PANEL DATA ANALYSIS: LINEAR MIXED-EFFECT MODELS

Introduction and Motivation

Panel Data Econometrics vs. Mixed Models

- Pane data econometrics has its counterparts in the statistic literature on *mixed effect*, *hierarchical models*, or *models for longitudinal data*.

- In part they are just differences in terminologies used, but in part there are substantial distintion.

- For mixed models, $ has the long-standing **nlme** package (Pinheiro et al. 2007) and the more recent **lme4** package (Bates 2007).

- In this course, we will use the **lme4** package.

Introduction and Motivation

pitch

- This is a pretty common example: modeling \$ as a function of age

$$pitch_i = age_i + \epsilon_i$$

where $i = 1, 2, \ldots, n$

- We called "age" a fixed effect, and ϵ as "error term" to represent the deviations from predictions due to "random" factors that we are not able to control for experimentally.

- Note that the use of the term "fixed effect" here, which may be confusing to you.

- In the mixed model, we add one or more random effects. These random effects give structure to the error term ϵ, which in the linear regression model should not have any structure.

- Now, consider the study Winter & Grawunder, 2012. "The Phonetic Profile of Korean Formality", Journal of Phonetics. They studied the relationship between *pitch* and *politness*, and they add $ as an additional fixed effect. Using *R*, the model is

$$pitch \sim politness + \epsilon$$

where *politness* is entered as a binary variable with two levels.

- Add an additional fixed effect: *sex*

$$pitch \sim politness + sex + \epsilon$$

- Their study took multiple measures per subjects: each subject gave multiple polite responses and multiple informal responses.

- Multiple measures per subjects immediately would violate the independence assumption (in linear regression models),as multiple responses from the same subject cannot be regarded as independent from each other

- Importantly, every person has a slightly different voice pitch; thus, this is an individual-specific factor that affects all responses from the same subject, rendering these different responses inter-dependent rather than independent.

- We will model this individual-specific factor as a random effect for the subjects, allowing for a different "baseline" pitch value for each subject. Note that this type of situation is very common in many different applications in practice.

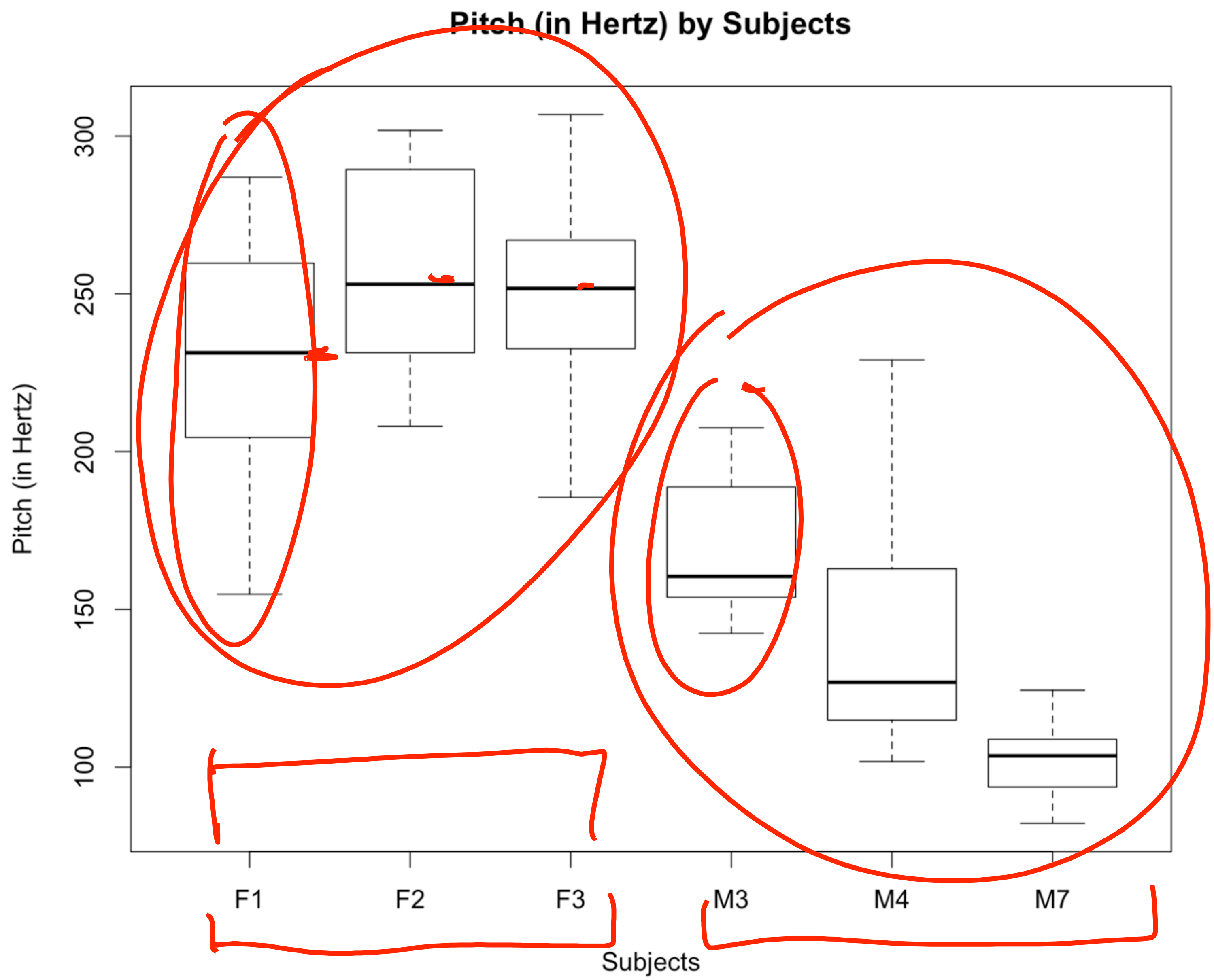
Let's look at the data:

```
> politeness= read.csv("http://www.bodowinter.com/tutorial/politeness_data.csv")
> str(politeness)
'data.frame':	84 obs. of  5 variables:
 $ subject  : Factor w/ 6 levels "F1","F2","F3",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 1 1 1 1 1 ...
 $ scenario : int  1 1 2 2 3 3 4 4 5 5 ...
 $ attitude : Factor w/ 2 levels "inf","pol": 2 1 2 1 2 1 2 1 2 1 ...
 $ frequency: num  213 204 285 260 204 ...
```

```
> table(politeness$subject)

F1 F2 F3 M3 M4 M7
14 14 14 14 14 14
```

```
> which(is.na(politeness$frequency))
[1] 39
```


Pitch (in Hertz) by Subjects
Pitch (in Hertz)
300
250
200
150
100
F1
F2
F3
M3
M4
M7
Subjects

- We can model these individual differences using different (random) intercepts for each subject: each subject has a different estimated intercept, and the mixed model can be used to estimate these intercepts.

- In the mixed model, we add one or more random effects to the fixed effects. These random effects give a structure to the error term ϵ.

- In our example, a random effect for "subject" is added, and this characterizes idiosyncratic variation that is due to individual differences.

- The updated model is

$$pitch \sim politness + sex + (1|subject) + \epsilon$$

- This formula, in R notation, models "an intercept that's different for each subject" and "1" stands for the intercept.

- Think of this formula as telling the model that it should expect multiple responses per subject and the responses depend on each subject's baseline level.

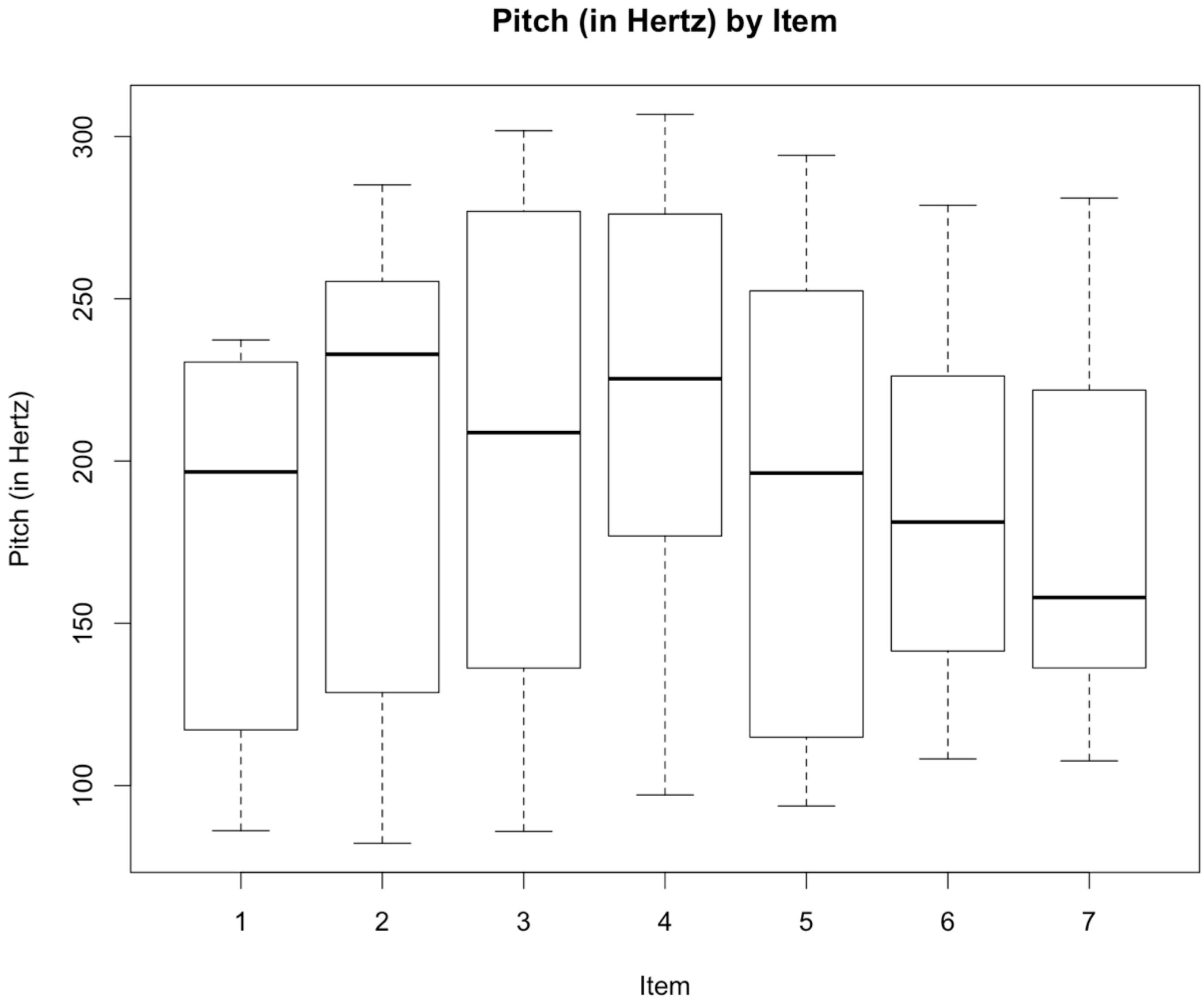
- Importantly, this framework resolves the non-independence that stems from having multiple responses by the same subject.

- In Winter and Grawunder's study, there were different items. One item, for example, was an "asking for a favor" scenario. Subjects had to imagine asking a professor for a favor (polite condition), or asking a peer for a favor (informal condition). Another item was an "excusing for coming too late" scenario, which was similarly divided between polite and informal. There were 7 different items in their study.

- Hence, we should also expect "*item-specific*" variation.

- Most importantly, the different responses to one item cannot be regarded as independent. If these interdependencies are not accounted for, the independence assumption would be violated.

By-Item Variation

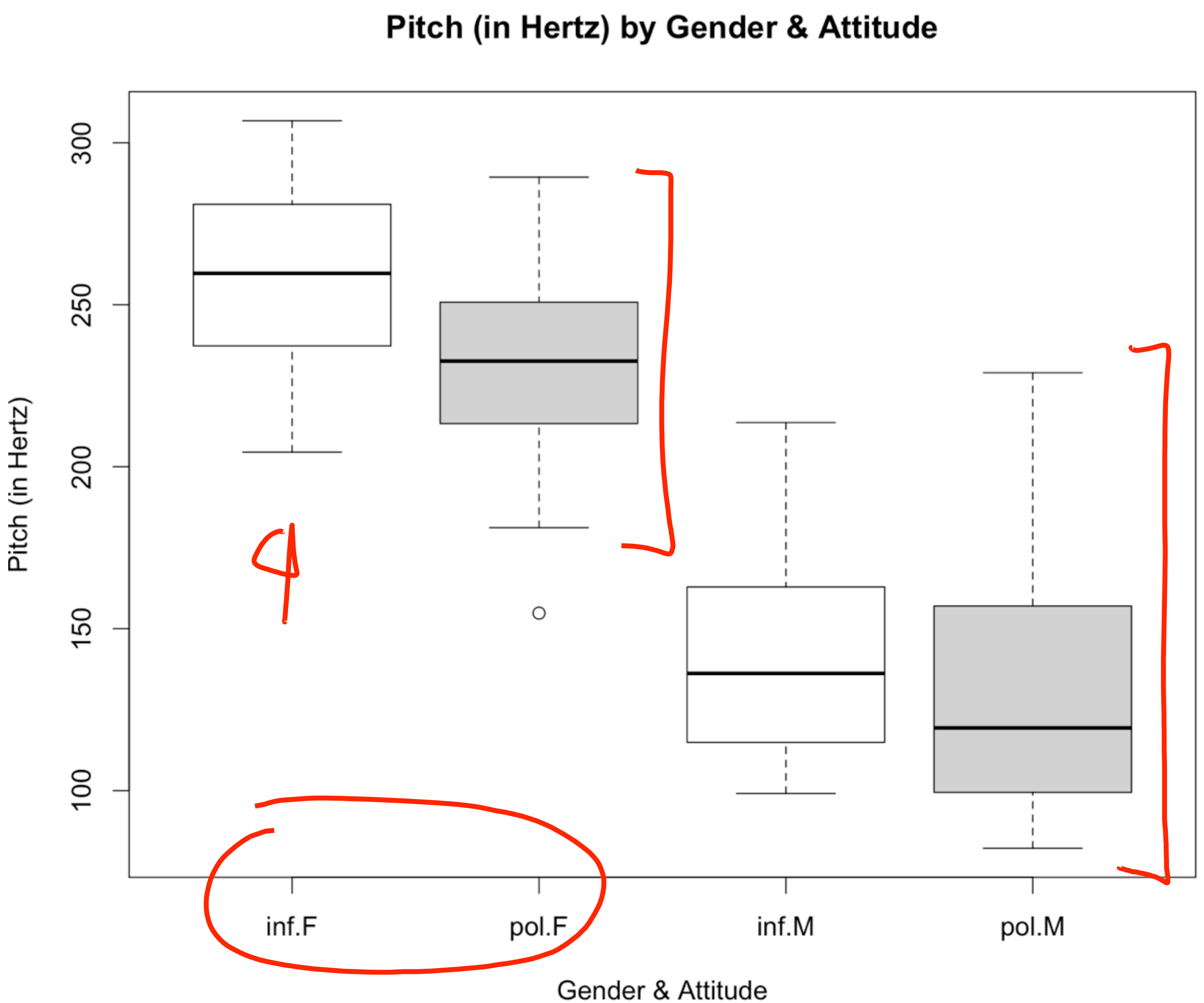


- The model now becomes

$$pitch \sim politness + sex + (1|subject) + (1|item) + \epsilon$$

- Note that in addition to different intercepts for different subjects, the model includes different intercepts for different items. As such, non-independencies are accounted for - the model knows that there are multiple responses per subject and per item, and there are by-subject and by-item variation in overall pitch levels.

- Let's look at the relationship between politeness and pitch using a boxplot:


Pitch (in Hertz) by Gender & Attitude
Pitch (in Hertz)
300
250
200
150
100
inf.F
pol.F
inf.M
pol.M
Gender & Attitude
4

- Estimate the model using functions in R's lme4 package:

```
politeness.lmm = lmer(frequency ~ attitude + (1|subject) + (1|scenario),
data=politeness)
summary(politeness.lmm)
```

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

REML criterion at convergence: 793.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2006 -0.5817 -0.0639  0.5625  3.4385

Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept)  219     14.80
 subject  (Intercept) 4015     63.36
 Residual              646     25.42
Number of obs: 83, groups:  scenario, 7; subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  202.588     26.754   7.572
attitudepol  -19.695      5.585  -3.527

Correlation of Fixed Effects:
            (Intr)
attitudepol -0.103
```

- One of the biggest differences between this model output and those we have seen earlier in this course is the inclusion of the Random effects:

- Have a look at the column standard deviation under random effects. This measures of how much variability in the dependent measure there is due to scenarios and subjects (our two random effects).

- The scenario ("item") has much less variability than subject, which we have already seen in the boxplots from above. THere is more idiosyncratic differences between subjects than between items, as expected.

- "Residual"stands for the variability that's not due to either scenario or subject. This corresponds to ϵ, the "random" deviations from the predicted values that are not due to subjects and items.

- It reflects the fact that each and every utterance has some factors that affect pitch that are outside of the purview of our experiment.

- The fixed effects output mirrors those in linear regression models

- The coefficient "*attitudepol*" is the slope for the categorical effect of politeness. -19.695 means that the pitch on average goes down by -19.695 Hz when going from "informal" to "polite". That is, pitch is lower in polite speech than in informal speech, by about 20 Hz.

- The model intercept 202.588 Hz is the average of our data for the informal condition, without distinguishing the gender difference. As we didn't inform our model that there are two sexes in our dataset, the intercept is particularly off, in between the voice pitch of males and females.

- Let's add gender to the model and re-estimate it again.

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

REML criterion at convergence: 775.5

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.2591 -0.6236 -0.0772  0.5388  3.4795

Random effects:
 Groups   Name        Variance Std.Dev.
 scenario (Intercept) 219.5    14.81
 subject  (Intercept) 615.6    24.81
 Residual             645.9    25.41
Number of obs: 83, groups:  scenario, 7; subject, 6

Fixed effects:
            Estimate Std. Error t value
(Intercept)  256.846     16.116  15.938
attitudepol  -19.721      5.584  -3.532
genderM     -108.516     21.013  -5.164

Correlation of Fixed Effects:
            (Intr) atttdp
attitudepol -0.173
genderM     -0.652  0.004
```

- We added "gender" as a fixed effect because the relationship between sex and pitch is systematic and predictable
- Compared to the last model without the gender fixed effect, the variation associated with the random effect "subject" dropped considerably.
- This is because the variation that's due to gender was confounded with the variation that's due to subject.
- Under the Fixed effects, we see that males and females differ by about 109 Hz. The intercept is now much higher (256.846 Hz), as it now represents the female category under the informal condition.
- The coefficient for the effect of attitude didn't change much.

- Importantly, *p-values* for mixed models are not as straightforward as they are for the linear regression model.

- We can use the **Likelihood Ratio Test** as a means to attain p-values. To implement it in *R*, we will have to re-run the model twice, but setting the argument *REML* to FALSE.

```
politeness.null = lmer(frequency ~ gender + (1|subject) + (1|scenario),
data=politeness, REML=FALSE)

politeness.full = lmer(frequency ~ attitude + gender + (1|subject) +
(1|scenario), data=politeness, REML=FALSE)

anova(politeness.null,politeness.full)
```

```
Data: politeness
Models:
politeness.null: frequency ~ gender + (1 | subject) + (1 | scenario)
politeness.full: frequency ~ attitude + gender + (1 | subject) + (1 | scenario)
                Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
politeness.null  5 816.72 828.81 -403.36   806.72
politeness.full  6 807.10 821.61 -397.55   795.10 11.618      1  0.0006532 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We would report the result the following way: "… politeness affected pitch ($\chi^2(1) = 11.62, p = 0.00065$), lowering it by about $19.7Hz \pm 5.6$ (standard errors) …"

- Note that we kept the predictor “gender”" in the model. The only change between the full model and the null model that we compared in the likelihood ratio test was the factor of interest, *politeness*. In this particular test, think of “gender” as a control variable and of “attitude” as the test variable.

- **Interaction:** What happens if we have an interaction term? Suppose we predicted “attitude” to have an effect on pitch that is somehow modulated through “gender”.

- For example, it could be that speaking politely versus informally has the opposite effect for men and women. Or, it could be that women show a difference and men don’t (or vice versa)

full model:

$$frequency \sim attitude * gender$$

reduced model:

$$frequency \sim attitude + gender$$

- Comparison of the above models in a likelihood ratio test using the $anova()$ function produces a p-value that gives the significance of the interaction.
- If it is significant, attitude and gender are significantly inter-dependent on each other.

Berkeley SCHOOL OF INFORMATION