practical_exercise_2, Methods 3, 2021, autumn semester

Linus Backström

29.9.2021

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
```

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" describes the subjects, with each individual having a unique name. "Gender" describes gender (female or male in this dataset). "Scenario" describes the difference items in the study, such as "asking for a favor" and "excusing for coming too late". "Attitude" describes the two conditions: informal and polite. "total_duration" describes the duration of each response in seconds, while "f0mn" describes the pitch, measured in Hz.

```
i. Also consider whether any of the variables in _politeness_ should be encoded as factors or have the 
```

```
class(politeness$subject)
```

```
## [1] "character"
```

```
class(politeness$gender)
```

```
## [1] "character"
```

```r
class(politeness$scenario)
```

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

```r
class(politeness$attitude)
```

```
## [1] "character"
```

```r
class(politeness$total_duration)
```

```
## [1] "numeric"
```

```r
class(politeness$f0mn)
```

```
## [1] "numeric"
```

```r
class(politeness$hiss_count)
```

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

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

```r
df_f1 <- politeness %>%
  dplyr::filter(politeness$subject == 'F1')

df_f1$scenario_int <- as.integer(df_f1$scenario)

lm_i <- lm(f0mn~scenario_int, data=df_f1)
lm_f <- lm(f0mn~scenario, data=df_f1)
```

```
i. Include the model matrices, $X$ from the General Linear Model, for these two models in your report ar
```

```r
X_i <- model.matrix(lm_i)
X_f <- model.matrix(lm_f)

X_i
```

```
##    (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
```

```
X_f
```

```
##    (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"
```

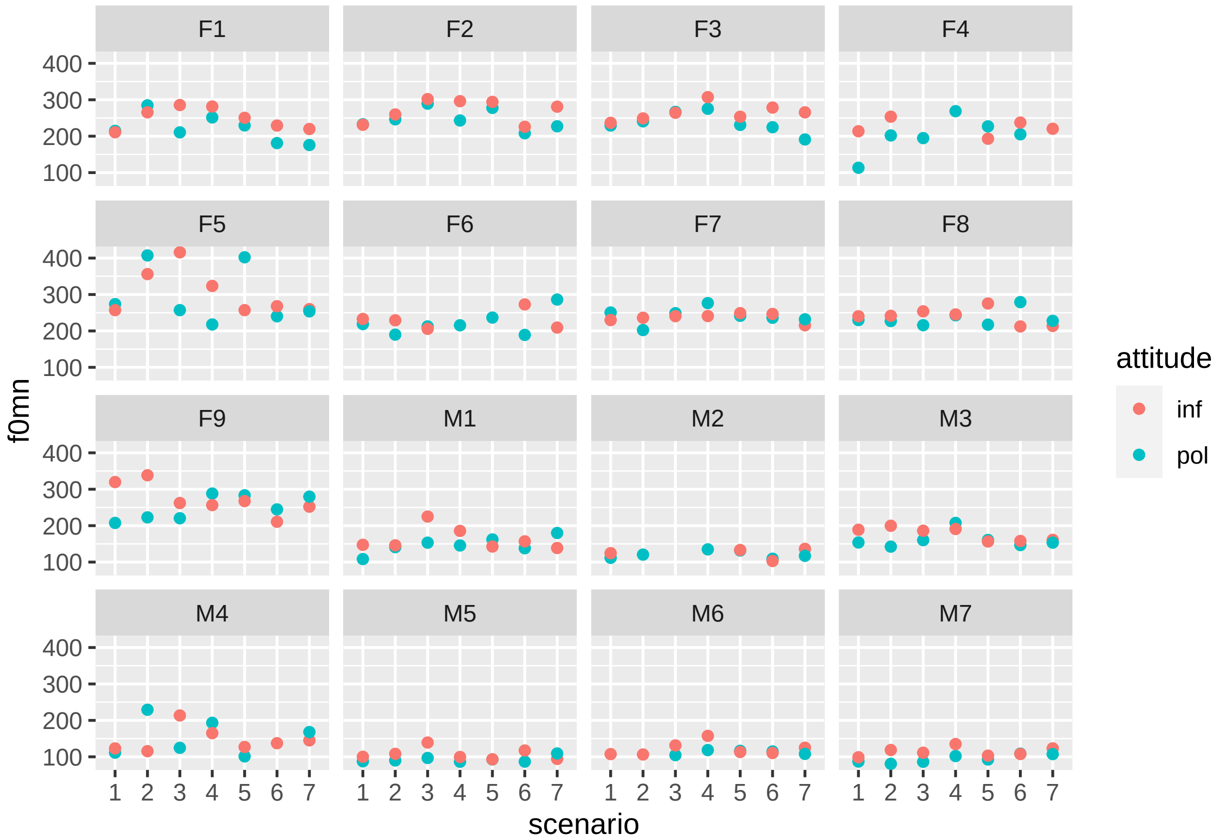
```
ii. Which coding of _scenario_, as a factor or not, is more fitting?
```

Coding it as a factor makes more sense, because the variable is categorical, not continuous. In other words, scenario 6, i.e. item 6, isn't 6x more than scenario 1.

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(scenario, f0mn, color = attitude))+
  geom_point()+
  facet_wrap(politeness$subject)
```

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




```
i. Describe the differences between subjects
```

Males have lower voices overall, and it looks like there may be more variation in pitch height within the female 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

```
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
   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 eac
```

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + (1 | scenario)
##    Data: politeness
##
## REML criterion at convergence: 2144.3
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2314 -0.6033 -0.1599  0.4893  4.2069
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  scenario (Intercept)   91.77   9.579
##  Residual             1478.25  38.448
## Number of obs: 212, groups:  scenario, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  247.786      5.033   49.23
## genderM     -115.875      5.338  -21.71
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.455
```

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

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + (1 | subject)
##    Data: politeness
##
## REML criterion at convergence: 2091.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2200 -0.5402 -0.1385  0.4358  3.8184
```

```
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept)  595.1   24.39
##  Residual             1026.7   32.04
## Number of obs: 212, groups:  subject, 16
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.525      8.641  28.531
## genderM     -115.181     13.080  -8.806
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.661
```

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

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

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: f0mn ~ gender + (1 | scenario) + (1 | subject)
##    Data: politeness
##
## REML criterion at convergence: 2082.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.0131 -0.5373 -0.1089  0.4381  3.7558
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 588.83   24.266
##  scenario (Intercept)  96.17    9.807
##  Residual             939.92   30.658
## Number of obs: 212, groups:  subject, 16; scenario, 7
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  246.765      9.327   26.46
## genderM     -115.175     12.955   -8.89
##
## Correlation of Fixed Effects:
##         (Intr)
## genderM -0.606
```

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

```r
AIC(m1, m2, m3, m4)
```

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

The last model (m4) with two intercepts has the lowest residual standard deviation, and also the lowest AIC.

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

```
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
```

"Subject" explains the most variance.

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_

```
df_3v <- politeness %>%
  select(subject, gender, f0mn) %>%
  filter(!is.na(f0mn)) %>%
  group_by(subject, gender) %>%
  summarise('f0mn' = mean(f0mn))
```

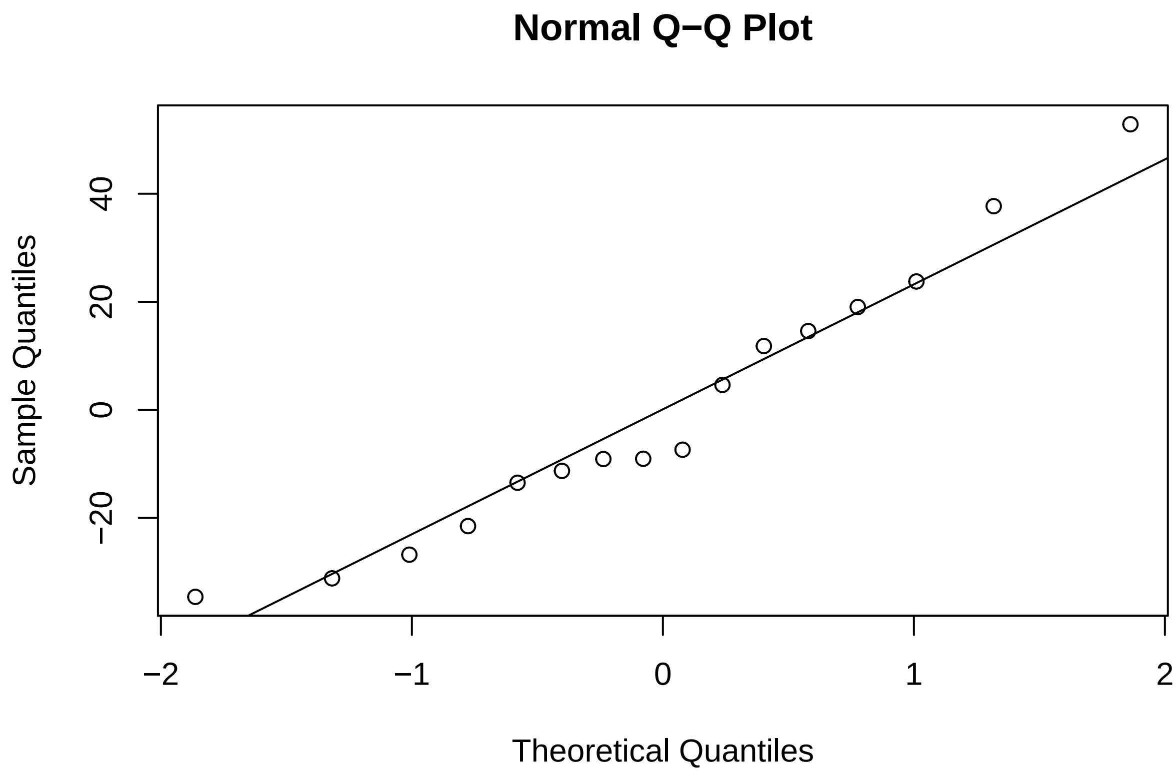
```
## `summarise()` has grouped output by 'subject'. You can override using the `.groups` argument.
```

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

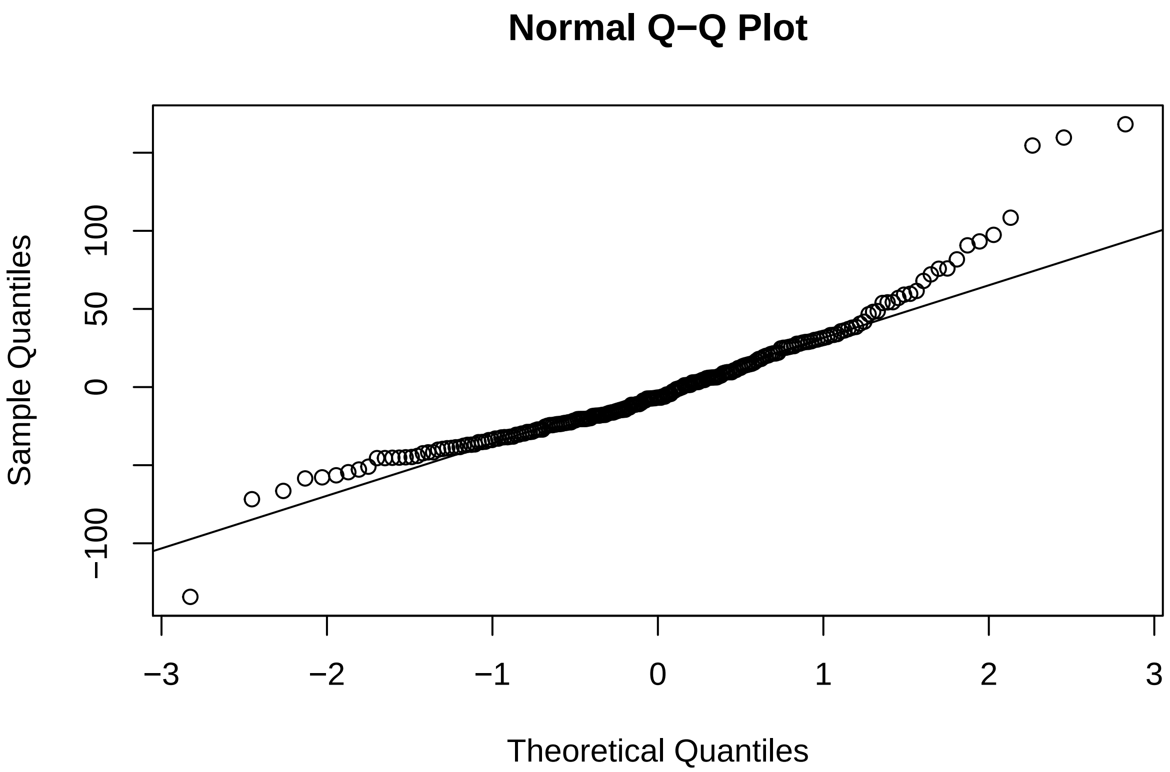
```
m5 <- lm(f0mn~gender, data=df_3v)
```

```
iii. make Quantile-Quantile plots, comparing theoretical quantiles to the sample quantiles) using `qqnor
```

```
qqnorm(residuals(m5))
qqline(residuals(m5))
```



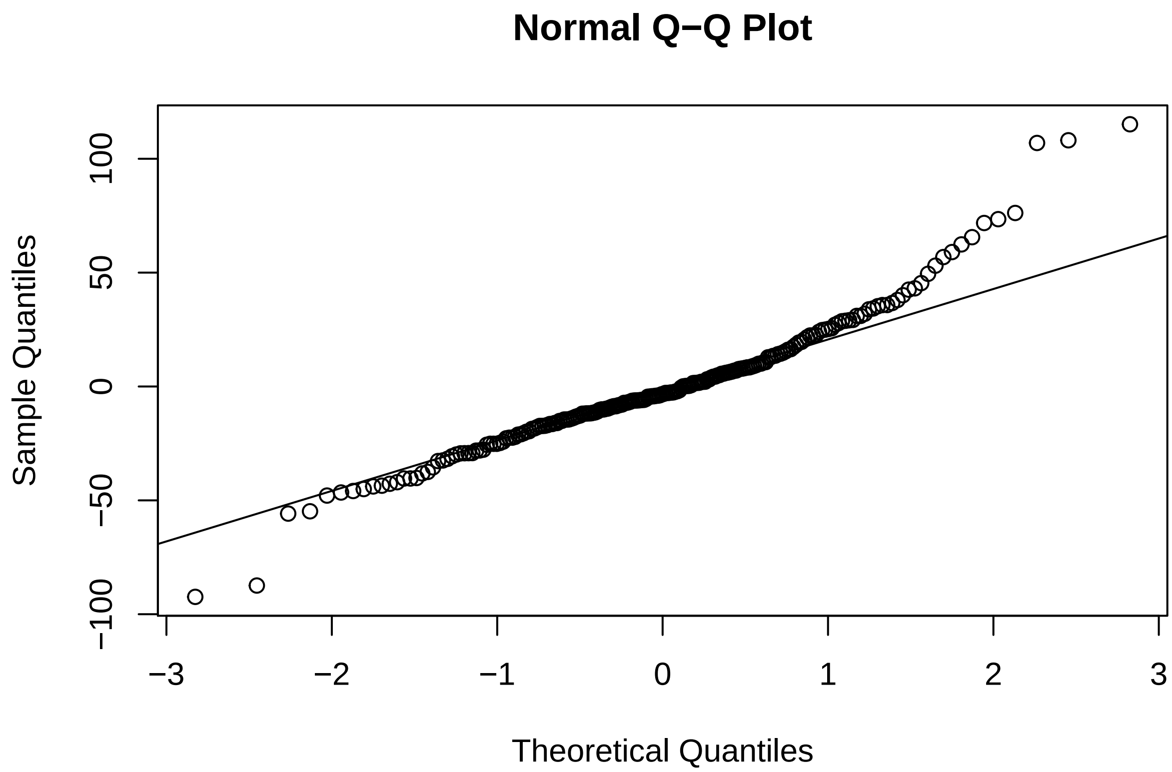

```r
qqnorm(residuals(m1))
qqline(residuals(m1))
```



m5 has less data points so it's hard to say. Maybe m1 is better.

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

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

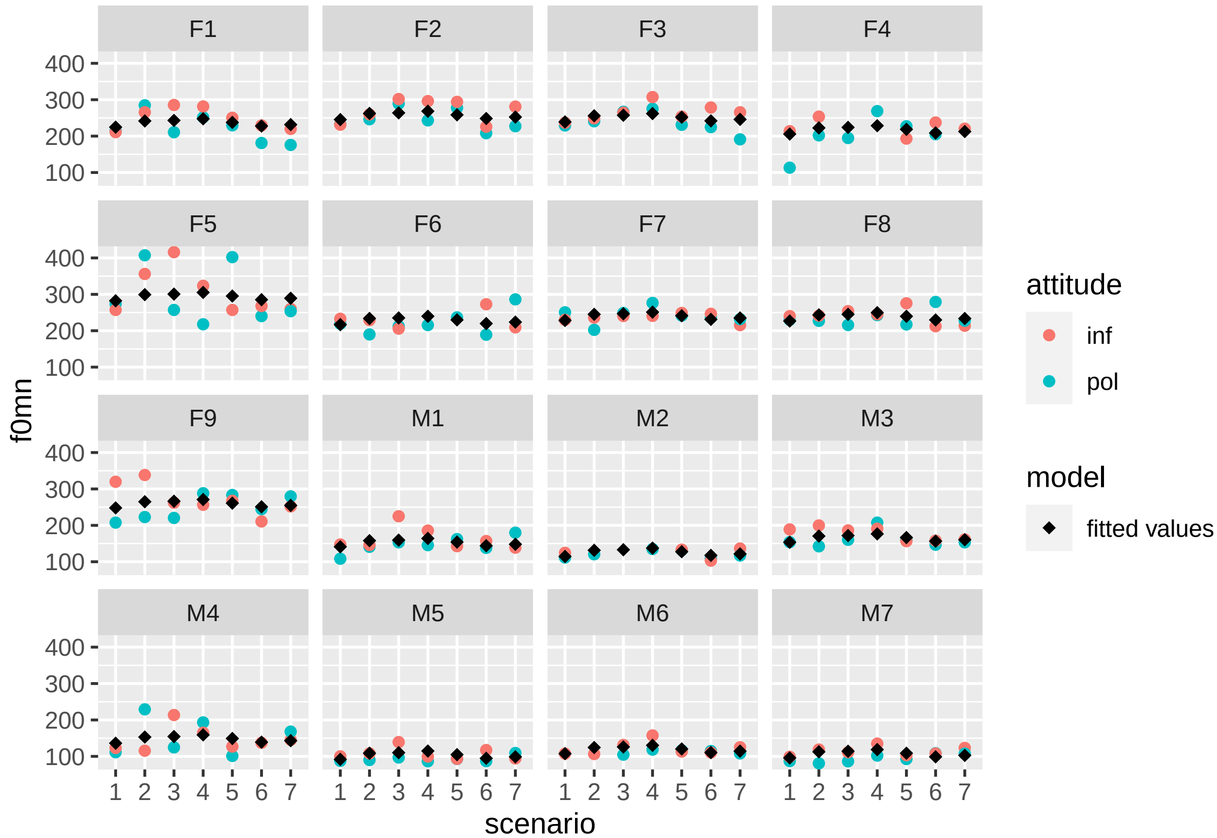


It looks alright.

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)

```
ff <- fixef(m4)
rf <- ranef(m4)
rf <- as.data.frame(rf)

politeness$effect_gender <- 0.0
politeness[politeness$gender == "F", ]$effect_gender <- ff[1]
politeness[politeness$gender == "M", ]$effect_gender <- ff[1] + ff[2]

politeness$intercept_subject <- left_join(politeness, rf, by = c("subject" = "grp"), copy = TRUE, keep =
politeness$intercept_scenario <- left_join(politeness, rf, by = c("scenario" = "grp"), copy = TRUE, keep

politeness$predicted <- politeness$effect_gender + politeness$intercept_subject + politeness$intercept_s

politeness %>% ggplot(aes(scenario, f0mn, color = attitude)) +
    geom_point() +
    geom_point(aes(y = predicted, shape = "fitted values"), color = "black", size = 2) +
    scale_shape_manual(name = "model", values = c(18)) +
    facet_wrap(vars(subject))
```

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



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*

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

```
## 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)  254.398      9.597  26.507
## genderM     -115.437     12.881  -8.962
## attitudepol  -14.819      4.096  -3.618
##
## Correlation of Fixed Effects:
##             (Intr) gendrM
## genderM     -0.587
## attitudepol -0.220  0.006
```

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

```r
m7 <- lmer(f0mn~gender*attitude + (1|scenario) + (1|subject), data=politeness)
summary(m7)
```

```
## 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)          255.618      9.761  26.186
## genderM             -118.232     13.531  -8.738
## attitudepol          -17.192      5.423  -3.170
## genderM:attitudepol    5.544      8.284   0.669
##
## Correlation of Fixed Effects:
##             (Intr) gendrM atttdp
## genderM     -0.606
## attitudepol -0.286  0.206
## gndrM:tttdp  0.187 -0.309 -0.654
```

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

The Korean men's pitch does not decrease as much as it does for the women when they are polite.

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.

```r
m_1 <- lmer(f0mn~gender + (1|scenario) + (1|subject), data=politeness)
m_2 <- lmer(f0mn~gender + attitude + (1|scenario) + (1|subject), data=politeness)
m_3 <- lmer(f0mn~gender*attitude + (1|scenario) + (1|subject), data=politeness)

anova(m_1, m_2, m_3)
```

```
## refitting model(s) with ML (instead of REML)

## Data: politeness
## Models:
## m_1: f0mn ~ gender + (1 | scenario) + (1 | subject)
## m_2: f0mn ~ gender + attitude + (1 | scenario) + (1 | subject)
## m_3: f0mn ~ gender * attitude + (1 | scenario) + (1 | subject)
##     npar    AIC    BIC  logLik deviance   Chisq Df Pr(>Chisq)
## m_1    5 2105.2 2122.0 -1047.6   2095.2
## m_2    6 2094.5 2114.6 -1041.2   2082.5 12.6868  1  0.0003683 ***
## m_3    7 2096.0 2119.5 -1041.0   2082.0  0.4551  1  0.4998998
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
sum(residuals(m_1)^2)
```

```
## [1] 181392.4
```

```r
sum(residuals(m_2)^2)
```

```
## [1] 169253.2
```

```r
sum(residuals(m_3)^2)
```

```
## [1] 168903.6
```

```r
AIC(m_1, m_2, m_3)
```

```
##     df      AIC
## m_1  5 2092.482
## m_2  6 2077.131
## m_3  7 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:

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)

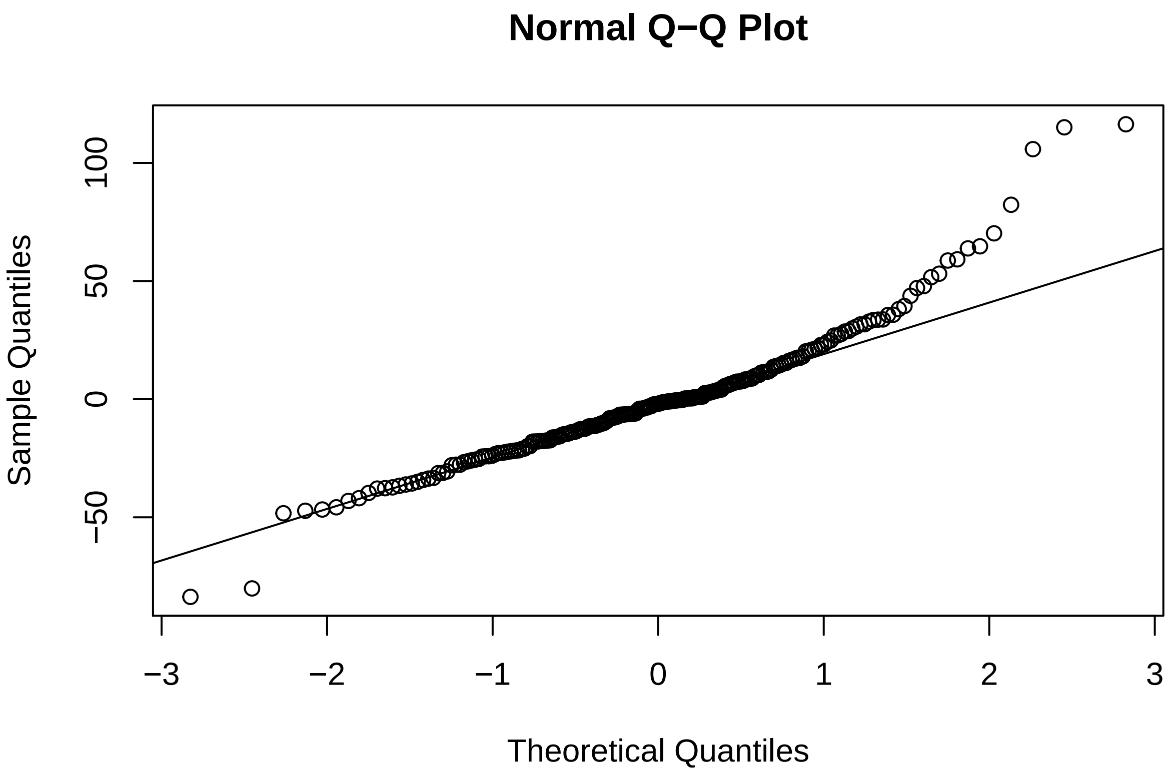
iv. describe the variance components of the second level (if any)

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

I used R (R Core Team, 2020) and lme4 (Bates, Maechler, Bolker & Walker, 2015) to perform a linear mixed effects analysis on the relationship between pitch and politeness. As fixed effects, I entered politeness and gender, including the interaction between them. As random effects, I used intercepts for subjects and scenarios. Random intercepts were used because I assumed that different subjects would have different baselines for the pitch of their voice, and that different scenarios would also have different baselines (e.g. excusing for being late may call for a different pitch than asking fore a favor).

Male subjects had lower pitch than female subjects. Subjects had lower pitch when speaking with the polite condition, and this effect was stronger for female subjects.

```r
qqnorm(residuals(m_3))
qqline(residuals(m_3))
```


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