attitude <- table(politeness$attitude)/n * 100) #generating percentages of gender
```

```
##
## inf pol
##  50  50
```

```
ls.str(politeness) #general information
```

```
## attitude :  chr [1:224] "pol" "inf" "pol" "inf" "pol" "inf" "pol" "inf" "pol" "inf" ...
## f0mn :  num [1:224] 215 211 285 266 211 ...
## gender :  Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
## hiss_count :  int [1:224] 2 0 0 0 0 0 1 0 1 0 ...
## scenario :  Factor w/ 7 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
## subject :  chr [1:224] "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" ...
## total_duration :  num [1:224] 18.39 13.55 5.22 4.25 6.79 ...
```

```
politeness$attitude <- as.factor(politeness$attitude)
```

```
#Attitude variable plot
barplot(percent_attitude,ylim=c(0,100), ylab="percent",main="Barplot of Attitude")
```
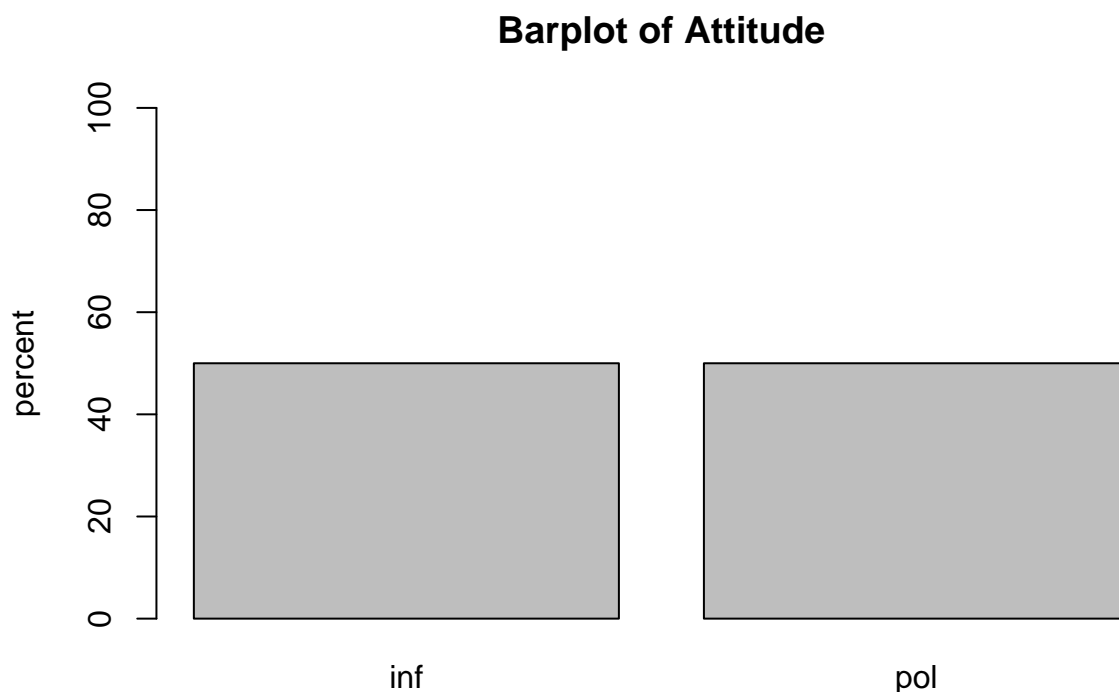

#####Total duration variable The total duration variable is a numeric variable representing duration of each response (pitch). The distribution of the variable is positively skewed (right-skewed distribution).The

minimum value in the variable is 0.988 and the maximum value in the variable is 101.375, in addition to this the mean of the variable is 24.176.

```
#total duration variable plot
ggplot(politeness, aes(total_duration))+geom_density()
```
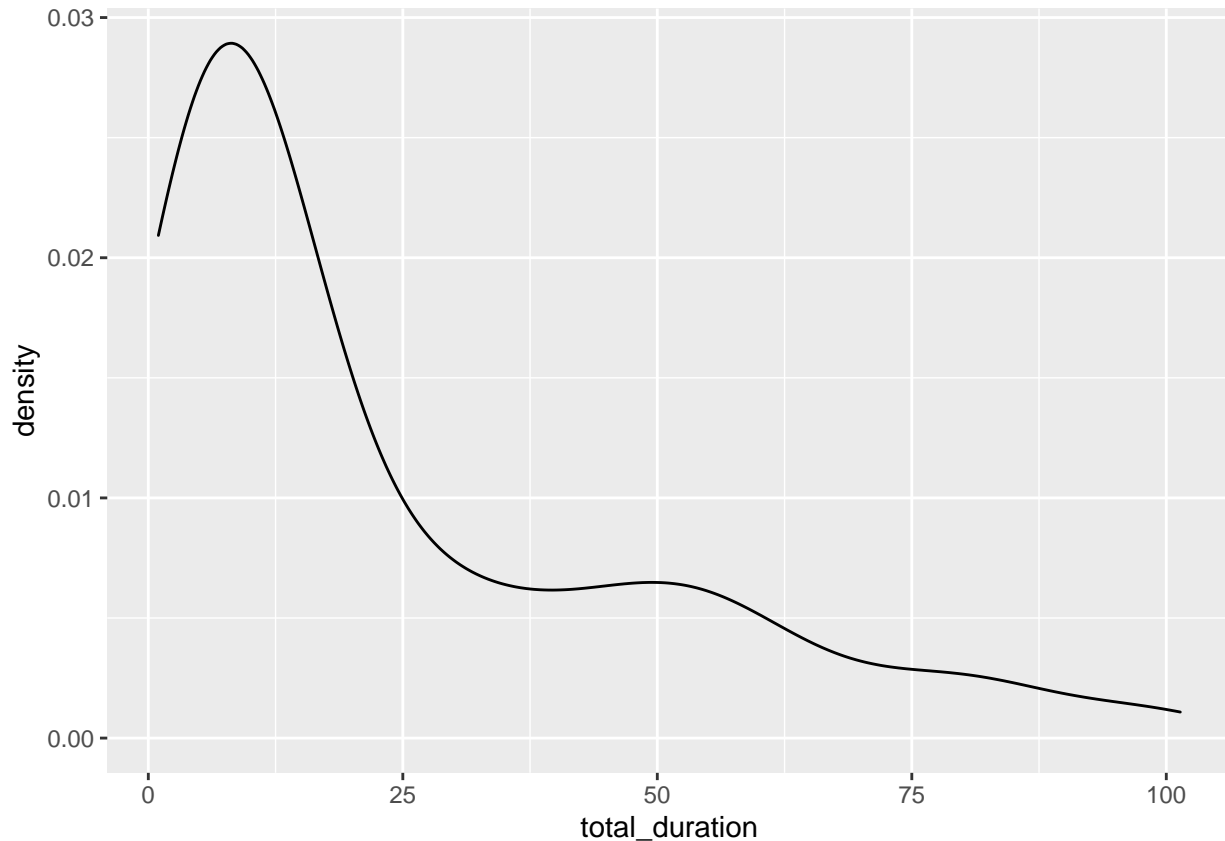


```
#additional information
ls.str(politeness) #general information
```

```
## attitude :  Factor w/ 2 levels "inf","pol": 2 1 2 1 2 1 2 1 2 1 ...
## f0mn :  num [1:224] 215 211 285 266 211 ...
## gender :  Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
## hiss_count :  int [1:224] 2 0 0 0 0 0 1 0 1 0 ...
## scenario :  Factor w/ 7 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
## subject :  chr [1:224] "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" ...
## total_duration :  num [1:224] 18.39 13.55 5.22 4.25 6.79 ...
```

```
summary(total_duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.988   6.140  12.640  24.176  39.373 101.375
```

#####F0mn variable F0mn represents the frequency of the voice in hertz i.e the pitch of voice in hertz. This variable is numeric and the distribution of the data is bimodal. The minimum value in the variable

is 80.8 hz and the maximum value in the variable is 415.8 hz, in addition to this the mean of the variable is 197.9 hz. The f0mn variable contains some NA's, and for analytic purposes, these NA's will be replaced with a mean of f0mn for each subject.

```
#f0mn variable plot
ggplot(politeness, aes(f0mn))+geom_density()
```

```
## Warning: Removed 12 rows containing non-finite values (stat_density).
```
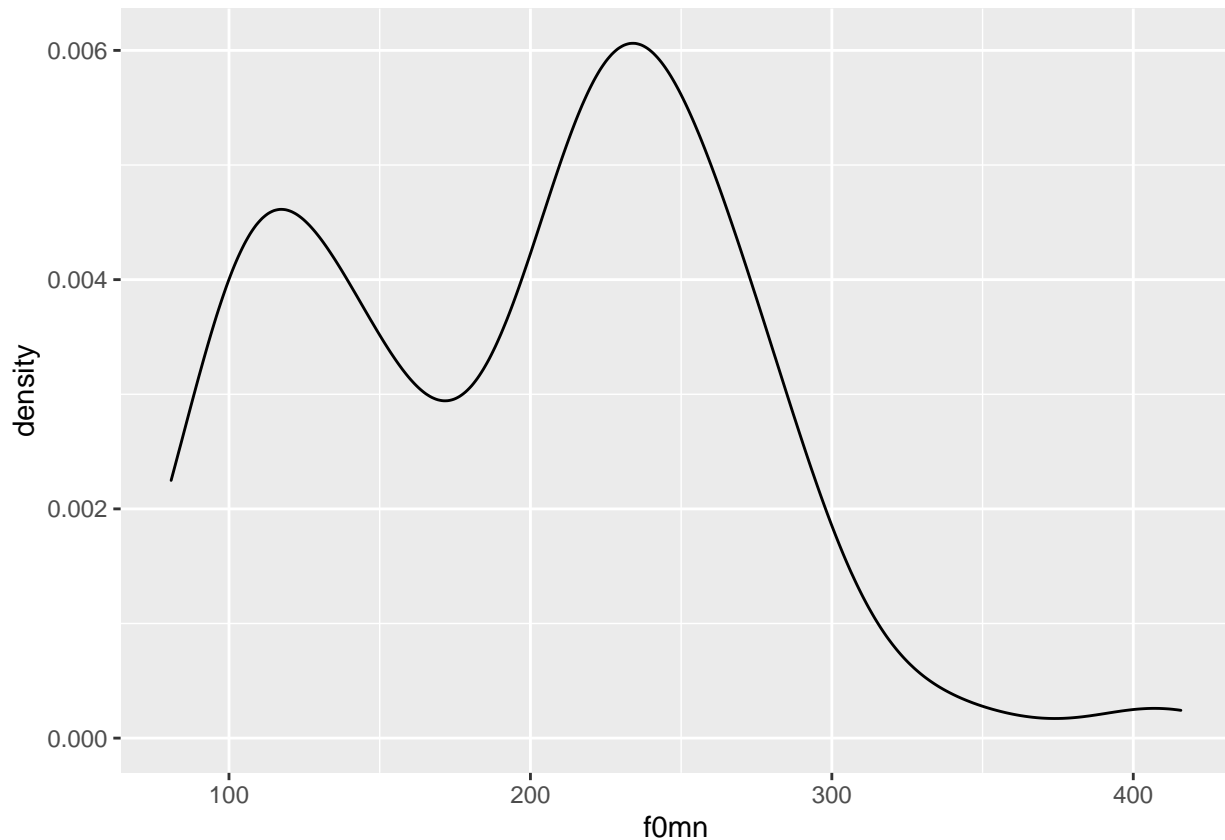


```
#additional information
ls.str(politeness) #general information
```

```
## attitude :  Factor w/ 2 levels "inf","pol": 2 1 2 1 2 1 2 1 2 1 ...
## f0mn :  num [1:224] 215 211 285 266 211 ...
## gender :  Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
## hiss_count :  int [1:224] 2 0 0 0 0 0 1 0 1 0 ...
## scenario :  Factor w/ 7 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
## subject :  chr [1:224] "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" ...
## total_duration :  num [1:224] 18.39 13.55 5.22 4.25 6.79 ...
```

```
summary(f0mn)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    80.8   134.5   211.6   197.9   247.1   415.8      12
```

```
#replacing NA's with mean pr participant
politeness <- politeness %>%
  group_by(subject) %>%
  mutate(f0mn = ifelse(is.na(f0mn), mean(f0mn, na.rm = TRUE), f0mn))
```

#####Hiss count variable The hiss count variable relates to the amount of loud "hissing" breath intake. This variable is an integer variable with a minimum value hissings of 0 and a maximum hissing value of 5. The distribution of the data is positively skewed (right-skewed distribution).

```
#hiss count variable plot
ggplot(politeness, aes(hiss_count))+geom_density()
```
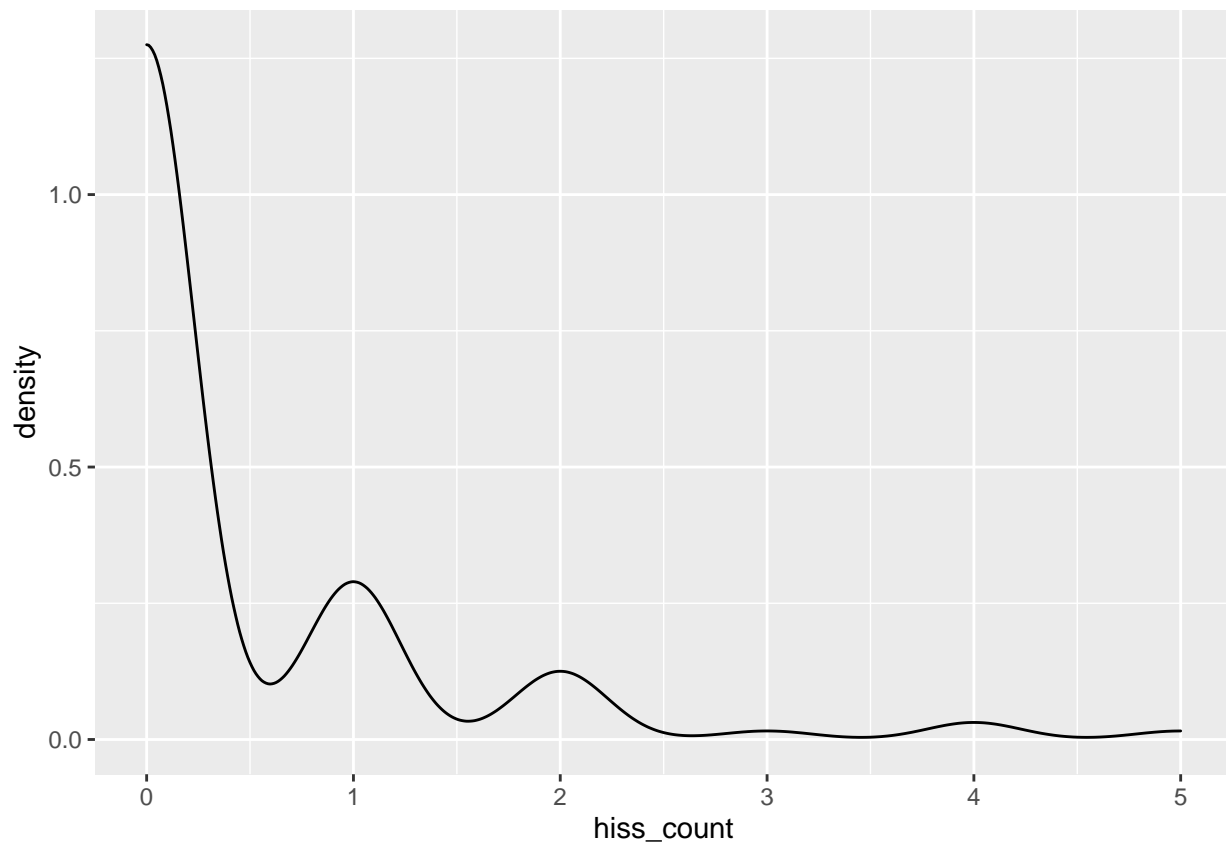


```
#additional information:
ls.str(politeness) #general information
```

```
## attitude :  Factor w/ 2 levels "inf","pol": 2 1 2 1 2 1 2 1 2 1 ...
## f0mn :  num [1:224] 215 211 285 266 211 ...
## gender :  Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
## hiss_count :  int [1:224] 2 0 0 0 0 0 1 0 1 0 ...
## scenario :  Factor w/ 7 levels "1","2","3","4",..: 1 1 2 2 3 3 4 4 5 5 ...
## subject :  chr [1:224] "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" "F1" ...
## total_duration :  num [1:224] 18.39 13.55 5.22 4.25 6.79 ...
```

7

```r
summary(hiss_count)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0000  0.0000  0.4509  1.0000  5.0000
```

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?

###Integer VS Factor Linear Model

#####Model 1 - scenario as integer Model 1 expresses f0mn as dependent on scenario as an integer. When the "scenario" variable is integer, it is treated as numeric by R, meaning we would interpret the slope in the summary of the model as follows: per increment of scenario the pitch will go down by 6.886 hz. This is a faulty interpretation, as the 2x2 matrix show that we would then interpret scenario 7, as being 7 times more than scenario 1, and this is not the case.

#####Model 2 - scenario as factor Model 2 expresses f0mn as dependent on scenario encoded as a factor. When the "scenario" variable is factor, it is treated categorically by R, which enables us to interpret the output of the model 2 correctly as follows; going from scenario 1 to scenario 2 the pitch increases by 62.40 hz, going from scenario 1 to scenario 3 the pitch increases by 35.35 hz etc. The model matrix of model 2 shows multiple columns (one for each scenario), meaning that we can look at each scenario individually, and thereby investigate their meaning properly.

Thus, coding the "scenario" variable as a factor provides the most accurate information, and eliminates any hierarchical relations in the variable.

```r
#creating a data frame with only subject F1
df_F1 <- politeness %>%
  filter(subject == "F1") %>%
  mutate(scenario_factor = as.factor(scenario)) %>%
  mutate(scenario_integer = as.integer(scenario))

#model 1 with "f0mn" as y and "scenario" as x (INTEGER)
model_integer <- lm(f0mn~scenario_integer, data=df_F1)
summary(model_integer)
```

```
##
## Call:
## lm(formula = f0mn ~ scenario_integer, data = df_F1)
##
## 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_integer   -6.886      4.610  -1.494    0.161
## ---
```

8

```
## 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
```

```r
model1_X <- model.matrix(model_integer)
model1_X
```

```
##    (Intercept) scenario_integer
## 1            1                1
## 2            1                1
## 3            1                2
## 4            1                2
## 5            1                3
## 6            1                3
## 7            1                4
## 8            1                4
## 9            1                5
## 10           1                5
## 11           1                6
## 12           1                6
## 13           1                7
## 14           1                7
## attr(,"assign")
## [1] 0 1
```

```r
#model 2 with "f0mn" as y and "scenario" as x (FACTOR)
model_factor <- lm(f0mn~scenario_factor, data=df_F1)
summary(model_factor)
```

```
##
## Call:
## lm(formula = f0mn ~ scenario_factor, data = df_F1)
##
## 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 ***
## scenario_factor2    62.40      28.78   2.168   0.0668 .
## scenario_factor3    35.35      28.78   1.228   0.2591
## scenario_factor4    53.75      28.78   1.867   0.1041
## scenario_factor5    27.30      28.78   0.948   0.3745
## scenario_factor6    -7.55      28.78  -0.262   0.8006
## scenario_factor7   -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
```

9

```
model2_X <- model.matrix(model_factor)
model2_X
```

```
##    (Intercept) scenario_factor2 scenario_factor3 scenario_factor4
## 1            1                0                0                0
## 2            1                0                0                0
## 3            1                1                0                0
## 4            1                1                0                0
## 5            1                0                1                0
## 6            1                0                1                0
## 7            1                0                0                1
## 8            1                0                0                1
## 9            1                0                0                0
## 10           1                0                0                0
## 11           1                0                0                0
## 12           1                0                0                0
## 13           1                0                0                0
## 14           1                0                0                0
##    scenario_factor5 scenario_factor6 scenario_factor7
## 1                 0                0                0
## 2                 0                0                0
## 3                 0                0                0
## 4                 0                0                0
## 5                 0                0                0
## 6                 0                0                0
## 7                 0                0                0
## 8                 0                0                0
## 9                 1                0                0
## 10                1                0                0
## 11                0                1                0
## 12                0                1                0
## 13                0                0                1
## 14                0                0                1
## attr(,"assign")
## [1] 0 1 1 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$scenario_factor
## [1] "contr.treatment"
```

```
#as concluded, coding "scenario" as factor makes most sense
politeness$scenario <- as.factor(politeness$scenario)
```

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

###Creating Subplots The main difference between each subject is the pitch of voice in hertz. The subplots show that the male subjects tend to have lower pitch of voice, and the female subjects tend to have higher pitch of voice. Likewise, the subplots demonstrates how there are individual differences within pitch of voice, which we need to account for when modeling the data.

10

```
#plot
subplot <- ggplot(data = politeness, aes(x = scenario, y = f0mn, color = attitude)) +
  geom_point()+
  facet_wrap(~subject);print(subplot)
```
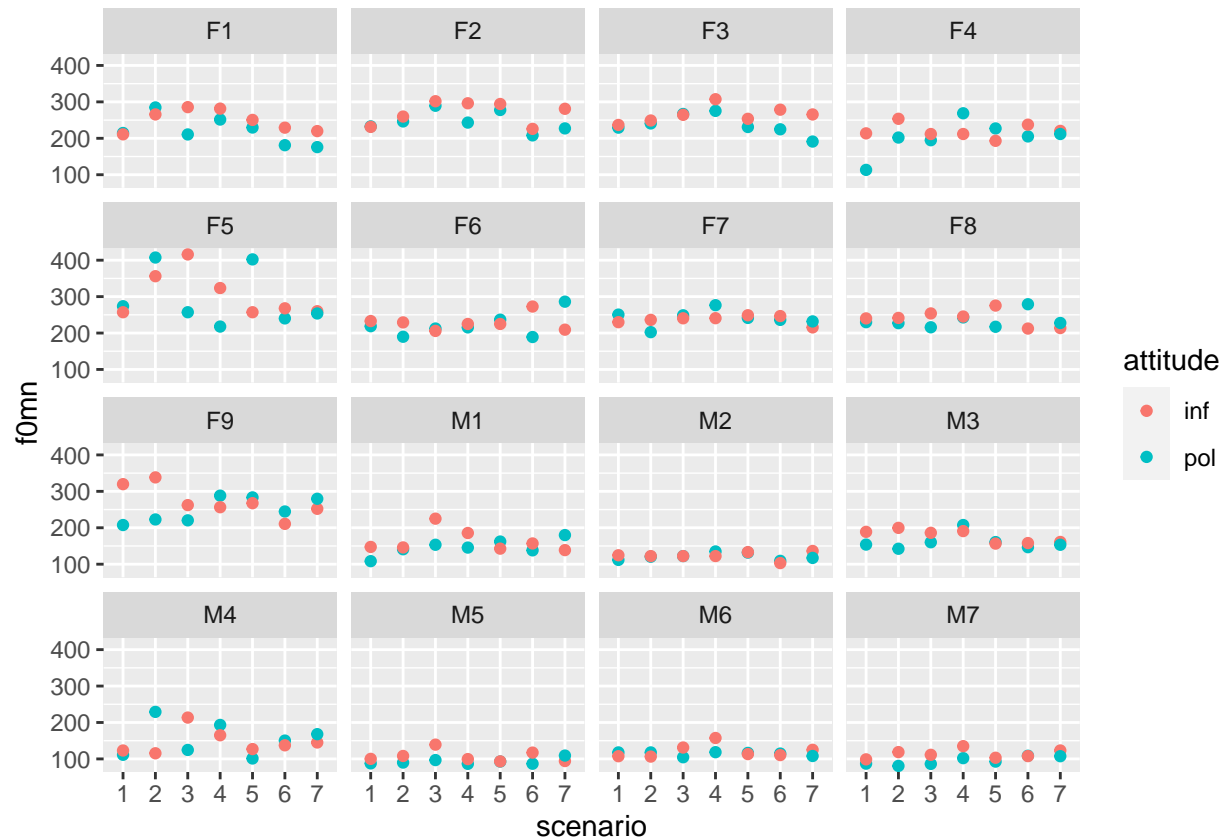

Exercise 2 - comparison of models

1) Build four models and do some comparisons

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

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

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

    iv. a two-level model that models intercepts for both *scenario* and *subject*

    v. which of the models has the lowest residual standard deviation, also compare the Akaike Information Criterion `AIC`?

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

###Building Four Models and Compare Them

When comparing the four models by residual standard deviation and Akaike Information Criterion, the fourth model (model 4) has the lowest residual standard deviation of 29.86727, likewise does the fourth model have the lowest AIC value of 2198.778. When comparing only the second-levels effects (model 2 and model 3) by their Conditional R2 (R2c) which states how much variance is explained by both random and fixed effects, it is shown that model 2 has an R2c of 0.698216 and model 3 has an R2c of 0.8002449, and thus stating that the second-level effect "subject" explains the most variance.

11

```
#model 1 - single level model that models _f0mn_ as dependent on _gender_
model1 <- lm(f0mn ~ gender, data=politeness)
summary(model1)
```

```
##
## Call:
## lm(formula = f0mn ~ gender, data = politeness)
##
## Residuals:
##       Min       1Q   Median       3Q      Max
## -133.070  -24.128   -6.728   19.530  169.430
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  246.370      3.449   71.43   <2e-16 ***
## genderM     -115.092      5.215  -22.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.72 on 222 degrees of freedom
## Multiple R-squared:  0.6869, Adjusted R-squared:  0.6855
## F-statistic: 487.1 on 1 and 222 DF,  p-value: < 2.2e-16
```

```
sigma(model1)   #38.71852
```

```
## [1] 38.71852
```

```
#model 2 - two-level model that adds a second level on top of i. where unique intercepts are modelled f
model2 <- lmer(f0mn ~ gender+ (1|scenario), data=politeness)
summary(model2)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2259.7
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2787 -0.6107 -0.1521  0.5110  4.3331
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  scenario (Intercept)   64.64   8.04
##  Residual             1443.22  37.99
## Number of obs: 224, groups:  scenario, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.370      4.549   54.16
## genderM     -115.092      5.117  -22.49
##
```

```
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.492
```

```r
sigma(model2)   #37.98968
```

```
## [1] 37.98968
```

```r
#model 3 - two-level model that only has _subject_ as an intercept
model3 <- lmer(f0mn ~ gender+ (1|subject), data=politeness)
summary(model3)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + (1 | subject)
##    Data: politeness
##
## REML criterion at convergence: 2197.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2799 -0.5085 -0.1146  0.3978  3.9225
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 602.0    24.54
##  Residual             967.6    31.11
## Number of obs: 224, groups:  subject, 16
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.370      8.635  28.530
## genderM     -115.092     13.055  -8.816
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.661
```

```r
sigma(model3)   #31.10665
```

```
## [1] 31.10665
```

```r
#model 4 - two-level model that models intercepts for both scenario and subject
model4 <- lmer(f0mn ~ gender + (1|subject) + (1|scenario), data=politeness)
summary(model4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + (1 | subject) + (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2188.8
##
```

```
## Scaled residuals:
##      Min      1Q  Median      3Q     Max
## -3.0458 -0.5030 -0.1125  0.4029  3.8668
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  subject   (Intercept) 607.40   24.646
##  scenario  (Intercept)  81.87    9.048
##  Residual              892.05   29.867
## Number of obs: 224, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.370      9.288  26.526
## genderM     -115.092     13.055  -8.816
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.615
```

```r
sigma(model4)  #29.86727 LOWEST
```

```
## [1] 29.86727
```

```r
#comparing AIC values
AIC(model1, model2, model3, model4)
```

```
##        df      AIC
## model1  3 2277.706
## model2  4 2267.733
## model3  4 2205.469
## model4  5 2198.778
```

```r
#comparing conditional R2 within the second levels
MuMIn::r.squaredGLMM(model2) #R2c = 0.698216
```

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

```
##           R2m      R2c
## [1,] 0.6846987 0.698216
```

```r
MuMIn::r.squaredGLMM(model3) #R2c = 0.8002449
```

```
##           R2m       R2c
## [1,] 0.6759677 0.8002449
```

```r
MuMIn::r.squaredGLMM(model4) #R2c = 0.816289
```

```
##           R2m      R2c
## [1,] 0.6743396 0.816289
```

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_

    ii. build a single-level model that models *f0mn* as dependent on *gender* using this new dataset

   iii. make Quantile-Quantile plots, comparing theoretical quantiles to the sample quantiles) using `qqnorm` and `qqline` for the new single-level model and compare it to the old single-level model (from 1).i). Which model's residuals (ϵ) fulfil the assumptions of the General Linear Model better?)

   iv. Also make a quantile-quantile plot for the residuals of the multilevel model with two intercepts. Does it look alright?

Why is our single-level model bad?

When comparing the Normal Q-Q-plots for the new single-level model and the old single-level model neither of them see to fulfill the assumptions of normally distributed residuals completely, as the dots deviate a little from the line in both plots. Especially the Normal Q-Q-plots for the old single-level model seems to show signs of pattern with its bending end, which indicates a right skewed tail.

Normality of residuals is not the only violation of assumptions. Likewise, does the data-set contain repeated measures per participant, thus violating the assumption of independence. A mixed effects model is therefore the appropriate model to use, and not a single-level linear model.

Both a residual plot with residuals VS fitted and a Q-Q-plot of the model were assessed. The residual plot of the multilevel model with two intercepts (model 4) looks fine, as its residuals are distributed as a cloud of dots. The Q-Q-plot for the residuals of model 4, looks a little better than model 1 and the single-level model, but still has a right skewed tail. This violation of normality of residuals can be accepted since linear models are relatively robust against violations of the assumptions of normality. This model does not violate the assumption of independence, which is more important.

```
#new df with subject, gender & average f0mn per subject
df2 <- politeness %>%
  group_by(subject, gender) %>%
  summarize(mean_f0mn= mean(f0mn))
```

```
## `summarise()` regrouping output by 'subject' (override with `.groups` argument)
```

```
#single level model from the new dataset
single_level_model <- lm(mean_f0mn ~ gender, data= df2,)
summary(single_level_model)
```

```
##
## Call:
## lm(formula = mean_f0mn ~ gender, data = df2)
##
## 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
```
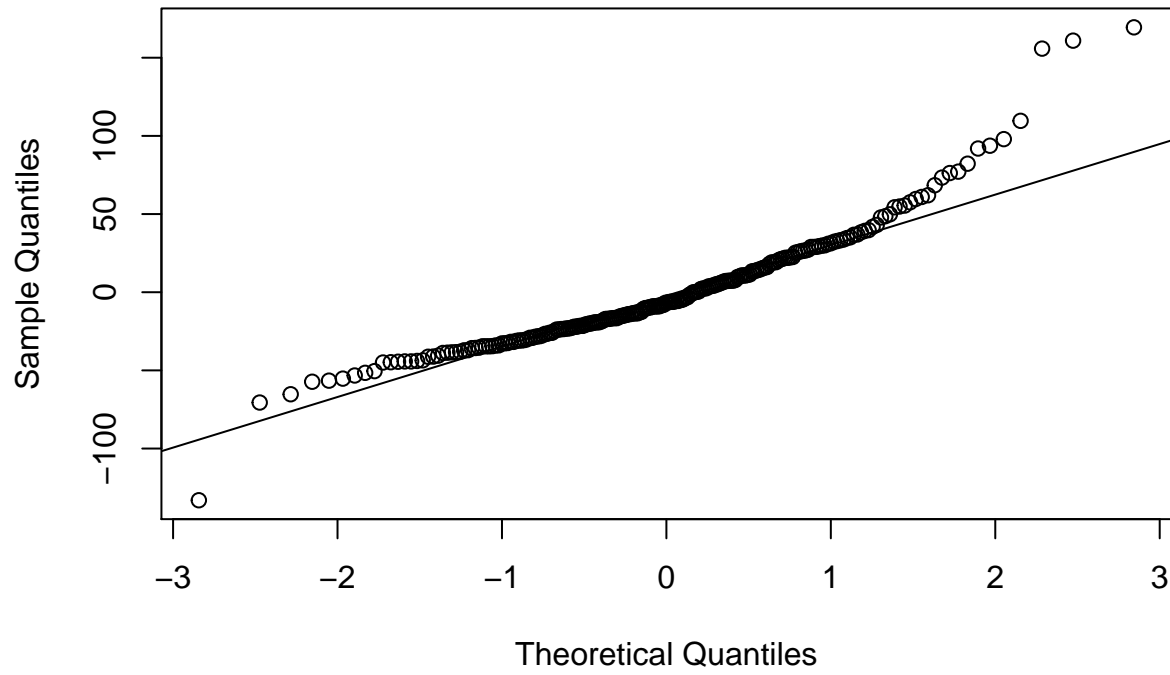
```
#qq-plot of single-level-model
qqnorm(resid(single_level_model))
qqline(resid(single_level_model))
```
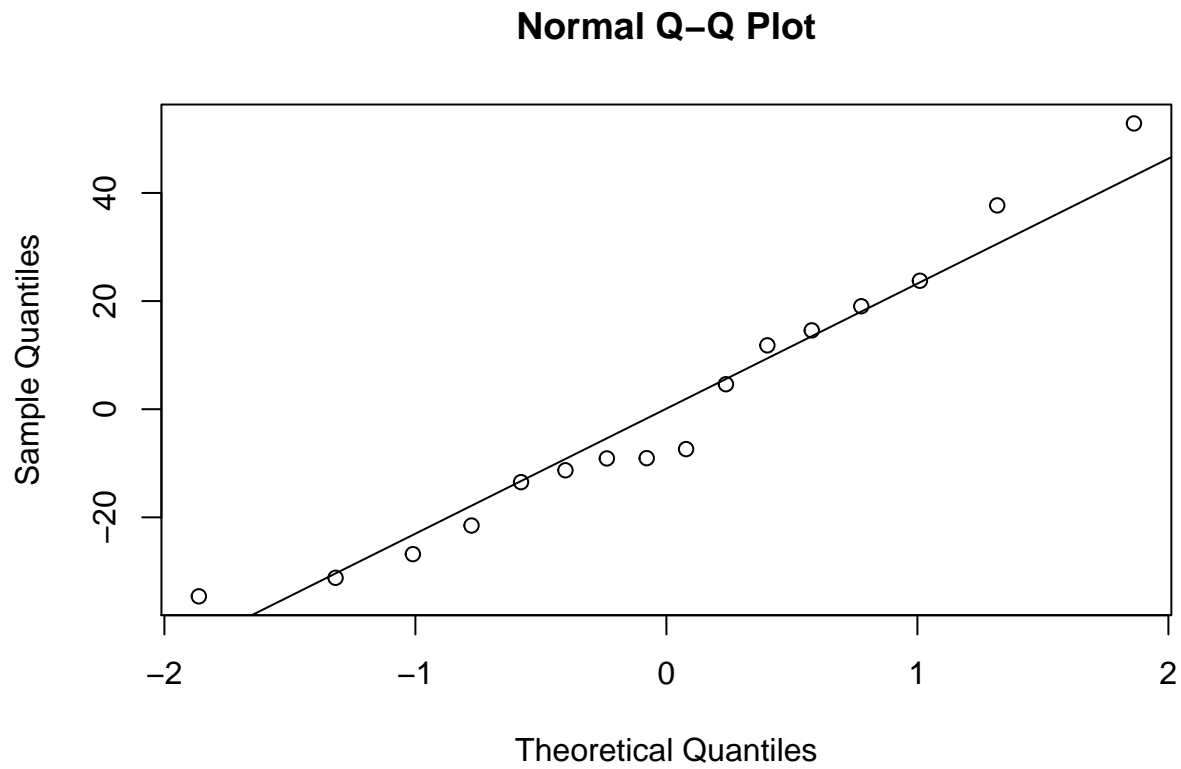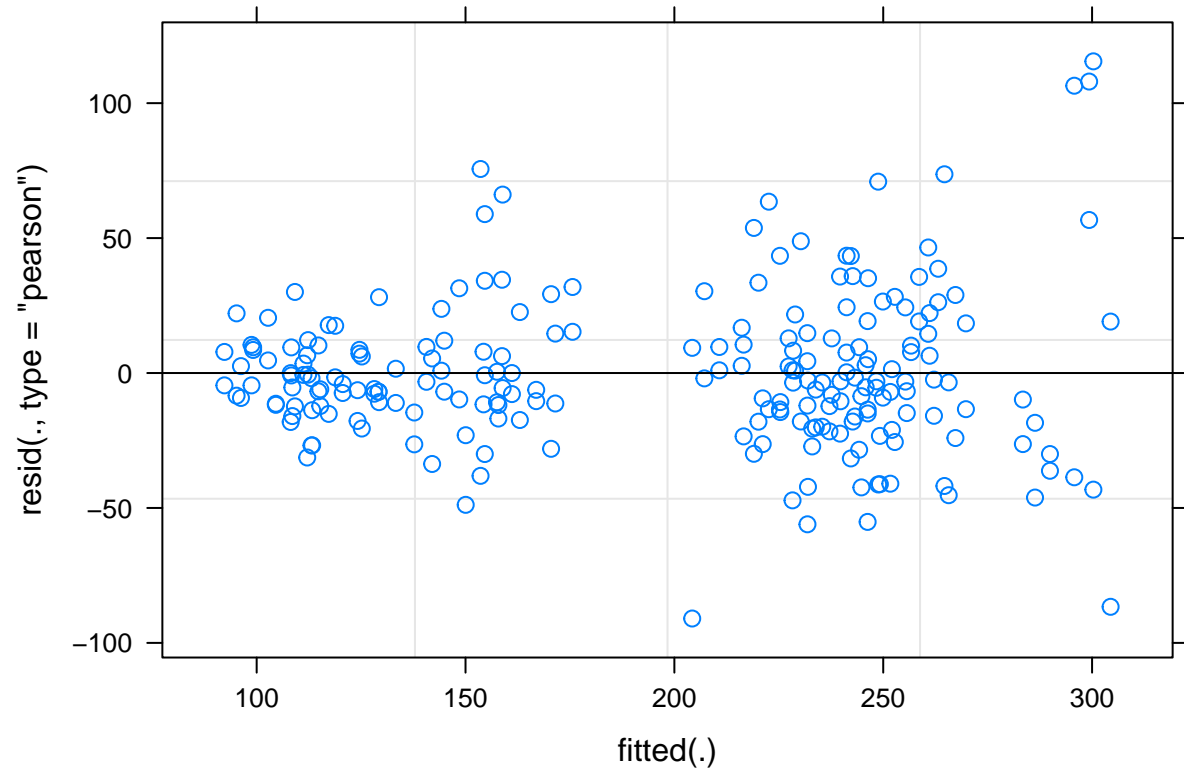
Normal Q–Q Plot


```
#qq-plot of model 1
qqnorm(resid(model1))
qqline(resid(model1))
```
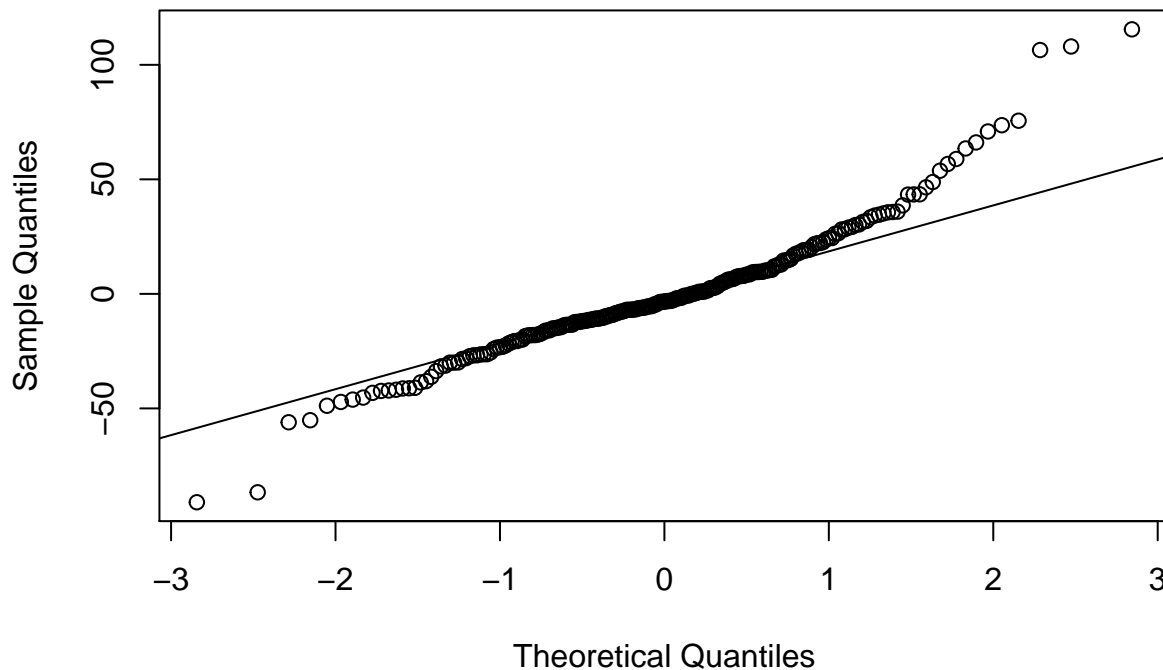
Normal Q–Q Plot


```
#quantile-quantile plot for the residuals of the multilevel model with two intercepts
plot(model4)
```

```
qqnorm(resid(model4))
qqline(resid(model4))
```

Normal Q–Q Plot

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)

###Plotting the two-intercepts model

```
#two-intercept model
model4 <- lmer(f0mn ~ gender + (1|subject) + (1|scenario), data=politeness)
summary(model4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + (1 | subject) + (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2188.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0458 -0.5030 -0.1125  0.4029  3.8668
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 607.40   24.646
```

```
##   scenario (Intercept)  81.87     9.048
##   Residual             892.05    29.867
## Number of obs: 224, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.370      9.288  26.526
## genderM     -115.092     13.055  -8.816
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.615
```

```
#merging predicted values for model 4 into df
politeness$yhat <- predict(model4)

#plot
subplot2 <- ggplot( politeness, aes(x = scenario, y = f0mn, color = attitude)) + geom_point()+geom_poin
```
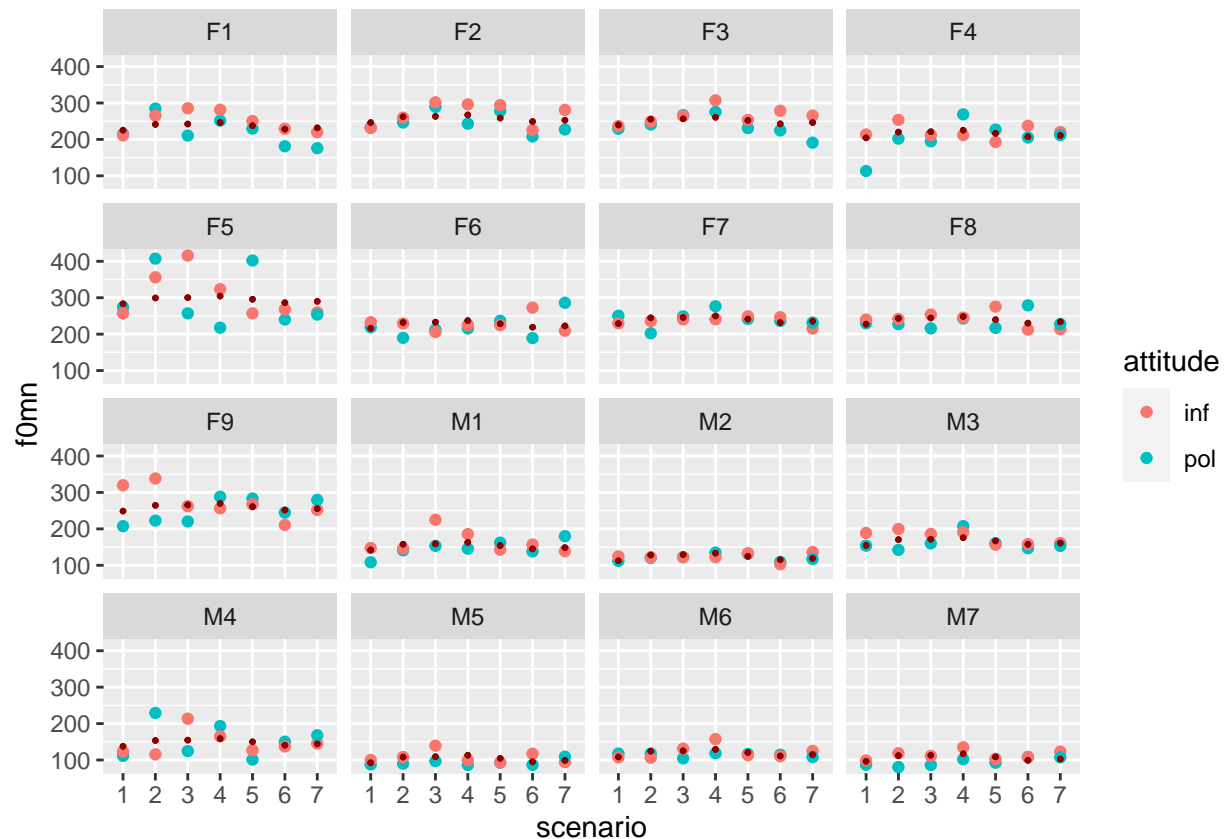

Exercise 3 - now with attitude

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*

    ii. make a separate model that besides the main effects of *attitude* and *gender* also include their interaction

    iii. describe what the interaction term in the model says about Korean men's pitch when they are polite relative to Korean women's pitch when they are polite (you don't have to judge whether it is interesting)

Model with interaction

"Attitudepol" indicates that the pitch generally decreases with -15.646 hz when going from informal to polite, but the interaction effect (gender*attitude) shows that for Males the pitch in decreases with LESS than for Females i.e 5,781 hz less in the polite condition. Thus, Korean men's pitch when they are polite are higher than Korean women's pitch when they are polite, in relation to them selves (as we have modelled indvidual intercepts for each subject). So the change in pitch is lower for Males than for Females in the polite condition.

```
#model 5 with attitude as a main effect besides gender
model5 <- lmer(f0mn ~ gender + attitude + (1|subject) + (1|scenario), data=politeness)
summary(model5)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + attitude + (1 | subject) + (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2173.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8861 -0.5948 -0.0981  0.4015  3.9216
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 610.51   24.709
##  scenario (Intercept)  83.23    9.123
##  Residual             848.56   29.130
## Number of obs: 224, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  252.928      9.500  26.624
## genderM     -115.092     13.055  -8.816
## attitudepol  -13.116      3.893  -3.370
##
## Correlation of Fixed Effects:
##             (Intr) gendrM
## genderM     -0.601
## attitudepol -0.205  0.000
```

```
#model 6 with an interaction between attitude and gender
model6 <- lmer(f0mn ~ gender + attitude+ gender*attitude + (1|subject) + (1|scenario), data=politeness)
summary(model6)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + attitude + gender * attitude + (1 | subject) +
##     (1 | scenario)
##    Data: politeness
```

```
##
## REML criterion at convergence: 2166.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8399 -0.5643 -0.0944  0.4264  3.9610
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 610.37   24.71
##  scenario (Intercept)  83.17    9.12
##  Residual             850.50   29.16
## Number of obs: 224, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        254.193      9.654  26.330
## genderM           -117.983     13.633  -8.654
## attitudepol        -15.646      5.196  -3.011
## genderM:attitudepol   5.781      7.856   0.736
##
## Correlation of Fixed Effects:
##            (Intr) gendrM atttdp
## genderM    -0.618
## attitudepol -0.269  0.191
## gndrM:tttdp  0.178 -0.288 -0.661
```

```
fixef(model6)
```

```
##        (Intercept)             genderM         attitudepol genderM:attitudepol
##          254.19276          -117.98276           -15.64572             5.78136
```

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.

Comparing the three new models

When comparing the three new 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. All of them having unique intercepts for *subject* and *scenario*), the residual variance, residual standard deviation and AIC, looks as followed: Model 4 has a sigma = 29.86727, an AIC = 2198.778, and a R2c = 0.816289. Model 5 has a sigma = 29.13006, an AIC = 2185.125, and a R2c = 0.8253963. Model 6 has a sigma = 29.16333, an AIC = 2180.623, and a R2c = 0.8251343.

Model 5 has the lowest sigma and Model 6 has the lowest AIC, but they all don't vary much in their sigma, AIC and R2c.

```
#model 4 - two-level model that models intercepts for both scenario and subject
model4 <- lmer(f0mn ~ gender + (1|subject) + (1|scenario), data=politeness)
summary(model4)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + (1 | subject) + (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2188.8
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0458 -0.5030 -0.1125  0.4029  3.8668
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 607.40   24.646
##  scenario (Intercept)  81.87    9.048
##  Residual             892.05   29.867
## Number of obs: 224, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.370      9.288  26.526
## genderM     -115.092     13.055  -8.816
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.615
```

```r
sigma(model4)   #29.86727
```

```
## [1] 29.86727
```

```r
#model 5 with attitude as a main effect besides gender
model5 <- lmer(f0mn ~ gender + attitude + (1|subject) + (1|scenario), data=politeness)
summary(model5)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + attitude + (1 | subject) + (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2173.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8861 -0.5948 -0.0981  0.4015  3.9216
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 610.51   24.709
##  scenario (Intercept)  83.23    9.123
##  Residual             848.56   29.130
## Number of obs: 224, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##             Estimate Std. Error t value
```

```
## (Intercept)    252.928       9.500  26.624
## genderM        -115.092      13.055  -8.816
## attitudepol    -13.116        3.893  -3.370
##
## Correlation of Fixed Effects:
##            (Intr) gendrM
## genderM    -0.601
## attitudepol -0.205  0.000
```

```
sigma(model5)  #29.13006
```

```
## [1] 29.13006
```

```
#model 6 with an interaction between attitude and gender
model6 <- lmer(f0mn ~ gender + attitude+ gender*attitude + (1|subject) + (1|scenario), data=politeness)
summary(model6)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + attitude + gender * attitude + (1 | subject) +
##     (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2166.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8399 -0.5643 -0.0944  0.4264  3.9610
##
## Random effects:
##  Groups    Name        Variance Std.Dev.
##  subject   (Intercept) 610.37   24.71
##  scenario  (Intercept)  83.17    9.12
##  Residual              850.50   29.16
## Number of obs: 224, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##                    Estimate Std. Error t value
## (Intercept)         254.193      9.654  26.330
## genderM            -117.983     13.633  -8.654
## attitudepol         -15.646      5.196  -3.011
## genderM:attitudepol   5.781      7.856   0.736
##
## Correlation of Fixed Effects:
##            (Intr) gendrM atttdp
## genderM    -0.618
## attitudepol -0.269  0.191
## gndrM:tttdp  0.178 -0.288 -0.661
```

```
sigma(model6)  #29.16333
```

```
## [1] 29.16333
```

```
#comparing AIC values
AIC(model4, model5, model6)
```

```
##        df      AIC
## model4  5 2198.778
## model5  6 2185.125
## model6  7 2180.623
```

```
#comparing conditional R2 within the second levels
MuMIn::r.squaredGLMM(model4) #R2c = 0.816289
```

```
##              R2m      R2c
## [1,] 0.6743396 0.816289
```

```
MuMIn::r.squaredGLMM(model5) #R2c = 0.8253963
```

```
##              R2m      R2c
## [1,] 0.6826496 0.8253963
```

```
MuMIn::r.squaredGLMM(model6) #R2c = 0.8251343
```

```
##              R2m      R2c
## [1,] 0.6825399 0.8251343
```

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

  ii. what can you conclude about the effect of gender and attitude on pitch (if anything)?

  iii. motivate why you would include separate intercepts for subjects and scenarios (if you think they should be included)

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

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

Choose the best model and report

#####The model Based on the previous analysis of the data and modeling, the best model would be model 6, f0mn ~ gender + attitude+ gender*attitude + (1|subject) + (1|scenario). The variables used in the model are; f0mn, gender, attitude, subject, and scenario. "f0mn" represents the pitch of voice in hertz, "gender" represents which gender each subject identifies as (Female or Male), "attitude" represents two types of conditions for each trial (Informal or Polite), "subject" represents each subject's individual participant ID, and "scenario" represents 7 different trials, each with 2 different conditions, hence 14 rows per subject.

The summary output of model 6 shows how going from female (with an intercept of 254.193 hz) to male, the pitch of voice decreases by 117.983 hz. When going from the attitude "informal" to "polite" the pitch of voice decreases by 15.646 hz. - Females in informal attitude condition has a pitch of 254.193 (intercept) -

Females in polite attitude condition has a pitch of 238.547 i.e. (254.193-15.646). - Males in informal attitude condition has a pitch of 136.21 hz i.e. (254.193-117.983) - Males in polite attitude condition has a pitch of 120.564 hz i.e. (254.193-(-117.983-15.646))

According to the interaction effect, males pitch decreases with LESS than for Females i.e 5,781 hz less in the polite condition. Thus, Korean men's pitch when they are polite are higher than Korean women's pitch when they are polite, in relation to them selves (as we have modelled indvidual intercepts for each subject). So the change in pitch is lower for Males than for Females in the polite condition.

Motivation for including seperate intercepts for subject and scenarios   It makes sense to include separate intercepts for the variables "subject" and "scenario", since we then assume that each subject and scenario has different baselines - a different average effect in pitch per subject and scenario. By making separate intercepts, we account for individual differences across subject and scenario.

Variance components   This model has a sigma = 29.16333, an AIC = 2180.623, and a R2c = 0.8251343. The model has a slightly better AIC value than the other models, but a slightly worse sigma and R2c than model 5.

Q-Q-plot of model   When comparing a Q-Q-plot from model 6 and Q-Q-plot from model 5, the values seem to fit the line slightly better for model 6, but it is still has a right skewed tail.

Conceptually model 6 would make sense, as one could argue that pitch would get affected depending on both gender and attitude (the interaction).

```
#model 6 with an interaction between attitude and gender
model6 <- lmer(f0mn ~ gender + attitude+ gender*attitude + (1|subject) + (1|scenario), data=politeness)
summary(model6)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + attitude + gender * attitude + (1 | subject) +
##     (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2166.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8399 -0.5643 -0.0944  0.4264  3.9610
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 610.37   24.71
##  scenario (Intercept)  83.17    9.12
##  Residual             850.50   29.16
## Number of obs: 224, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##                   Estimate Std. Error t value
## (Intercept)        254.193      9.654  26.330
## genderM           -117.983     13.633  -8.654
## attitudepol        -15.646      5.196  -3.011
## genderM:attitudepol  5.781      7.856   0.736
##
```
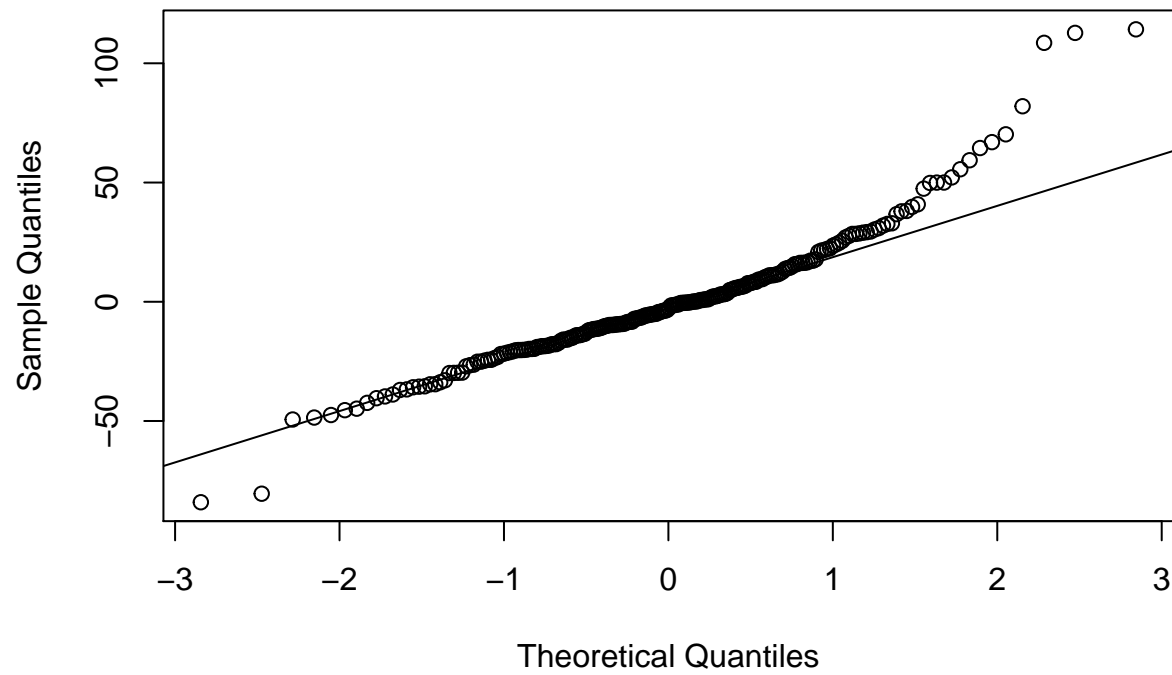
26

```
## Correlation of Fixed Effects:
##            (Intr) gendrM atttdp
## genderM    -0.618
## attitudepol -0.269  0.191
## gndrM:tttdp  0.178 -0.288 -0.661
```

```r
qqnorm(residuals(model6))
qqline(residuals(model6))
```
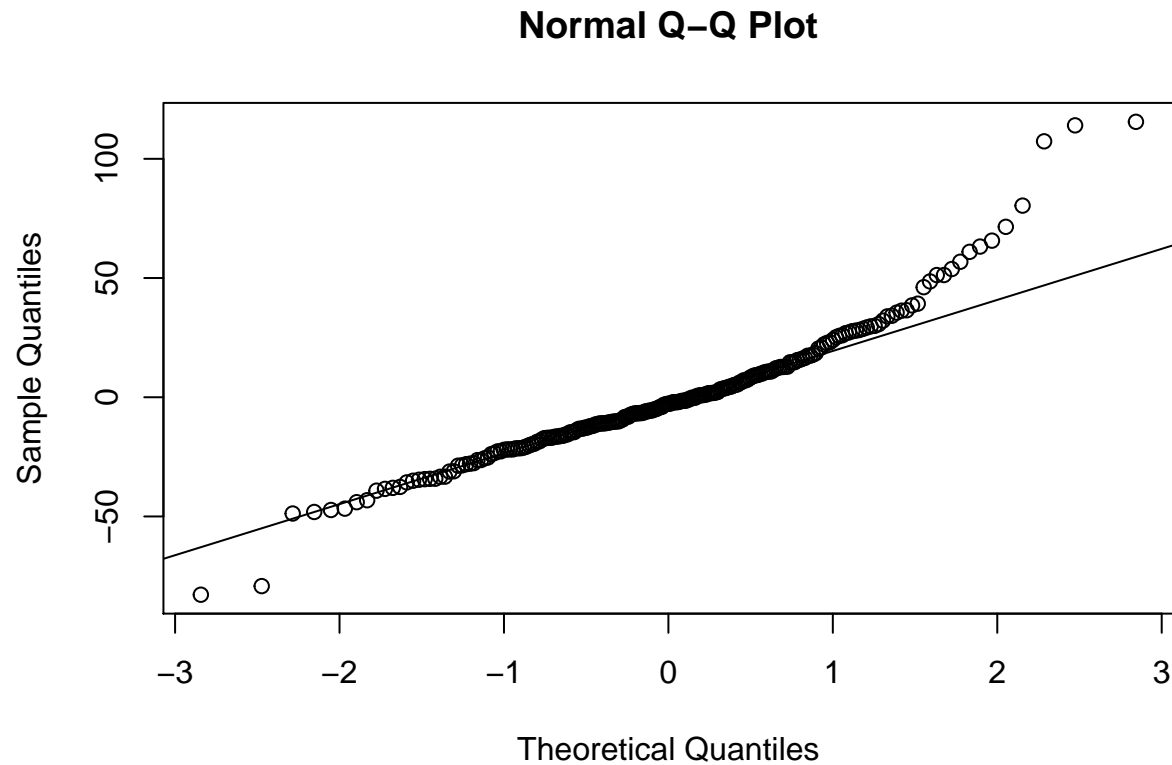
Normal Q–Q Plot


```r
#for comparison
qqnorm(resid(model5))
qqline(resid(model5))
```

Normal Q–Q Plot


```r
qqnorm(resid(model4))
qqline(resid(model4))
```

Normal Q−Q Plot