Portfolio Exercise 1, Methods 3, 2021, autumn semester

Alina Kereszt

13.12.2021

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

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

Winter and Grawunder (2012) studied the phonetic profile of Korean speakers in formal and informal scenarios.

- subject: The ID of the experimental participant. Factor.
- gender: Participant's gender. Factor.
- scenario: Experimental condition in the within-subject experiment. Factor.
- attitude: Is the scenario formal or informal. Factor.
- f0mn: Mean frequency of participants' voice pitch during the scenario (higher value indicates higher pitch). Numeric.
- total_duration: The length of the interaction within the scenario. Numeric.
- hiss_count: Number of hissing intakes of breath during the scenario. Numeric.

```
# Remove NAs for cleanliness and because some functions cannot handle them
politeness <- politeness %>%
  na.exclude(f0mn)
# While in some cases it may be appropriate to replace NAs with the mean value of the variabl
e, here it is NOT - the data points are not independent of each other!!!

# Ensure correct data format
politeness$subject <- as.factor(politeness$subject)
politeness$gender <- as.factor(politeness$gender)
politeness$scenario <- as.factor(politeness$scenario)
politeness$attitude <- as.factor(politeness$attitude)
politeness$total_duration <- as.numeric(politeness$total_duration)
politeness$f0mn <- as.numeric(politeness$f0mn)
politeness$hiss_count <- as.numeric(politeness$hiss_count)
```

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.

```
F1 <- politeness %>%
  filter(subject == "F1") # filter for subject _F1_

class(F1$scenario) #check data format of _scenario_ - "factor"
```

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

```
scenario_int <- as.integer(F1$scenario) #copy _scenario_ as integer

model_fac <- lm(f0mn ~ scenario, data = F1) #model where _scenario_ is factor
model_int <- lm(f0mn ~ scenario_int, data = F1) #model where _scenario_ is integer

model_fac$coefficients
```

```
## (Intercept)   scenario2   scenario3   scenario4   scenario5   scenario6
##      212.75       62.40       35.35       53.75       27.30       -7.55
##   scenario7
##      -14.95
```

```
model_int$coefficients
```

```
##  (Intercept) scenario_int
##   262.621429    -6.885714
```

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.

```
X_fac <- model.matrix(model_fac, data = F1)
X_int <- model.matrix(model_int, data = F1)

X_fac
```

```
##    (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"
```

```
X_int
```

```
##   (Intercept) scenario_int
## 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
```

When *scenario* is taken to be a factor, we suppose it to be categorical - the scenarios are discrete entities that don't lie on the same continuum. Therefore, a separate slope is calculated for each level.

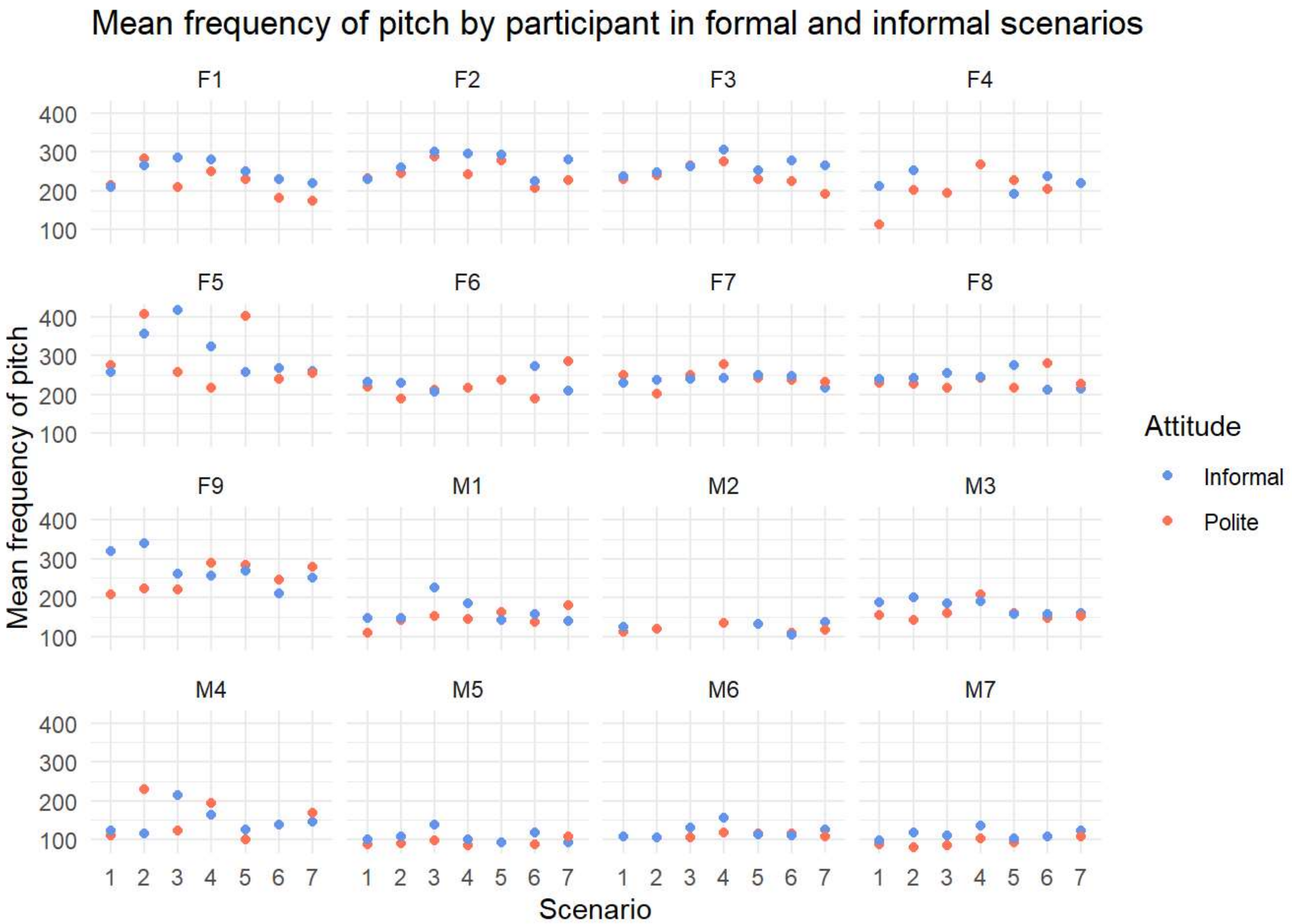
When *scenario* is taken to be an integer, we suppose it to be a continuous variable - the scenarios represent incrementally different values along the same continuum. Therefore, only a single slope is calculated to summarize the values *scenario* takes on.

ii. Which coding of *scenario*, as a factor or not, is more fitting?

As the scenarios are relatively independent of each other and likely create 'sub-groups' in the data, treating them as factors seems most reasonable.

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 color-coded according to *attitude*.

```
ggplot(data = politeness, aes(x = scenario, y = f0mn)) +
  geom_point(aes(colour = attitude)) +
  facet_wrap(~subject) +
  guides(color = guide_legend("Attitude")) +
  scale_color_manual(labels = c("Informal", "Polite"),
                     values = c("cornflowerblue", "coral1")) +
  labs(title = "Mean frequency of pitch by participant in formal and informal scenarios",
       x = "Scenario",
       y = "Mean frequency of pitch") +
  theme_minimal()
```



i. Describe the differences between subjects

Based on a visual inspection, the plots appear to indicate that individuals have a higher mean frequency of pitch in informal scenarios compared to polite ones, and that females in general have a higher mean frequency of pitch than males do.

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*;

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

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

```
m2 <- lme4::lmer(f0mn ~ gender + (1 | scenario), data = politeness)
```

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

```
m3 <- lme4::lmer(f0mn ~ gender + (1 | subject), data = politeness)
```

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

```
m4 <- lme4::lmer(f0mn ~ gender + (1 | scenario) + (1 | subject), data = politeness)
```

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

```
# Residual standard deviation of the models:
c(sigma(m1),
  sigma(m2),
  sigma(m3),
  sigma(m4)
)
```

```
## [1] 39.46268 38.44800 32.04287 30.65803
```

```
# AIC of the models:
AIC(m1, m2, m3, m4)
```

```
##    df      AIC
## m1  3 2163.971
## m2  4 2152.314
## m3  4 2099.626
## m4  5 2092.482
```

Model *m4* has both the lowest residual standard deviation (30.65803) and lowest AIC score (2092.482), although not much lower than the residual standard deviation and AIC score of model *m3*.

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

```
# The amount of variance explained by the model is represented by R^2, the coefficient of det
ermination.
rsq(m2)
```

```
## $model
## [1] 0.6989103
##
## $fixed
## [1] 0.6805506
##
## $random
## [1] 0.01835976
```

```
rsq(m3)
```

```
## $model
## [1] 0.7938981
##
## $fixed
## [1] 0.6804095
##
## $random
## [1] 0.1134886
```

Both models *m2* and *m3* have the same fixed effects (which thereby explain the same amount of variance). *Subject* (the random intercept in model *m3*) appears to explain almost 10 times as much variance as *scenario* (random intercept in model *m2*).

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*.

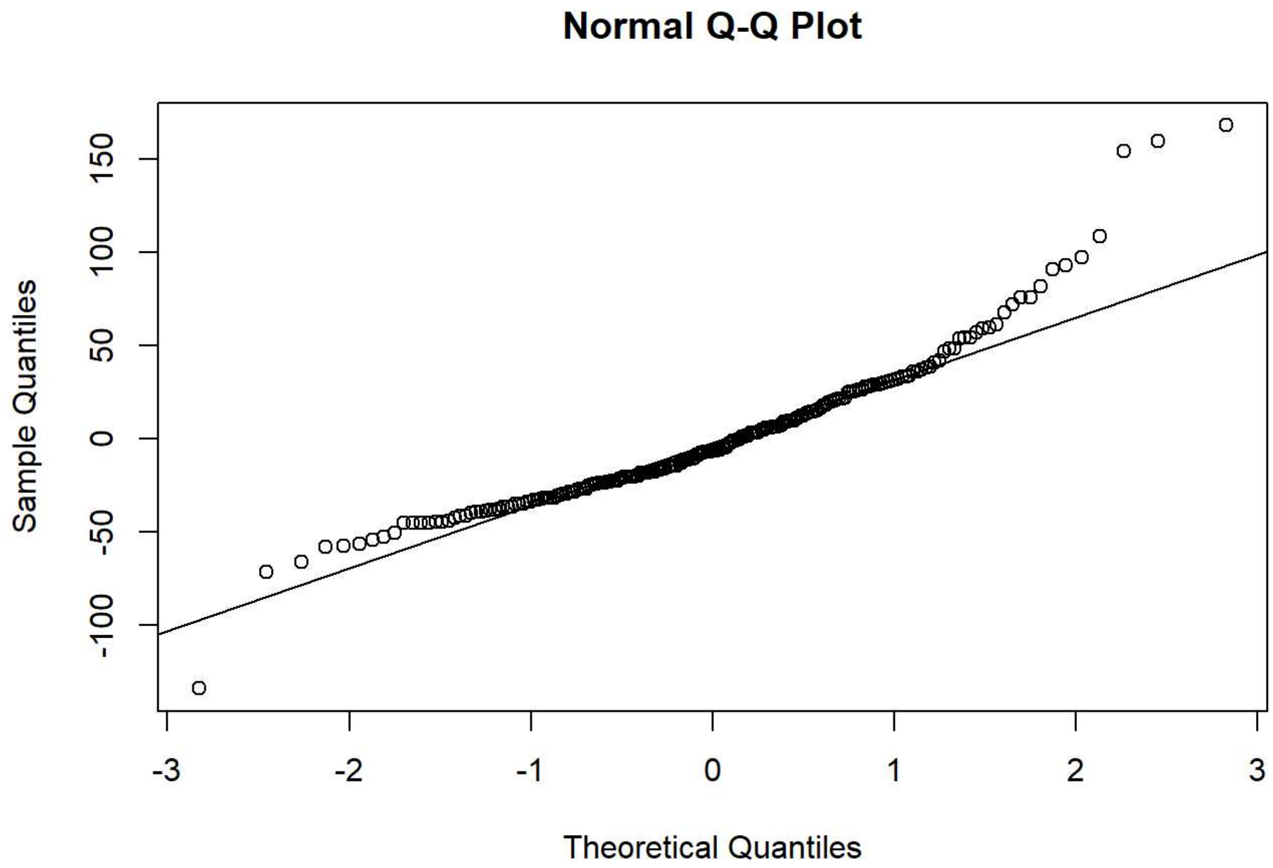
```
baddata <- politeness %>%
  group_by(subject, gender) %>%
  summarize(f0mn = mean(f0mn))
```

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

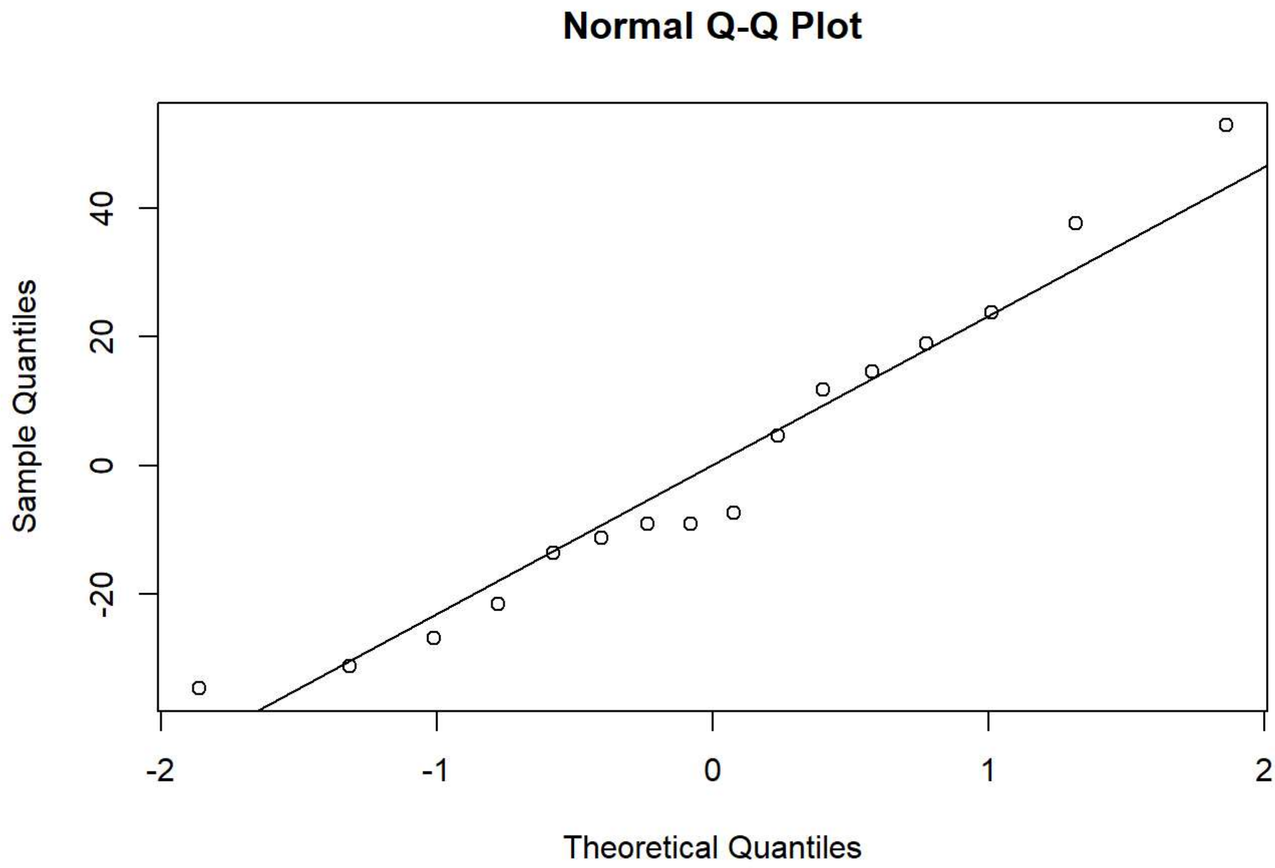
```
badmodel <- lm(f0mn ~ gender, data = baddata)
```

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) fulfill the assumptions of the General Linear Model better?)

```
qqnorm(residuals(m1))
qqline(residuals(m1))
```



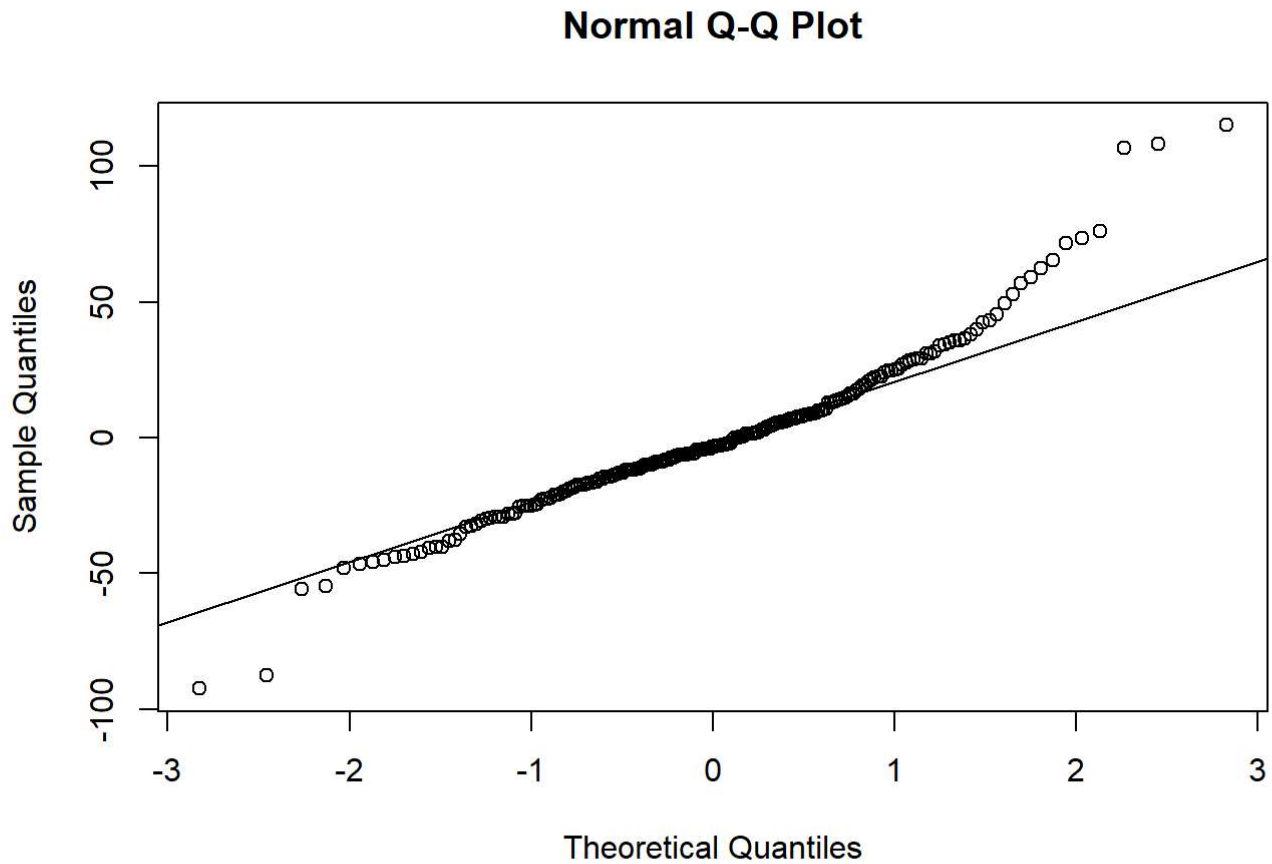

```
qqnorm(residuals(badmodel))
qqline(residuals(badmodel))
```



The residuals for both appear approximately normally distributed, however, there is skewness present in the residuals for model *m1* that is not present in the residuals for model *badmodel* (perhaps because by taking the mean of the *f0mn* values, less outliers or extreme values are taken into consideration).

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

```
qqnorm(residuals(m4))
qqline(residuals(m4))
```

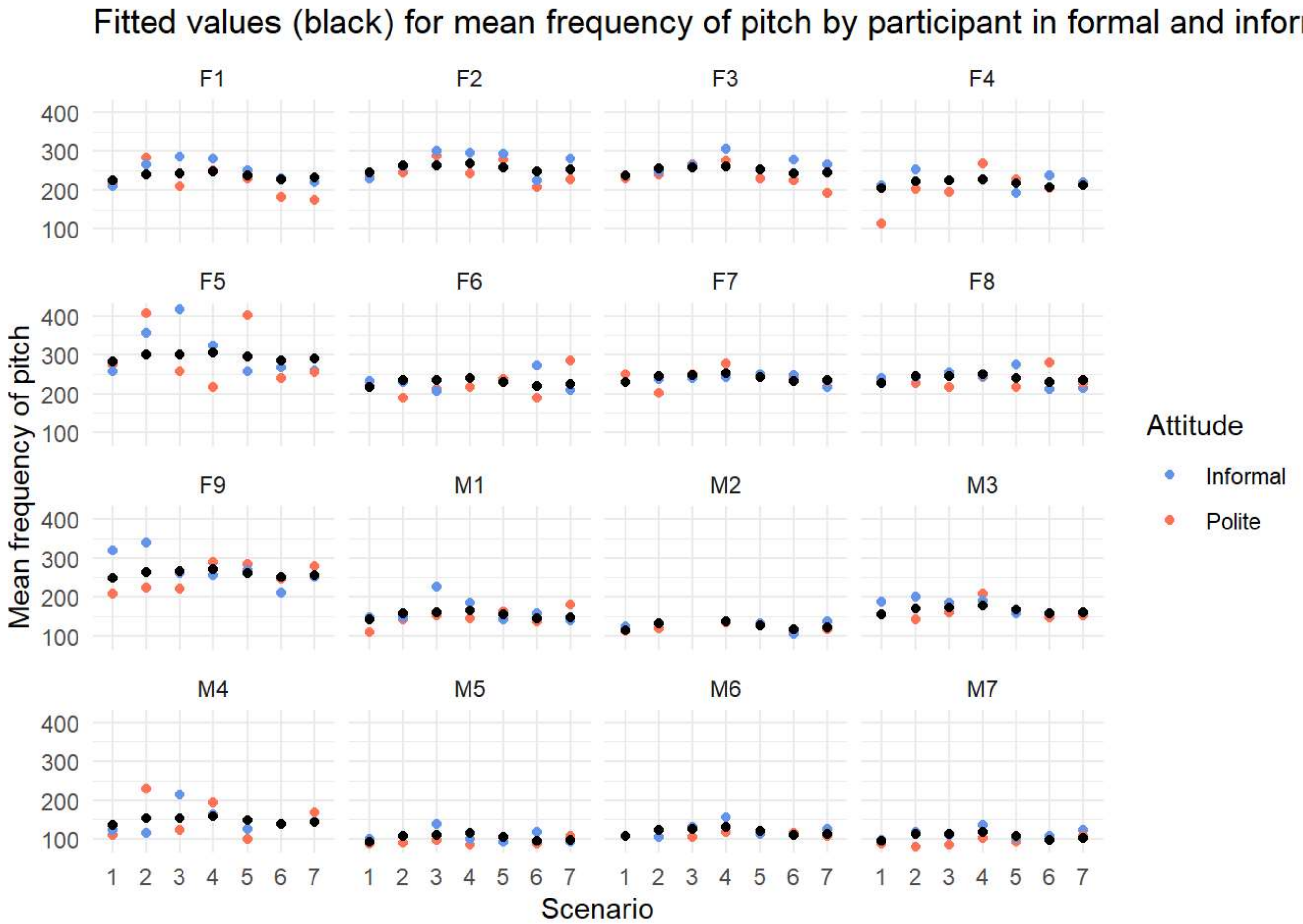


The residuals once again appear approximately normally distributed, albeit with some positive kurtosis (values concentrated around the mean to a greater extent than in a classical normal distribution). The data also shows skewness at the extremes.

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)

```
#Plotted:
ggplot(data = politeness, aes(x = scenario, y = f0mn, color = attitude, group = subject)) +
  geom_point() +
  geom_point(aes(x = scenario, y = fitted(m4)), color = "black") +
  facet_wrap(~subject) +
  guides(color = guide_legend("Attitude")) +
  scale_color_manual(labels = c("Informal", "Polite", "Fitted values"),
                     values = c("cornflowerblue", "coral1", "black")) +
  labs(title = "Fitted values (black) for mean frequency of pitch by participant in formal an
d informal scenarios",
       x = "Scenario",
       y = "Mean frequency of pitch") +
  theme_minimal()
```



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 <- lme4::lmer(f0mn ~ gender + attitude + (1 | scenario) + (1 | subject), data = politenes
s)
```

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

```
m6 <- lme4::lmer(f0mn ~ gender * attitude + (1 | scenario) + (1 | subject), data = politenes
s)
```

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).

The interaction appears to show that much like that of women, Korean mens' pitch decreases in polite situations, albeit to a somewhat lesser extent than that of women, as evidenced by the term 'genderM:attitudepol' being positive.

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.

```
# Residual variance:
c(sum(residuals(m4)^2),
  sum(residuals(m5)^2),
  sum(residuals(m6)^2)
)
```

```
## [1] 181392.4 169253.2 168903.6
```

```
# Residual standard deviation
c(sigma(m4),
  sigma(m5),
  sigma(m6)
)
```

```
## [1] 30.65803 29.71087 29.75684
```

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

```
##    df      AIC
## m4  5 2092.482
## m5  6 2077.131
## m6  7 2072.618
```

Model *m5* has the lowest residual standard deviation, although the difference from the residual standard deviation of *m6* is negligible. Model *m6* has the lowest residual standard deviation and AIC score, with the scores of model *m5* being almost as low.

Considering that these measures of model comparison yield similar results for both models *m5* and *m6*, we decided to choose model *m5* for our analysis, as we concluded that the added complexity of model *m6* did not constitute a significant improvement and may potentially involve overfitting.

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

REPORT

Winter and Grawunder (2012) studied the phonetic profile of Korean speakers in formal and informal scenarios in an experiment of both within- and between-participant design. The collected data contains the following variables: - subject: The ID of the experimental participant. Factor. - gender: Participant's gender. Factor. - scenario: Experimental condition in the within-subject experiment. Factor. - attitude: Is the scenario formal or informal. Factor. - f0mn: Mean frequency of participants' voice pitch during the scenario (higher value indicates higher pitch). Numeric. - total_duration: The length of the interaction within the scenario. Numeric. - hiss_count: Number of hissing intakes of breath during the scenario. Numeric.

We investigated the effects of gender and scenario on participant's pitch using the following linear mixed-effects model:

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

We decided to include separate intercepts for both *scenario* and *subject* as not doing so would have violated the assumption of independence for linear models: the structure of the data is hierarchical, and both the experimental condition and the individual participant introduce underlying categorization into the data.

We also decided to omit including an interaction between gender and scenario as the added complexity did not result in a significantly better fit.

```
##                Estimate Std. Error   t value
## (Intercept)  254.39833   9.597329 26.507201
## genderM     -115.43692  12.880973 -8.961817
## attitudepol  -14.81924   4.096110 -3.617882
```

The output shows that pitch is significantly modulated by gender and attitude, $p < .001$, with women using a higher pitch than men, and informal scenarios eliciting a higher pitch than polite ones. The second-level variables had a variance of 584.4 (subject) and 106.4 (scenario), compared to the residual variance of 882.7.

The residuals of the model are normally distributed: