practical_exercise_2, Methods 3, 2021, autumn semester

Anja, Astrid, Jessica, Juli, Magnus

22 sep 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 = participant Gender = gender Scenario = condition (e.g. "apologising for being late", "asking a professor for an extension on an assignment" etc) Attitude = informal or formal (polite) conditions Total_duration = of sentence/saying, in seconds f0mn = mean of the f0, which is something to do with frequency (pitch??) hiss_count = audible and nasal hissing/air-sucking in between talking #### i. Also consider whether any of the variables in *politeness* should be encoded as factors or have the factor encoding removed. Hint: `?factor`

```
glimpse(politeness)
```

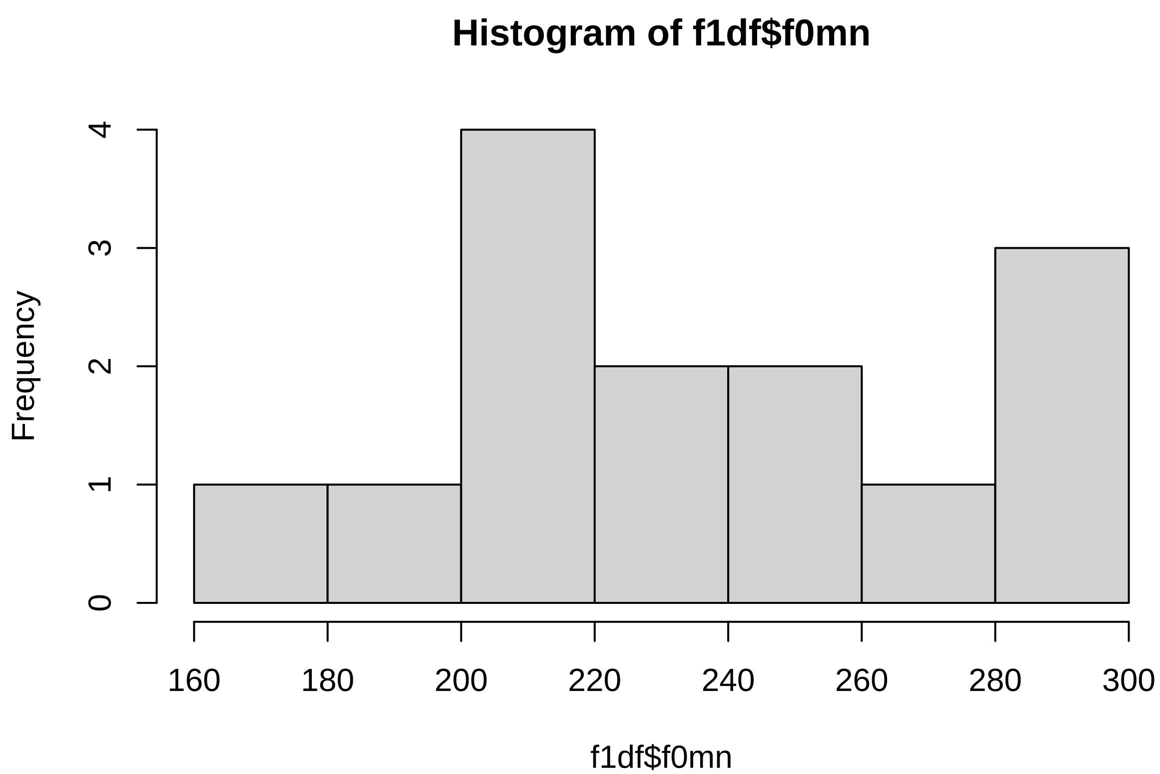
```
## Rows: 224
## Columns: 7
## $ subject        <chr> "F1", "F1", "F1", "F1", "F1", "F1", "F1", "F1", "F1"...
## $ gender         <chr> "F", "F", "F", "F", "F", "F", "F", "F", "F", "F", "F...
## $ scenario       <int> 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7, 7, 1, 1, 2, 2...
```

```
## $ attitude       <chr> "pol", "inf", "pol", "inf", "pol", "inf", "pol", "in...
## $ total_duration <dbl> 18.392, 13.551, 5.217, 4.247, 6.791, 4.126, 6.244, 3...
## $ f0mn           <dbl> 214.6, 210.9, 284.7, 265.6, 210.6, 285.6, 251.5, 281...
## $ hiss_count     <int> 2, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 2, 0, 0, 0, 0, 0...
```

```
politeness$subject <- as_factor(politeness$subject)
politeness$gender <- as_factor(politeness$gender)
politeness$attitude <- as_factor(politeness$attitude)
```

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

```
f1df <- politeness %>%
  filter(subject=="F1")
hist(f1df$f0mn)
```




```
#This is not normal, is that ok?
# as factor
f1df$scenario <- as.integer(f1df$scenario)
class(f1df$scenario)
```

```
## [1] "integer"
```

```
model_1_2_integer <- lm(f0mn ~ scenario, data = f1df)
#as integer
f1df$scenario <- as_factor(f1df$scenario)
class(f1df$scenario)
```

```
## [1] "factor"
```

```r
model_1_2_factor <- lm(f0mn ~ scenario, data = f1df)
```

Should not be an integer that it can do math with, it is a factor and 1 might as well be called "Talking to teacher"

```r

model.matrix(model_1_2_integer)
```

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

```
##    (Intercept) scenario
## 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
```

The matrix shows the scenario as x-values, as a datapoint that has a corresponding y-value(the frequency). The number as the output/value from the scenario

```r
model.matrix(model_1_2_factor)
```

```
##    (Intercept) scenario2 scenario3 scenario4 scenario5 scenario6 scenario7
## 1            1         0         0         0         0         0         0
## 2            1         0         0         0         0         0         0
## 3            1         1         0         0         0         0         0
## 4            1         1         0         0         0         0         0
## 5            1         0         1         0         0         0         0
## 6            1         0         1         0         0         0         0
## 7            1         0         0         1         0         0         0
## 8            1         0         0         1         0         0         0
## 9            1         0         0         0         1         0         0
## 10           1         0         0         0         1         0         0
## 11           1         0         0         0         0         1         0
## 12           1         0         0         0         0         1         0
## 13           1         0         0         0         0         0         1
## 14           1         0         0         0         0         0         1
## attr(,"assign")
## [1] 0 1 1 1 1 1 1
## attr(,"contrasts")
## attr(,"contrasts")$scenario
## [1] "contr.treatment"
```

Here we see true/false, whether the trial is from the scenario or not.

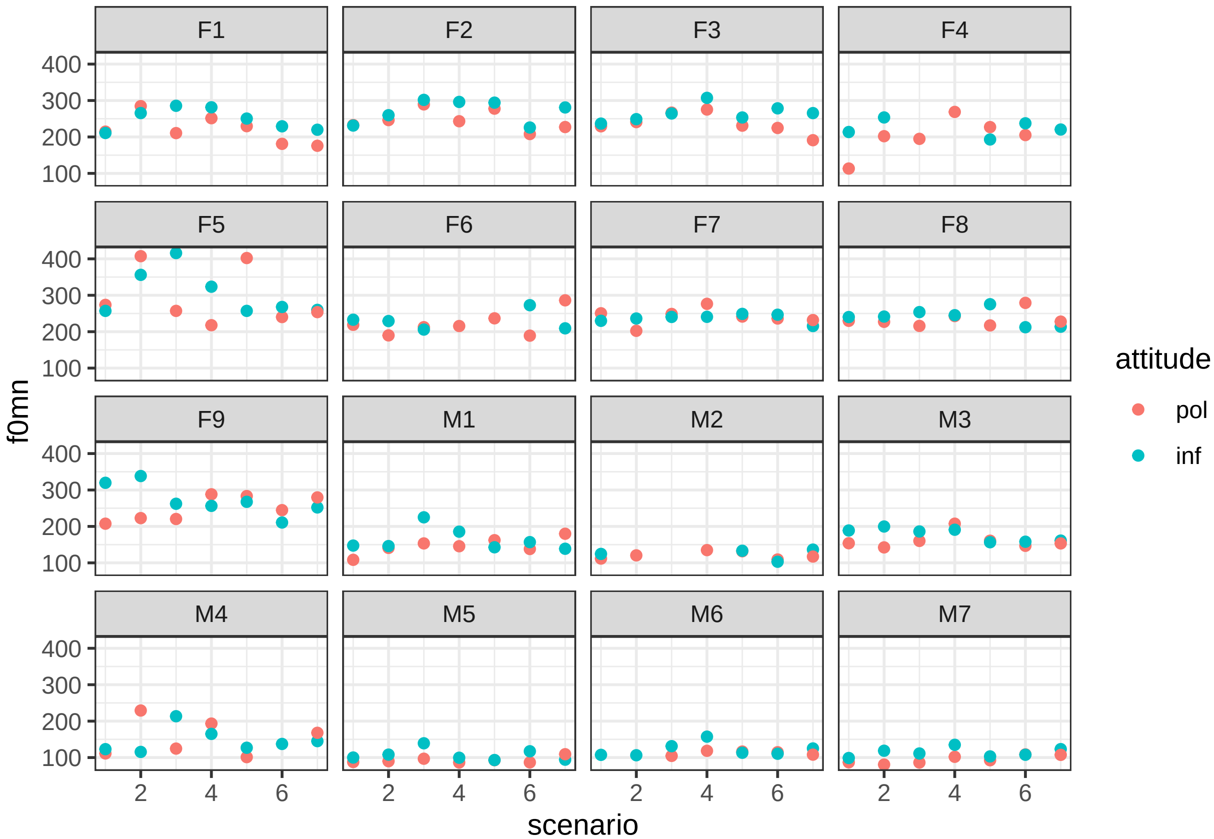
The integer encoding treats the scenarios as 'scores' or data points — which is not the way the scenario number should be treated in this case. The factor encoding treats it like TRUE/FALSE and in this way, we can work with the scenario column as an experiment condition.

ii. Which coding of *scenario*, as a factor or not, is more fitting? Factor is more fitting, 1 is not a value but a name for a specific senario. Coding as an integer makes R interpret the data incorrectly.

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*

```
ggplot(data = politeness, aes(x = scenario, y = f0mn, color = attitude)) +
  geom_point() +
  facet_wrap(~ subject) +
  theme_bw()
```

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



i. Describe the differences between subjects

- Males have lower frequency than females.
- Some people are not as affected by attitude as others.
- Looking at subject F5 as an example, the scenario seem to have a bigger impact on voice than attitude.
- For the hypotheses that Koreans lower their voices in formal conditions to hold, the red dots should be above the blue, which is hard to conclude from just the graphs, but might be true.

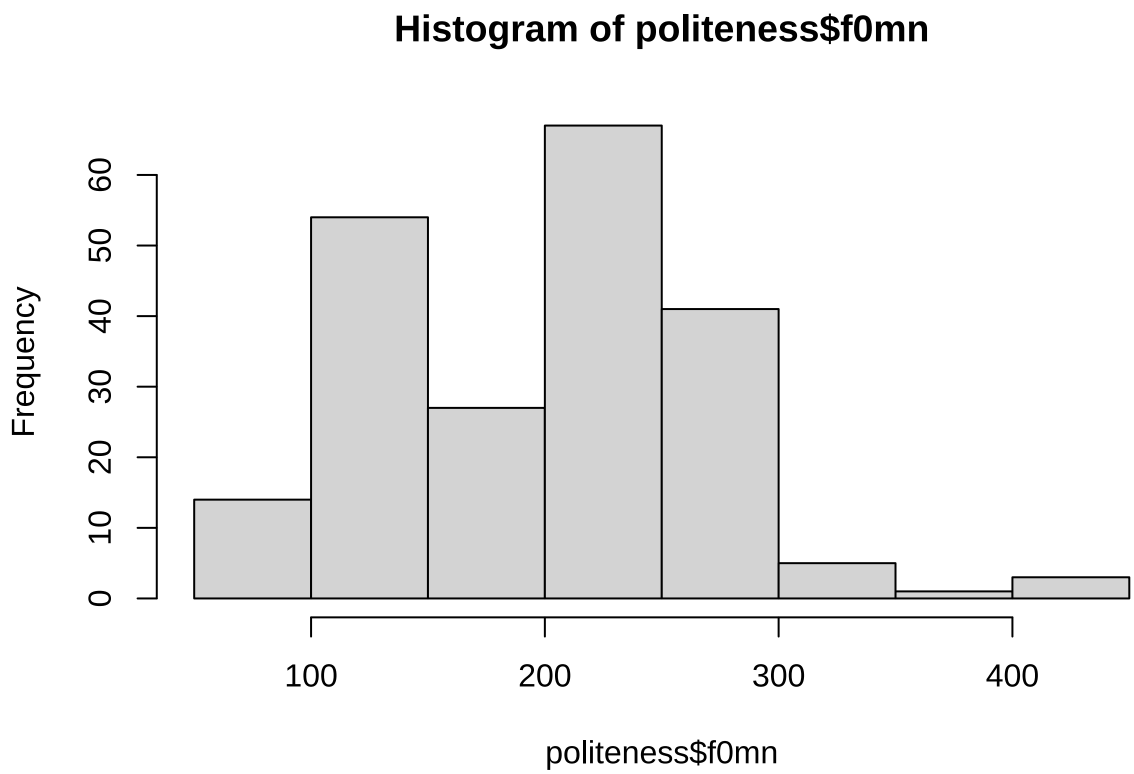
Exercise 2 - comparison of models

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

```
mixed.model <- lmer(formula=..., data=...)
example.formula <- formula(dep.variable ~ first.level.variable + (1 | second.level.variable))
```

1) Build four models and do some comparisons

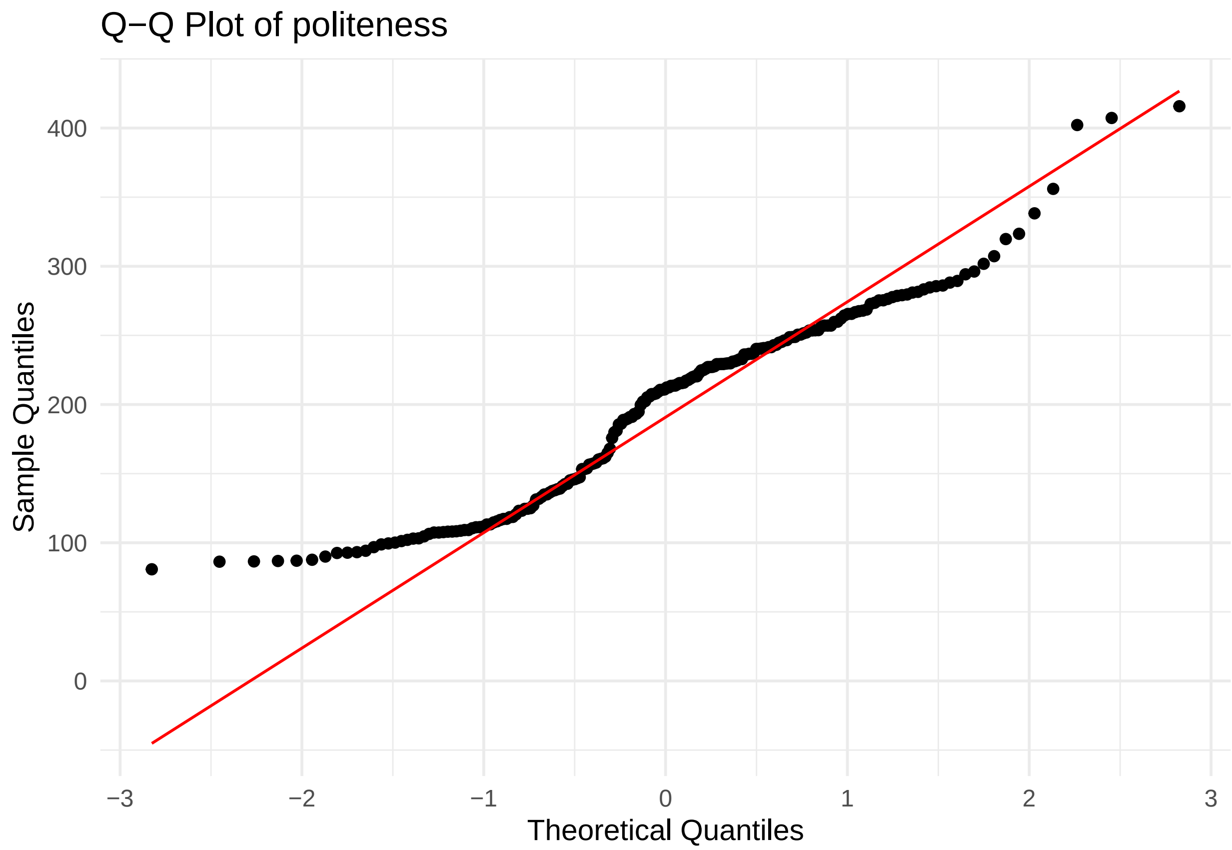
```
hist(politeness$f0mn)
```




```
ggplot(politeness, aes(sample=f0mn)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(x = "Theoretical Quantiles", y = "Sample Quantiles") +
  ggtitle("Q-Q Plot of politeness") +
  theme_minimal()
```

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

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



We wondered if the data was normal enough to model. . . ? We used the hist() and plotted it to visualize it.

```
model_2_1 <- lm(f0mn ~ gender, data = politeness)
```

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

```
model_2_2 <- lmer(f0mn ~ gender + (1 | scenario), data = politeness)
```

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

```
model_2_3 <- lmer(f0mn ~ gender + (1 | subject), data = politeness)
```

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

```
model_2_4 <- lmer(f0mn ~ gender + (1|subject) + (1|scenario), data = politeness)
```

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

```
# HOw to calculate the residual standard deviation of each model
sigma(model_2_1)
```

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

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

```
# Combining the residual standard deviation's into one table
SD_comparison <- cbind(sigma(model_2_1), sigma(model_2_2), sigma(model_2_3), sigma(model_2_4))
SD_comparison
```

```
##          [,1]   [,2]     [,3]     [,4]
## [1,] 39.46268 38.448 32.04287 30.65803
```

```
# Calculating the AIC for each model
AIC(logLik(model_2_1))
```

```
## [1] 2163.971
```

```
# Combinging the AIC's of all the models into one table
AIC_comparison <- cbind(AIC(logLik(model_2_1)), AIC(logLik(model_2_2)), AIC(logLik(model_2_3)), AIC(logL
AIC_comparison
```

```
##          [,1]     [,2]     [,3]     [,4]
## [1,] 2163.971 2152.314 2099.626 2092.482
```

From the tables we created, we can see that our model_2_4 as the lowest residual standard deviation and AIC.

```

r.squaredGLMM(model_2_1)
```

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

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

```
##            R2m       R2c
## [1,] 0.6795237 0.6795237
```

```
r_squared_comparison <- cbind(r.squaredGLMM(model_2_1), r.squaredGLMM(model_2_2), r.squaredGLMM(model_2_
r_squared_comparison
```

```
##            R2m       R2c       R2m       R2c       R2m       R2c       R2m
## [1,] 0.6795237 0.6795237 0.6779555 0.6967788 0.6681651 0.7899229 0.6677206
##            R2c
## [1,] 0.8077964
```

When looking at the R^2, we see again that model_2_4 has the highest value and exlains the most variance.

2) Why is our single-level model bad?

```

simple_df <- politeness %>%
  filter(!is.na(f0mn)) %>%
  group_by(subject, gender) %>%
  summarise(mean_f0mn = mean(f0mn))
```

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 (ignoring) *attitude* and_scenario_

```
## `summarise()` has grouped output by 'subject'. You can override using the `.groups` argument.
```

```r
simple_df
```

```
## # A tibble: 16 x 3
## # Groups:   subject [16]
##    subject gender mean_f0mn
##    <fct>   <fct>      <dbl>
##  1 F1      F           235.
##  2 F2      F           258.
##  3 F3      F           251.
##  4 F4      F           212.
##  5 F5      F           299.
##  6 F6      F           225.
##  7 F7      F           239.
##  8 F8      F           237.
##  9 F9      F           261.
## 10 M1      M           155.
## 11 M2      M           122.
## 12 M3      M           169.
## 13 M4      M           150.
## 14 M5      M           100.
## 15 M6      M           118.
## 16 M7      M           104.
```

```r
# There is so few N/A that we can kill them off in good faith
```

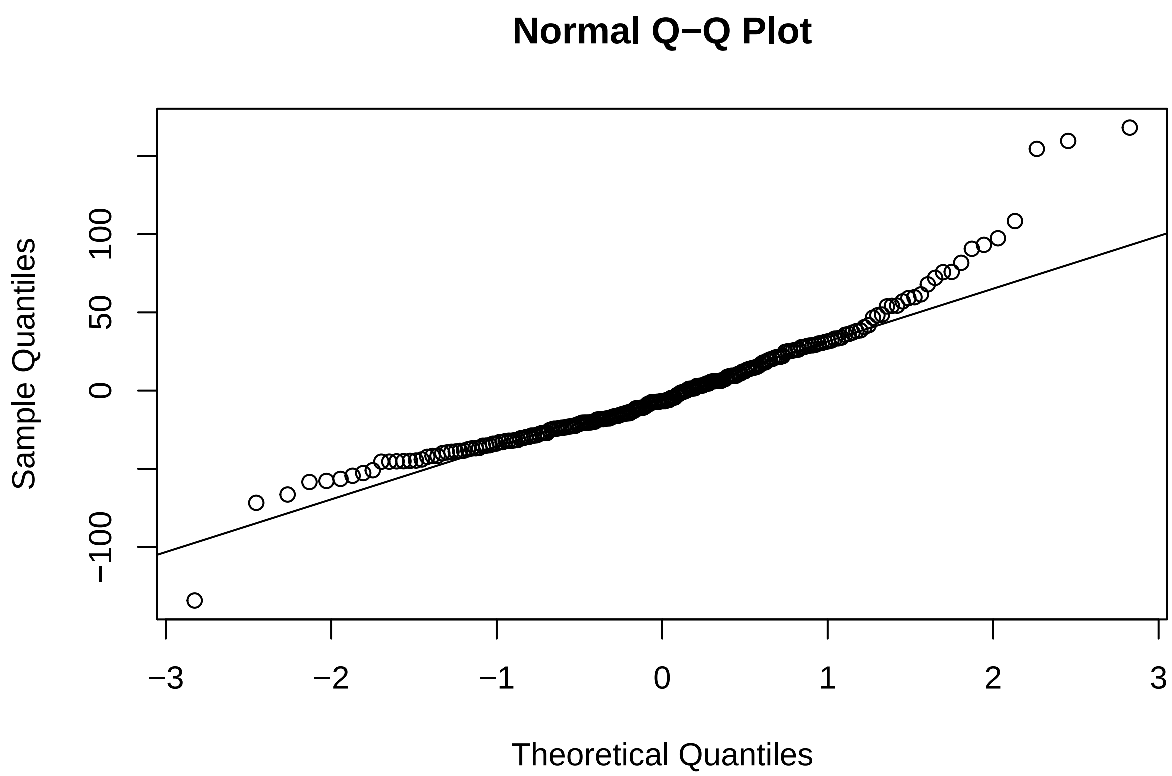
```r
model_2_5 <- lm(mean_f0mn ~ gender, data = simple_df)
model_2_5
```

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

```
##
## Call:
## lm(formula = mean_f0mn ~ gender, data = simple_df)
##
## Coefficients:
## (Intercept)      genderM
##       246.4       -115.1
```

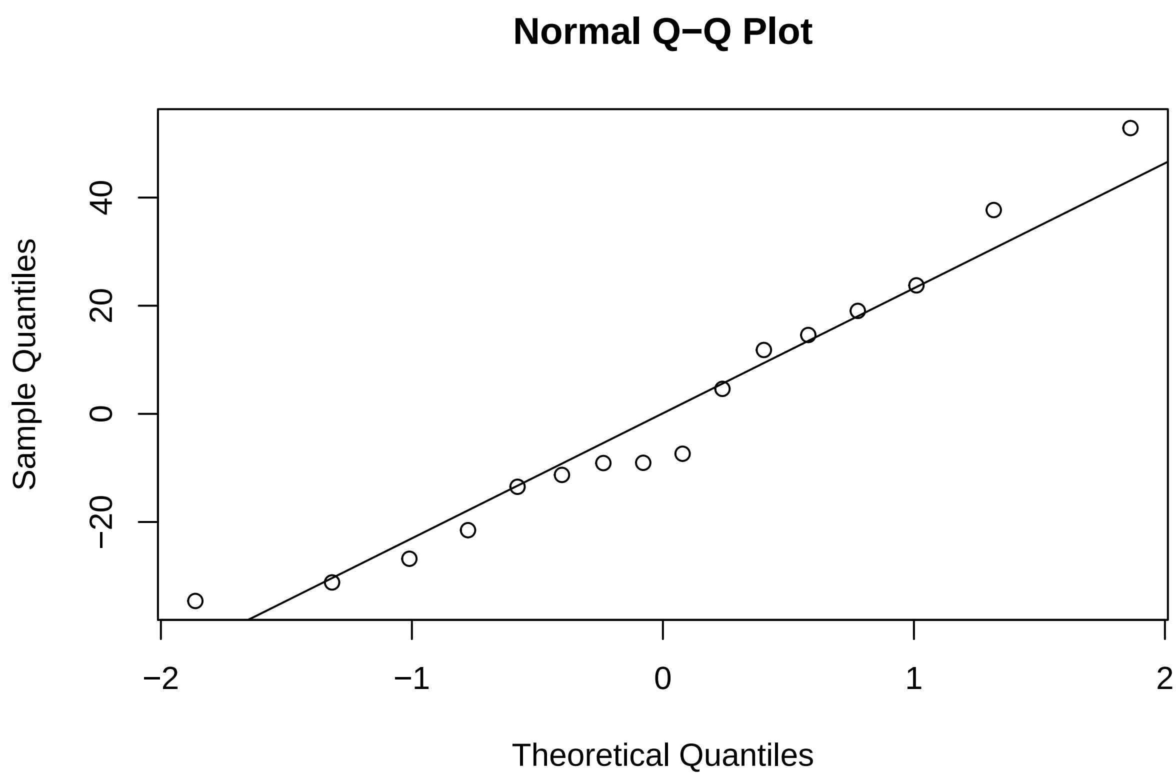
```r
qqnorm(residuals(model_2_1))
qqline(residuals(model_2_1))
```

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 bet-



ter?)

```
qqnorm(residuals(model_2_5))
qqline(residuals(model_2_5))
```

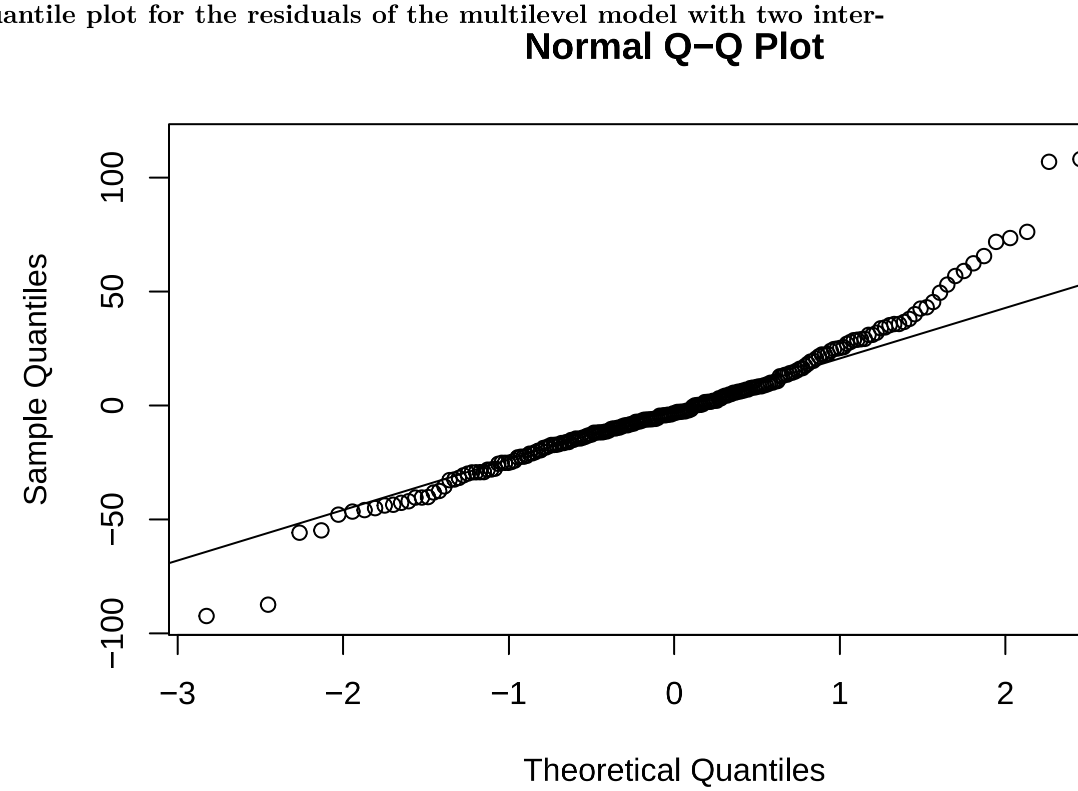



```
#The model from 1i fits better, cause so many dots are on the line. ?
# Comparison between m1 and the new model from the new data frame: Very difficult to determine visually
```

```
# Buuut perhaps m1 is a little bit better, because it has sooo many points right on the line
```

```
qqnorm(residuals(model_2_4))
qqline(residuals(model_2_4))
```

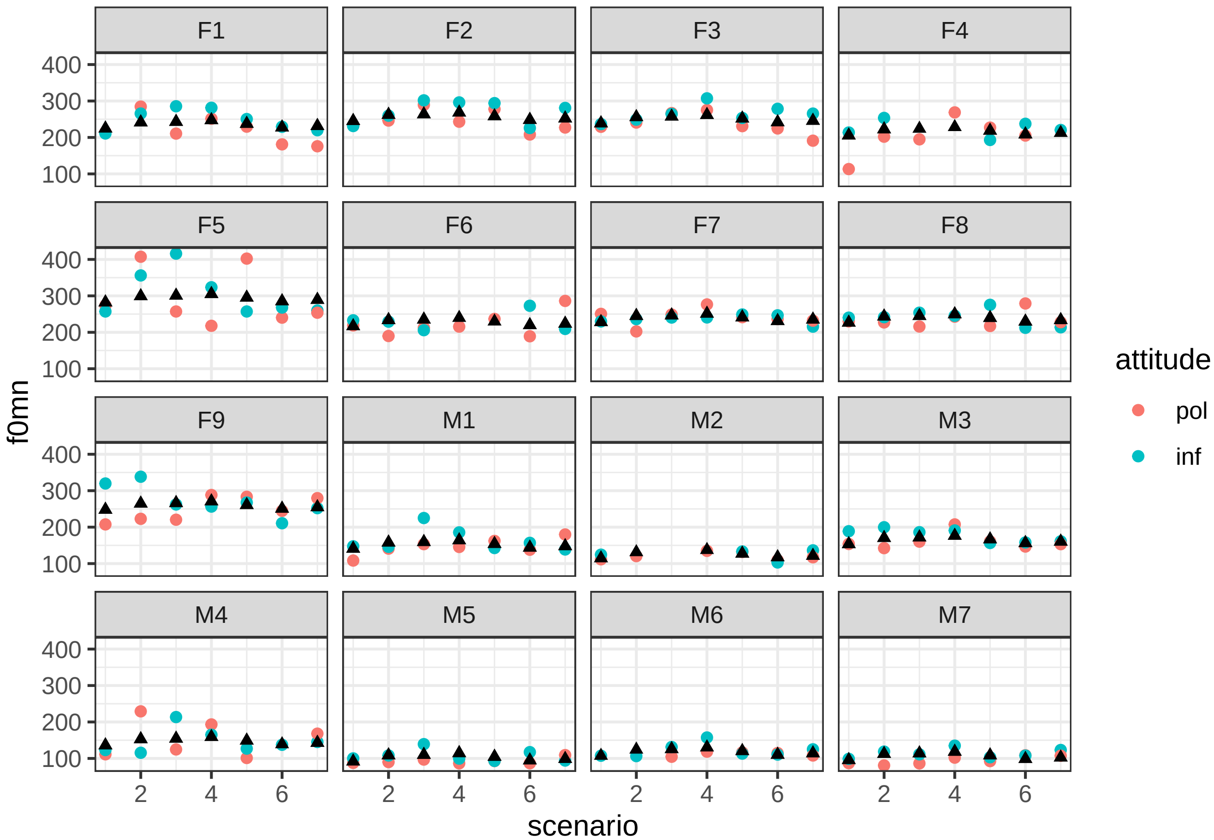
iv. Also make a quantile-quantile plot for the residuals of the multilevel model with two intercepts. Does it look alright?



3) Plotting the two-intercepts model

```
fitted <- fitted(model_2_4)
politeness_na_removed <- politeness %>%
  na.omit()
politeness_na_removed$fitted_f0mn <- fitted
ggplot(data = politeness_na_removed, aes(x = scenario, y = f0mn, color = attitude)) +
  geom_point() +
  geom_point(aes(scenario, fitted_f0mn), color = "black", shape = 17)+
  facet_wrap(~ subject) +
  theme_bw()
```

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)



Exercise 3 - now with attitude

1) Carry on with the model with the two unique intercepts fitted (*scenario* and *subject*).

```
#Adding attitude as a main effect
model_3_1 <- lmer(f0mn ~ gender + attitude + (1 | scenario) + (1 | subject), data = politeness)
summary(model_3_1)
```

i. now build a model that has *attitude* as a main effect besides *gender*

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + attitude + (1 | scenario) + (1 | subject)
##    Data: politeness
##
## REML criterion at convergence: 2065.1
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8511 -0.6081 -0.0602  0.4329  3.8745
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 585.6    24.20
##  scenario (Intercept) 106.7    10.33
##  Residual             882.7    29.71
## Number of obs: 212, groups:  subject, 16; scenario, 7
##
```

```
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  239.579      9.571  25.031
## genderM     -115.437     12.881  -8.962
## attitudeinf   14.819      4.096   3.618
##
## Correlation of Fixed Effects:
##             (Intr) gendrM
## genderM     -0.586
## attitudeinf -0.208 -0.006
```

```
#Adding the interaction of gender and attitude
model_3_2 <- lmer(f0mn ~ gender * attitude + (1 | scenario) + (1 | subject), data = politeness)
summary(model_3_2)
```

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender * attitude + (1 | scenario) + (1 | subject)
##    Data: politeness
##
## REML criterion at convergence: 2058.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8120 -0.5884 -0.0645  0.4014  3.9100
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 584.4    24.17
##  scenario (Intercept) 106.4    10.32
##  Residual             885.5    29.76
## Number of obs: 212, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)          238.426      9.718  24.535
## genderM             -112.687     13.511  -8.341
## attitudeinf           17.192      5.423   3.170
## genderM:attitudeinf   -5.544      8.284  -0.669
##
## Correlation of Fixed Effects:
##             (Intr) gendrM atttdn
## genderM     -0.604
## attitudeinf -0.271  0.195
## gndrM:tttdn  0.177 -0.304 -0.654
```

```
# Men have a bigger gap in pitch when being informal compared to polite (seen in difference between slo
```

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)

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.

```
#Is R^2 the residual variance?
r.squaredGLMM(model_2_4)
```

```
##            R2m       R2c
## [1,] 0.6677206 0.8077964
```

```
r.squaredGLMM(model_3_1)
```

```
##            R2m       R2c
## [1,] 0.6782542 0.8196777
```

```
r.squaredGLMM(model_3_2)
```

```
##           R2m       R2c
## [1,] 0.678249 0.8192531
```

```
#Residual standard deviation
sigma(model_2_4)
```

```
## [1] 30.65803
```

```
sigma(model_3_1)
```

```
## [1] 29.71087
```

```
sigma(model_3_2)
```

```
## [1] 29.75684
```

```
#AIC
AIC1 <- AIC(logLik(model_2_4))
AIC2 <- AIC(logLik(model_3_1))
AIC3 <- AIC(logLik(model_3_2))
AIC1
```

```
## [1] 2092.482
```

```
AIC2
```

```
## [1] 2077.131
```

```
AIC3
```

```
## [1] 2072.618
```

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:

```
#Data for this report is taken from Winter & Grawunder (2012)'s research article 'The phonetic profile
#Participants are given numbers and their gender is recorded (Subject and gender variables). The study
```

i. describe what the dataset consists of

```
#Males have lower pitch than females, being polite lowers pitch
```

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

```
# separate intercepts for subjects are needed due to individual differences in baseline pitch
# separate intercepts for scenarios are needed as we can't directly compare across all polite condition
```

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

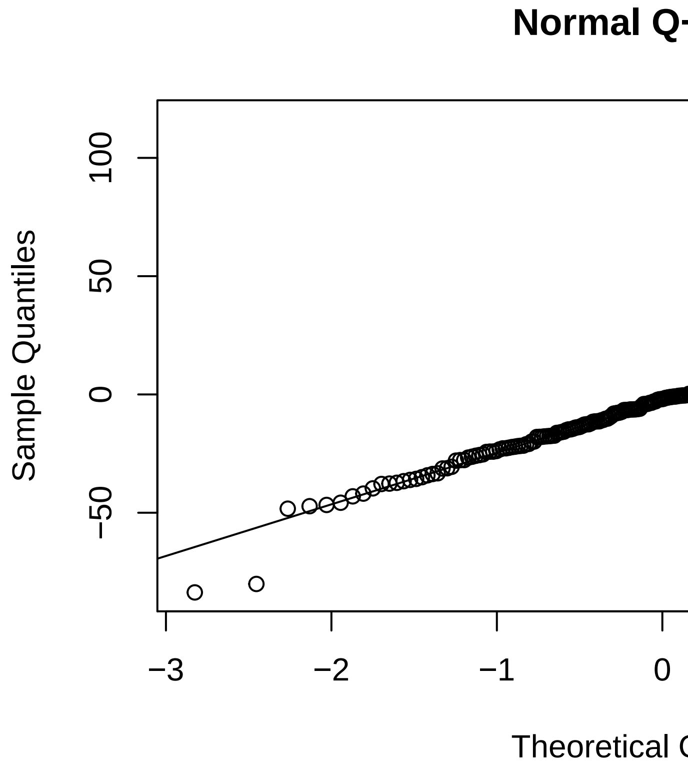
```
summary(model_3_2)
```

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender * attitude + (1 | scenario) + (1 | subject)
##    Data: politeness
##
## REML criterion at convergence: 2058.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.8120 -0.5884 -0.0645  0.4014  3.9100
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 584.4    24.17
##  scenario (Intercept) 106.4    10.32
##  Residual             885.5    29.76
## Number of obs: 212, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##                      Estimate Std. Error t value
## (Intercept)           238.426      9.718  24.535
## genderM              -112.687     13.511  -8.341
## attitudeinf            17.192      5.423   3.170
## genderM:attitudeinf    -5.544      8.284  -0.669
##
## Correlation of Fixed Effects:
##             (Intr) gendrM atttdn
## genderM     -0.604
## attitudeinf -0.271  0.195
## gndrM:tttdn  0.177 -0.304 -0.654
```

```
# Scaled residuals:
#     Min      1Q  Median      3Q     Max
# -2.8120 -0.5884 -0.0645  0.4014  3.9100
# Random effects:
#  Groups   Name        Variance Std.Dev.
#  subject  (Intercept) 584.4    24.17
#  scenario (Intercept) 106.4    10.32
#  Residual             885.5    29.76
# Number of obs: 212, groups:  subject, 16; scenario, 7
```

```
qqnorm(residuals(model_3_2))
qqline(residuals(model_3_2))
```



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