# Mixed Effects Models 1: Random Intercept

Yury Zablotski

Last updated on Dec 14, 2019 · 30 min read · 0 Comments · R

## Previous topics

- Repeated Measures ANOVA
- Multiple linear regression

## Why do we need it? What are the benefits?

- Repeated measurements violate the “independence of observations” assumption of linear models. This adds some random variation to the model and so, leads to unrealistic results. Mixed Effects Models (MEM) remove such (random) variation of repeated measurements from the model and so, secure realistic results.
- Compared to the Repeated Measures ANOVA, MEMs have no problem with a few missing values, show the degrees of freedom, are able to handle disballanced groups (different number of observations) and heteroscestadicity. Besides, MEM is more flexible, accurate, powerful and delivers more inferences.

## How do Mixed Effects Models work

**Mixed effects** come from **fixed (or between) effects**, which is a usual linear regression, and **random (within) effects**, which you might already know as the **error term** from the *Repeated Measures ANOVA*. If not, I'll explain both effects in the next chapter.

Load all needed packages to avoid interruptions.

```r
library(tidyverse) # data wrangling and visualization
library(lme4)      # "golden standard" for mixed-effects modelling in R (no p-values)
library(lmerTest)  # p-values for MEMs based on the Satterthwaite approximation
library(psycho)    # mainly for an "analyze()" function
library(broom)     # for tidy results
library(knitr)     # beautifying tables
library(sjPlot)    # for visualising MEMs
library(effects)   # for visualising MEMs
library(report)    # for describing models
library(emmeans)   # for post-hoc analysis
```

## Fixed and random effects

**Explanatory** variables (predictors) in a usual model always explain **some** of the variation of the response variable, but they never explain 100%. The remaining, or unexplained variance is the **error** of the model. Error comes from the fact that we can't collect all the needed data to explain everything, or we can't completely control the experiments. Simply put - we can't always control, and therefore explain, reality. For instance, we can't control weather or what people think. Often we don't have enough resources for an experiment, be it money or number of Guinea pigs, and are forced to collect **data from the same person (animal, process) repeatedly**. And although more data often leads to better conclusions, unfortunately, additional data may also add **more variation** or uncertainty.

Such additional variation might either (1) increase the error (unexplained variation) by "diluting" the real (fixed) effect of predictors, making them insignificant, or (2) reduce the error by over-explaining predictors and making predictors significant, while in (a fixed) reality they are not.

Particularly, opinions of different people on the same topic would certainly differ (vary). Opinions are **subjective**. But even the same person might slightly change the opinion depending on the circumstances (e.g. after an easy workday or after a stressful one), age or even who asks. And since there are so many different people in the world, we can only ask a small group of **random** individuals. Thus, you see that an opinion is very far from being **fixed**, since it will always vary. Such random variation is therefore called **random effect**. It's often called a **subjective variation** or **subjective error**. Random effects are not related to the mathematical randomness, though!

The height or gender of any person at a given time is **fixed**, because the possibilities are not unlimited, not random.

One of the 6 assumptions of the usual linear regression is **the independence of observations**. It's the strictest one. If we take repeated measurements of e.g. blood pressure from the same group of people on different days, we **violate this assumption heavily**, because we would have two different sources of variation (effects) in our data. Namely (1) **fixed** (for independent observations), which measures the variance of blood pressure between different people (e.g. choleric and melancholic), (2) and **random** (for dependent ones) which will contain the variation of each particular person on different days (think about a day after a heavy party and no sleep as opposed to a day after a chilled-out movie evening). Models, which are able to handle both kinds of effects (fixed and random) at the same time are - **mixed effects models (MEM)**.

## Fixed or random effects?

**Whether the variable is fixed or random does not depend on the variables themselves, but heavily depends on your research question! Sometimes, the same variable could be considered either a random or a fixed effect, and sometimes even both at the same time!**

If you want to study the effect of a particular variable on the response variable, use it as a fixed effect (or simply a categorical) variable. For instance, if we want to study the voice-frequency at seven particular *scenarios* and want to make predictions about them, then *scenario* would be fitted as a fixed effect.

If you don't care about the effect of a particular variable on the response variable, but you know that such variable can somehow influence the result, use it as a random effect. **Random effects are always categorical grouping factors for which we are trying to control**. Never treat a continuous variable as a random effect!

Random effects are often non-systematic, unpredictable and infinite, having a *random* influence on your data. In experiments, that's often *individuals*. Fixed effects are finite, for instance only two categories in the *gender* variable, and are expected to have a systematic and predictable influence on your data.

https://stats.stackexchange.com/questions/37647/what-is-the-minimum-recommended-number-of-groups-for-a-random-effects-factor
(this discussion recommends using at least 6)

## The golden rule is > 5

**The golden rule: random effect suppose to have at least five levels**. This is why sex in the politeness data, having only two levels: male or female, is a fixed effect. The problem with <5 levels is, that **estimating variance of few data points** (say 3) is very imprecise, so that you'll get a result, but not the one you can trust!

Ok, enough theory, let's look at a practical example.

response ~ 1: calc. the mean of the response
(1|subject): calc. one intercept for every subject

## MEM in R

A simple linear model without predictors calculates the mean of a response variable. This mean is called - *Intercept* and the model without predictors is called - **Intercept-only-model:** `response ~ 1`. Example would be a mean of blood pressure of several patients. However, two patients, e.g. choleric and melancholic, might have very different blood pressure. And if we only have two patients, we need to **repeat the measurements** on several days to get more data. If we don't account for repeated measurements, we'll finish up with a single mean (a single intercept), which won't cover subjective differences between them. Simply put, we'll loose information and will make a completely wrong conclusion. MEMs accounts for such **subjective variation** by calculating **several intercepts (means), one for each subject (patient)**, and thus MEM will catch the differences in blood pressure between a choleric and a melancholic. And since MEM can measure such subjective (individual) variation, it can also remove it from the fixed effects, making fixed effects more realistic.

Removing individual differences from the model solves the problem with the "independence of observations" assumption, which was caused by repeated measurements of subjects.

In *R* you would write a following formula for a mixed effects model: `blood_pressure ~ age + (1|subject)`, where *age* is a fixed effect we are interested in, and *subject* is a random effect. If you recall a formula of an **intercept only model** - `response ~ 1`, you'll remember that 1 in the formula **is the Intercept**. This makes the formula of the mixed model intuitively understandable: `(1|subject)` means `(one intercept for every subject)`.

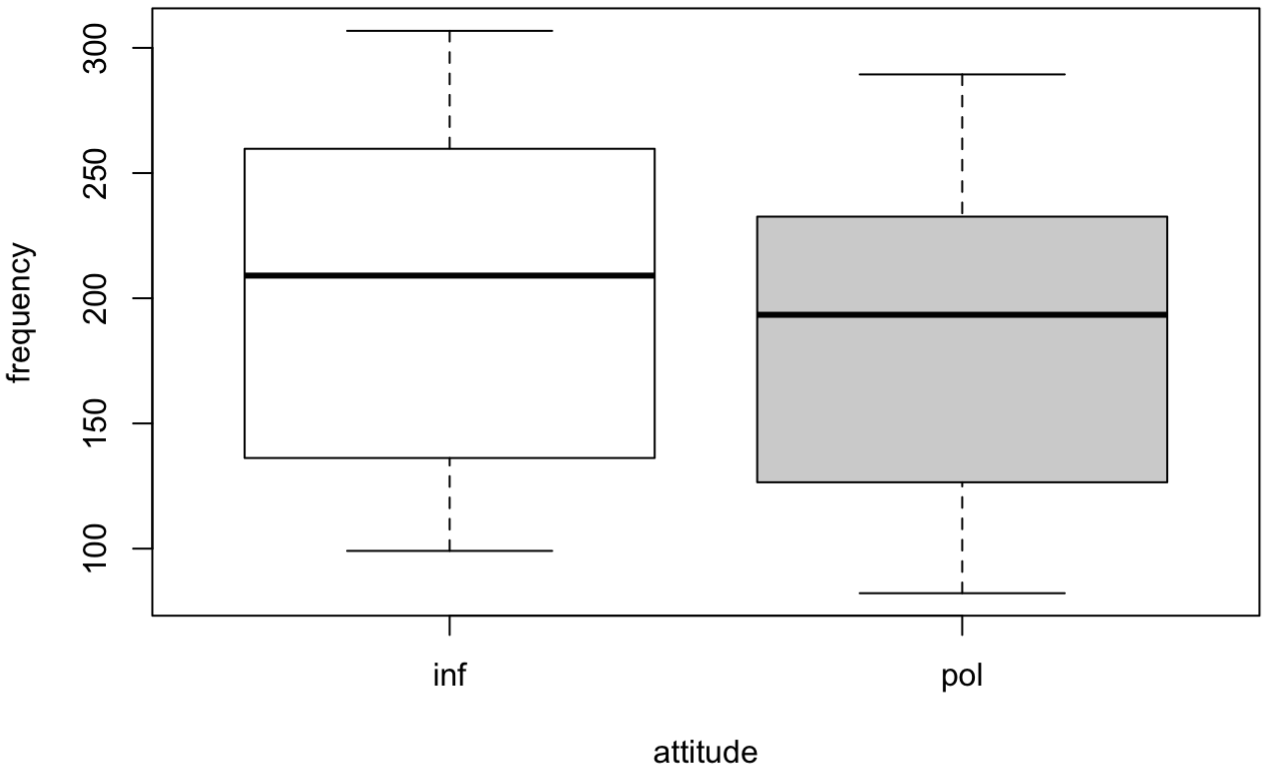
Let's reinforce this moment with an example, which studies the frequency of voice. The data is borrowed from Bodo Winter

```
politeness = read_csv("politeness_data.csv")
```

```
## Parsed with column specification:
## cols(
##   subject = col_character(),
##   gender = col_character(),
##   scenario = col_double(),
##   attitude = col_character(),
##   frequency = col_double()
## )
```

Visualize the data:

```
boxplot(frequency ~ attitude,
col=c("white","lightgray"),politeness)
```



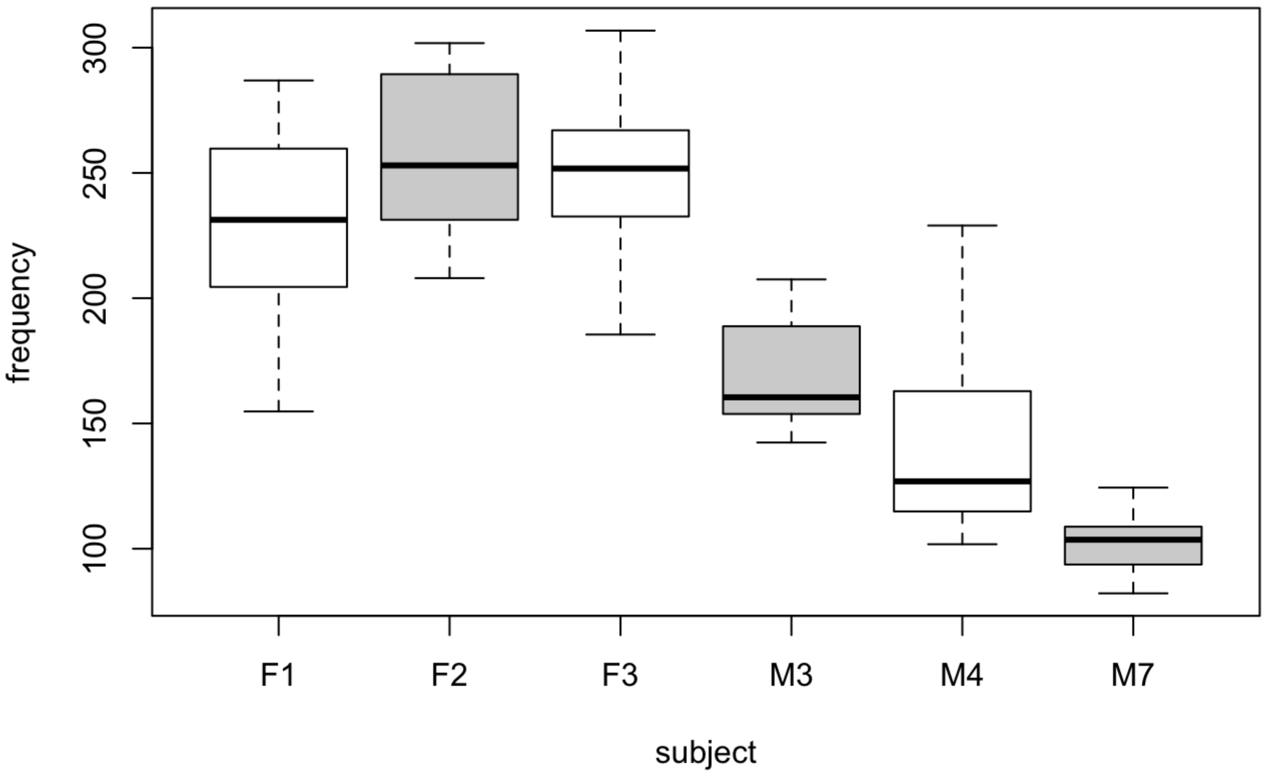
Our response variable - *voice-frequency* on the y-axis does not seem to differ in terms of *attitude* - polite vs. informal. And a simple linear model below confirms that there is no significant differences (p = 0.2) in between different *attitudes*:

```
politeness.model = lm(frequency ~ attitude, data=politeness)
summary(politeness.model) %>% tidy() %>% kable()
```

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 202.5881 | 10.07528 | 20.107444 | 0.000000 |
| attitudepol | -18.2320 | 14.33521 | -1.271833 | 0.207072 |

However, the boxplots above include the subjective variation of 6 different *subjects* (people) within the *attitude*-predictor, and if we extract this random effect of the *subjects* from the fixed effect of *attitude* by including this exact random effect of the *subjects* into the model in this form `(1|subject)`, the difference in *attitudes* suddenly becomes significant (p = 0.003):

```
boxplot(frequency ~ subject,
col=c("white","lightgray"),politeness)
```



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

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: frequency ~ attitude + (1 | subject)
##    Data: politeness
## 
## REML criterion at convergence: 804.7
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -2.2953 -0.6018 -0.2005  0.4774  3.1772 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  subject  (Intercept) 3982     63.10   
##  Residual              851     29.17   
## Number of obs: 83, groups:  subject, 6
## 
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)    
## (Intercept)  202.588     26.151   5.150   7.747 0.000501 ***
## attitudepol  -19.376      6.407  76.003  -3.024 0.003399 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Correlation of Fixed Effects:
##             (Intr)
## attitudepol -0.121
```

**Interpretation:** Summary of the MEM first includes information about the *random effects*. Here, the column *Groups* shows the random effect variables. The column *Name* indicates that this random effect only affects the intercept. The line *subject* refers to the variation due to subjects.The line *Residuals* is the general variation that cannot be explained by the model - the real error term. Important is that the standard error of the slope became much smaller (6.4) compared to the ordinary regression slope (14.3), thus, our model became more precise!

**The sum of the participant and residual variances in MEM is exactly the same as the residual variance of LM**.(Theoretically. I somehow could not prove it with my models? Please, comment below if you know how.)

Is the random effect important? Hell, yeah! And there are two ways to prove that:

1. The amount of variance explained by the *subject* is huge = 3982. How do we know that this number is huge? We can divide the variance for the *subject* by the total variance (variance of the random effect + residuals).

   3982/(3982 + 851) # 82 %

So, the differences between *subjects* explain ca. 82 % of the variance that's "left over" after the variance explained by our fixed effects (which, according to `report(politeness.model)` is only 1.93 %). Since random effect of *subject* explains the most of models variance, I conclude that this effect is significant.

2. And if we compare two models, one with a random effect of subjects (1|subject) and one without, we'll see that the model with the random effect is significantly ($p < 0.05$) better. And **since the only thing which differs between these two models is the random effect (1|subject), we can conclude that this random effect is significantly important**! Thus, we have to add this **random** effect into the (**fixed**) linear model, making it a **mixed-effect-model**.

Hold on! How does such comparison actually work?

## Why and how do we compare models?

The models are compared via the **Likelihood ratio test**, where **Likelihood is the probability of seeing the data you collected given your model.** The logic of the **likelihood ratio test** is to **compare the likelihoods, or Akaike Information Criteria (AIC), of two models with each other**. First, the model without the factor that you're interested in (the null model), secondly, the model with the factor that you're interested in.

It also works for interactions! Compare one model with "+" between two predictors and the other with "*" between them (star means - interaction). If the model with interaction is significantly better (lower AIC and significant p-value), then the interaction is important and have to be considered. By the way, if the interaction is not important it is still a valuable inference which can and should be reported in the publication.

```
m1 <- lm(frequency ~ attitude, data=politeness)
m2 <- lmer(frequency ~ attitude + (1|subject), data=politeness, REML = F)
anova(m2, m1)
```

```
## Data: politeness
## Models:
## m1: frequency ~ attitude
## m2: frequency ~ attitude + (1 | subject)
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m1    3 933.22 940.48 -463.61   927.22
## m2    4 826.51 836.18 -409.25   818.51 108.71  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To obtain the estimated random coefficients we use the `ranef()` functions. `fixef()` gives us only the fixed effects. The intercept of a particular subject is the sum of the estimated random intercept and the mean intercept value reported in the fixed-effects section of the summary: `fixef(m2)[1] + ranef(m2)$subject`. The `coef()` function adds them all for us. The slopes for every subject can be calculated by simply adding the fixed slope (-19.37) of our MEM to the individual intercept: `fixef(m2)[1] + fixef(m2)[2] + ranef(m2)$subject`.

The `dotplot()` function from `lattice` visualizes the random effects. The `condVar=T` argument in the `ranef()` allows to plot 95% confidence intervals of the estimated random effects.

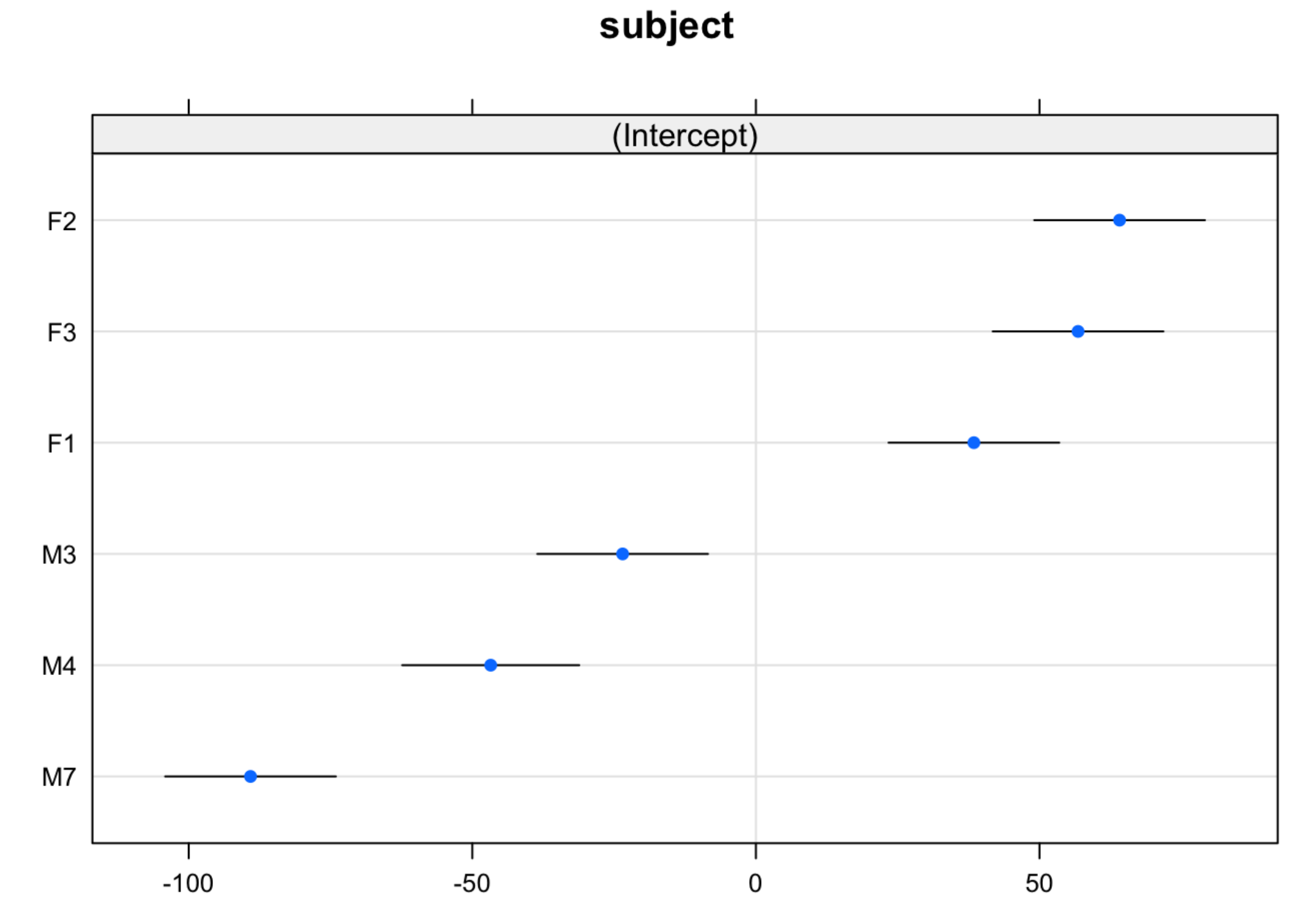
Now, having only 3 woman and 3 man with repeated measurements (see below), we have 6 different intercepts, one for each subject:

```
coef(m2)
```

```
## $subject
##    (Intercept) attitudepol
## F1    241.0250   -19.37241
## F2    266.7092   -19.37241
## F3    259.3919   -19.37241
## M3    179.0908   -19.37241
## M4    155.8311   -19.37241
## M7    113.4805   -19.37241
##
## attr(,"class")
## [1] "coef.mer"
```

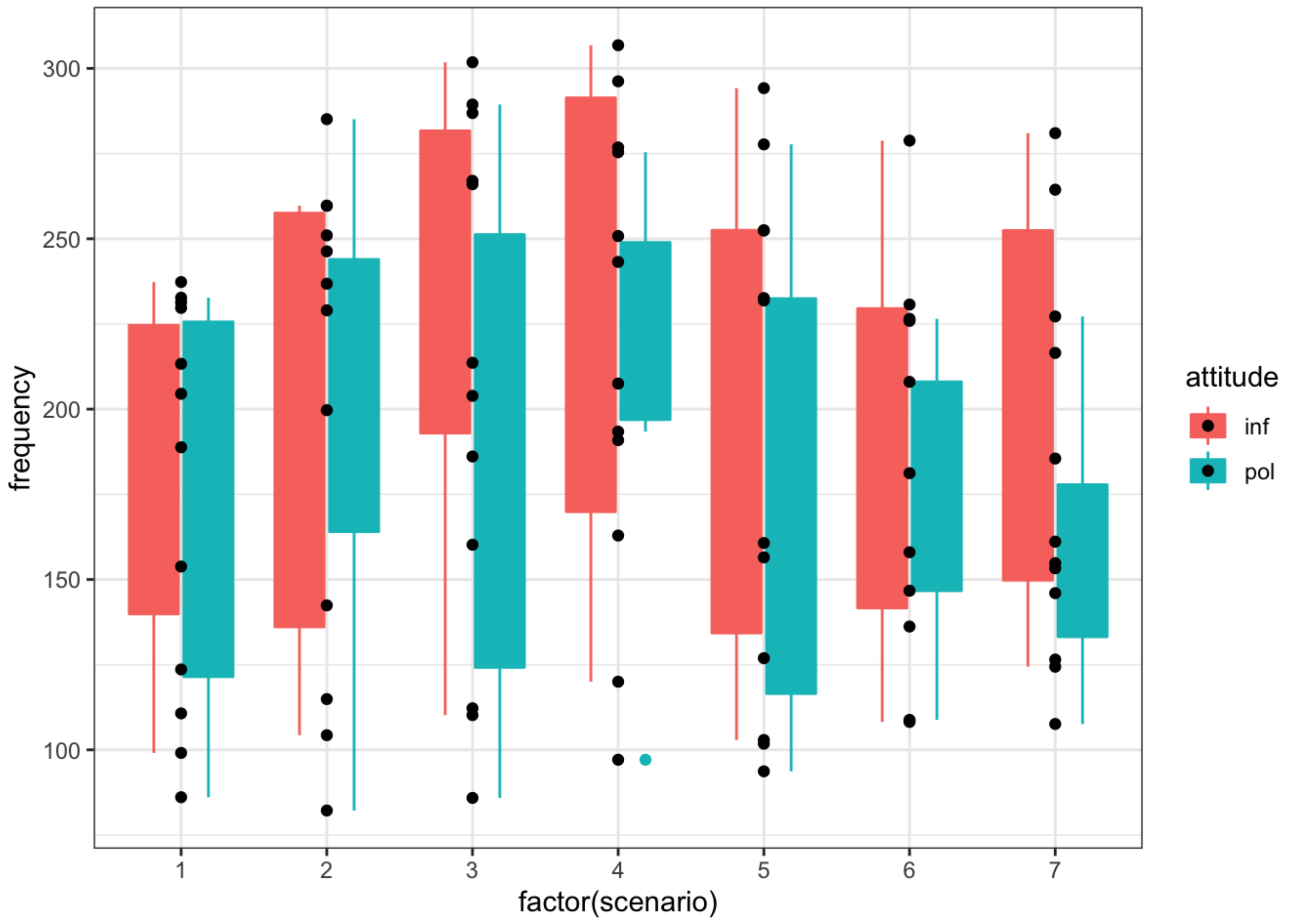
```
library(lattice)
dotplot(ranef(m2, condVar=T))
```

```
## $subject
```



Ok, we have seen, that the subjective variation due to repeated measurements is real and soo significant in our model, that it screwed up the real effect of the attitude. Similarly, there might be other kinds of variation which would continue to overwrite a real (fixed) effects of the attitude, if not removed from the fixed effects via random effects. For instance, if we visualize different scenarios, we see 6 repeated measurements for each scenario per attitude and they also vary a lot (big boxplots)

```
ggplot(politeness, aes(factor(scenario), frequency, fill = attitude))+
  geom_boxplot(aes(color = attitude))+
  geom_point()+
  theme_bw()
```



Thus, we first can check whether this random effect (1|scenario) is important for the model:

```
m1 <- lmer(frequency ~ attitude + (1|subject), data=politeness, REML = F)
m2 <- lmer(frequency ~ attitude + (1|subject) + (1|scenario), data=politeness, REML = F)
anova(m2, m1)
```

```
## Data: politeness
## Models:
## m1: frequency ~ attitude + (1 | subject)
## m2: frequency ~ attitude + (1 | subject) + (1 | scenario)
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m1    4 826.51 836.18 -409.25   818.51
## m2    5 817.04 829.13 -403.52   807.04 11.466  1  0.0007089 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

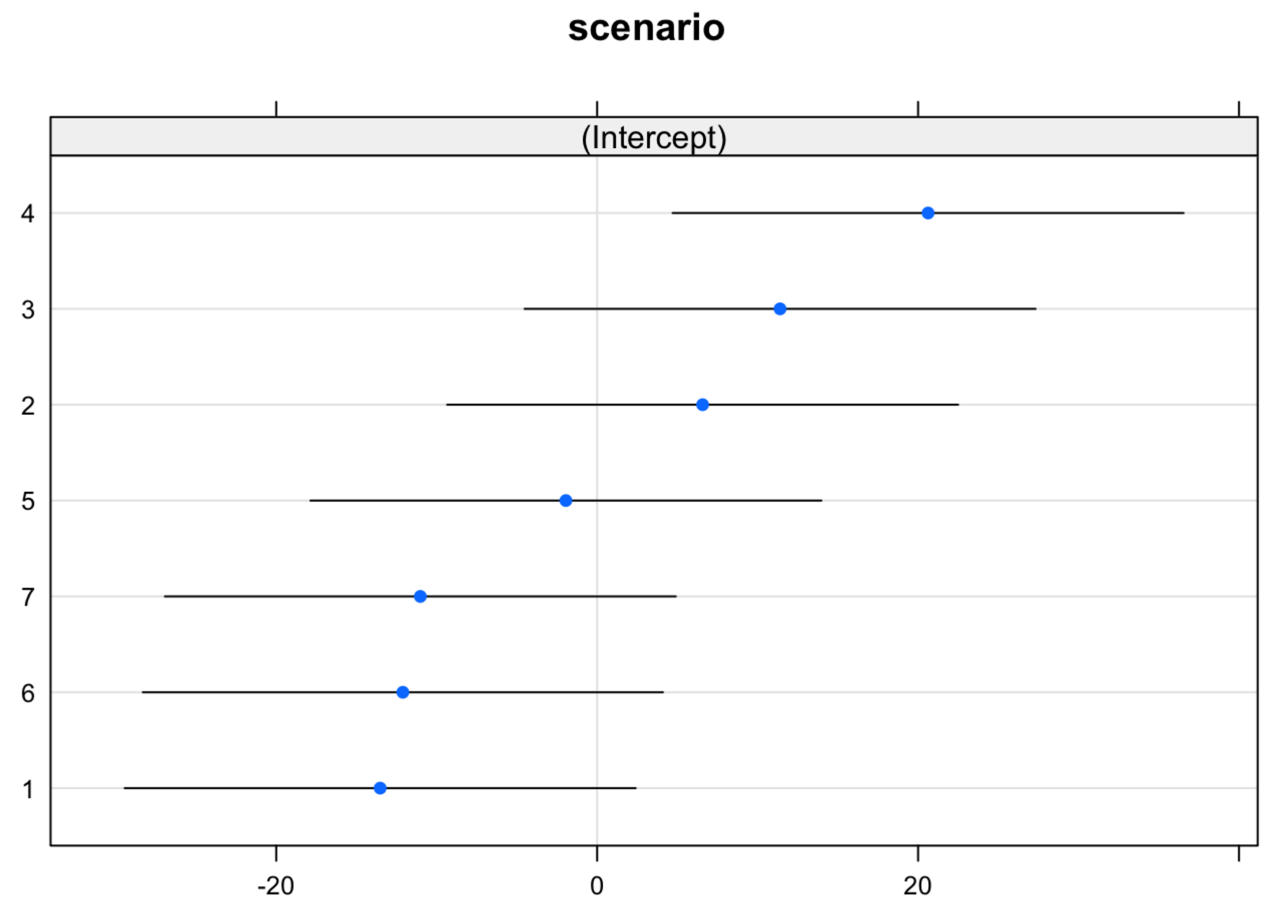
It looks like it is! And then we can have a look at the new 7 intercepts (for 7 scenarios):

```
coef(m2)
```

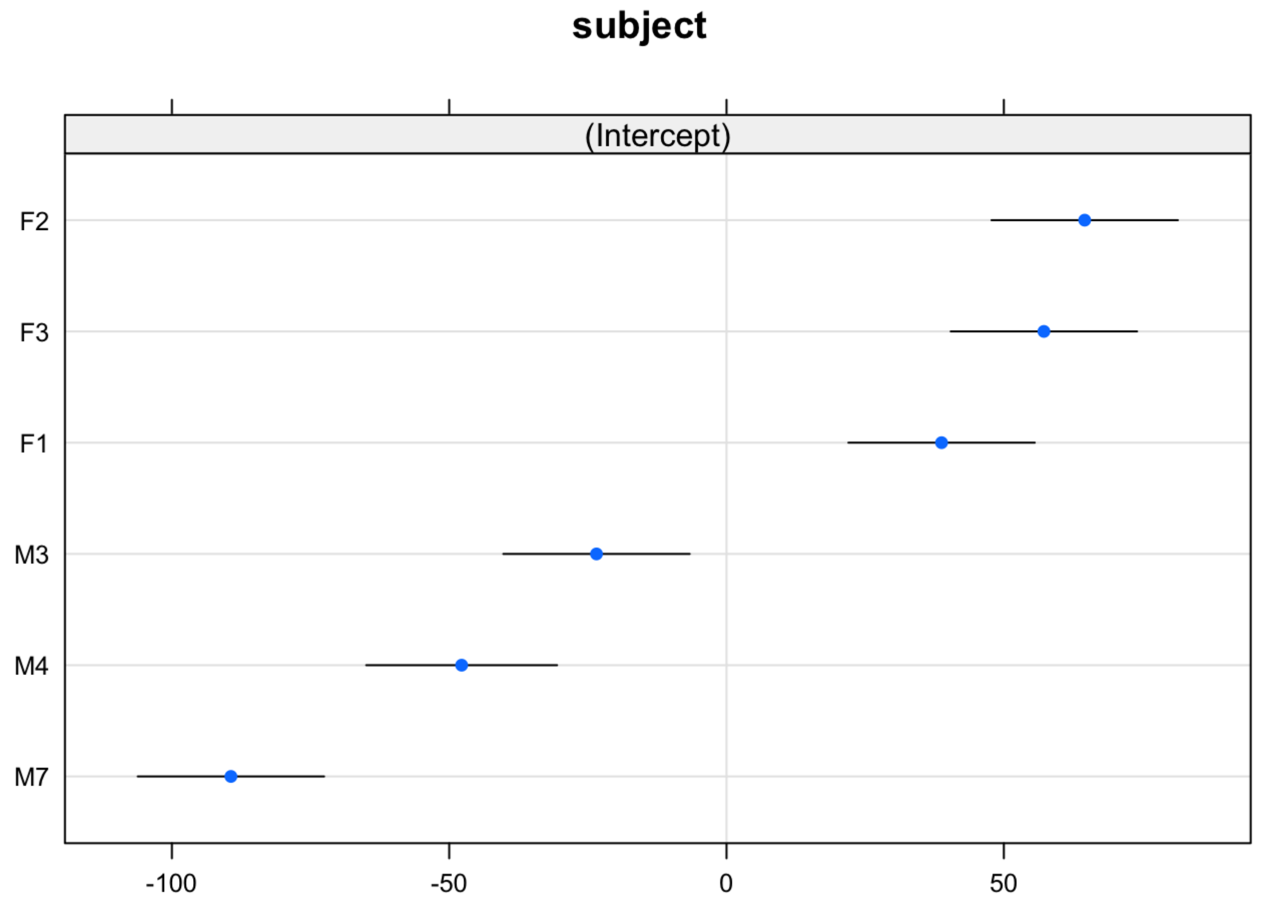
```
## $scenario
##   (Intercept) attitudepol
## 1    189.0660   -19.69216
## 2    209.1613   -19.69216
## 3    213.9877   -19.69216
## 4    223.2120   -19.69216
## 5    200.6399   -19.69216
## 6    190.4804   -19.69216
## 7    191.5695   -19.69216
##
## $subject
##    (Intercept) attitudepol
## F1    241.3580   -19.69216
## F2    267.1593   -19.69216
## F3    259.8087   -19.69216
## M3    179.1415   -19.69216
## M4    154.8292   -19.69216
## M7    113.2320   -19.69216
##
## attr(,"class")
## [1] "coef.mer"
```

```
dotplot(ranef(m2, condVar=T))
```

```
## $scenario
```



```
##
## $subject
```



If we take out this variance from the model (via random effects), we'll again make the fixed effects ("*attitude*" in our case) more realistic! This is why the p-value for the fixed effect is:

- not-significant (p = 0.2) if variation contained in both, *subject* and *scenario*, is included into the *attitude*
- significant (p = 0.003) when variation of *subjects* is excluded from *attitude*
- and even more significant (p = 0.0007) when variation of both *subjects* and *scenario* is excluded from the *attitude*:

```
summary(m2)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: frequency ~ attitude + (1 | subject) + (1 | scenario)
##    Data: politeness
##
##      AIC      BIC   logLik deviance df.resid
##    817.0    829.1   -403.5    807.0       78
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2127 -0.5906 -0.0598  0.5675  3.4584
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  scenario (Intercept)  216.8   14.72
##  subject  (Intercept) 3367.7   58.03
##  Residual              637.0   25.24
## Number of obs: 83, groups:  scenario, 7; subject, 6
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  202.588     24.646   6.794   8.220 9.03e-05 ***
## attitudepol  -19.692      5.546  70.994  -3.551 0.000686 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr)
## attitudepol -0.111
```

The variation between scenarios is 15 times smaller (3367.7 / 216.8) then the variation between subjects, which is another inference we can draw from our model, but scenarios are still significant, and we better account for them in our model!

## Why not just use everything as a fixed effects?

If we use a *subject* and *scenario* variables as fixed effects in a form of categorical variables, the categories (i.e. subject 1) would be estimated only from their own data. It **enormously increases the number of Intercepts and slopes** while **severally decreases a sample size**, because every category will be treated separately and other categories wouldn’t have any influence on it. Besides, it also increases chances of a Type I Error (false rejection of the null hypothesis) by carrying out multiple comparisons. Moreover, the data in particular categories are more similar to each other than the data from different categories - they are correlated.

In contrast, if included as random effects (in MEM), the categories (now - *subjects* or/and *scenarios*) will be **estimated from all the data, not only from the data that belongs to, i.e., subject 1**, which increases the sample size. Besides, MEM accounts for the correlations between data coming from the *subjects* and *scenarios*. MEM produces fewer parameters and avoids problems with multiple comparisons that we would encounter while using separate categories (separate regressions for each category).

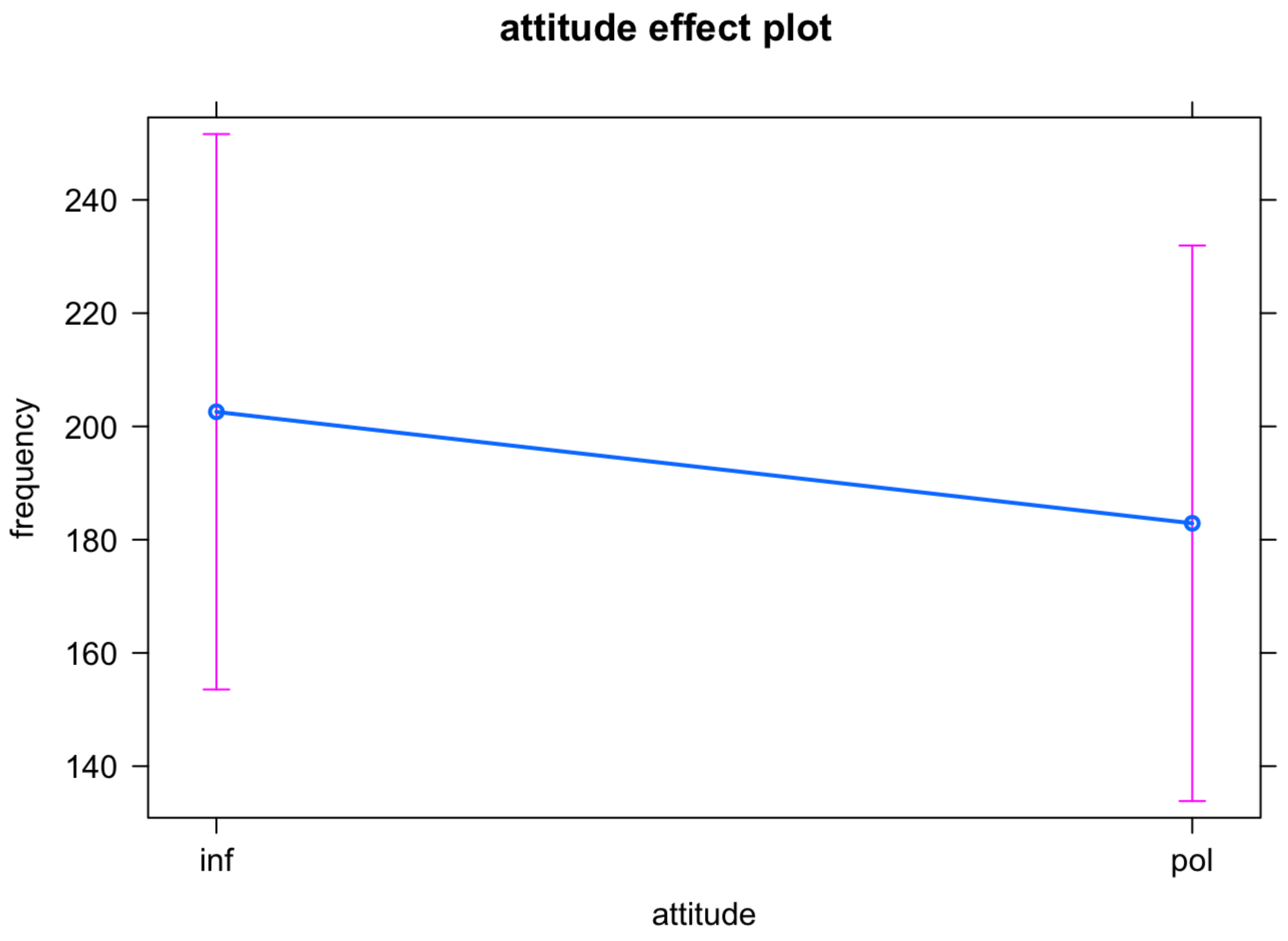
The estimates of MEM will fall in between a model that ignores the subject variation completely, and the one that includes individual subjects as predictors.

**However, as mentioned before, whether the variable is fixed or random heavily depends on your research question!**

## Visualize model results

One picture is worth a thousand words!

```
plot(allEffects(m2))
```



## How to report the results

The `report()` function from the `report` package provides a useful interpretation of the model:

```
report(m2) %>% text_long()
```

```
## We fitted a linear mixed model (estimated using ML and nloptwrap optimizer) to predict frequency with a
##
##   - The effect of attitudepol is negative and can be considered as small and significant (beta = -19.69
```

We fitted a linear mixed model (estimated using ML and nloptwrap optimizer) to predict frequency with attitude (formula = frequency ~ attitude). The model included subject and scenario as random effects (formula = ~1 | subject + ~1 | scenario). Standardized parameters were obtained by fitting the model on a standardized version of the dataset. Effect sizes were labelled following Funder's (2019) recommendations.The model's total explanatory power is substantial (conditional R2 = 0.85) and the part related to the fixed effects alone (marginal R2) is of 0.02. The model's intercept, corresponding to frequency = 0, attitude = inf, subject = F1 and scenario = 0, is at 192.74 (SE = 24.49, 95% CI [144.74, 240.75], p < .001). Within this model:

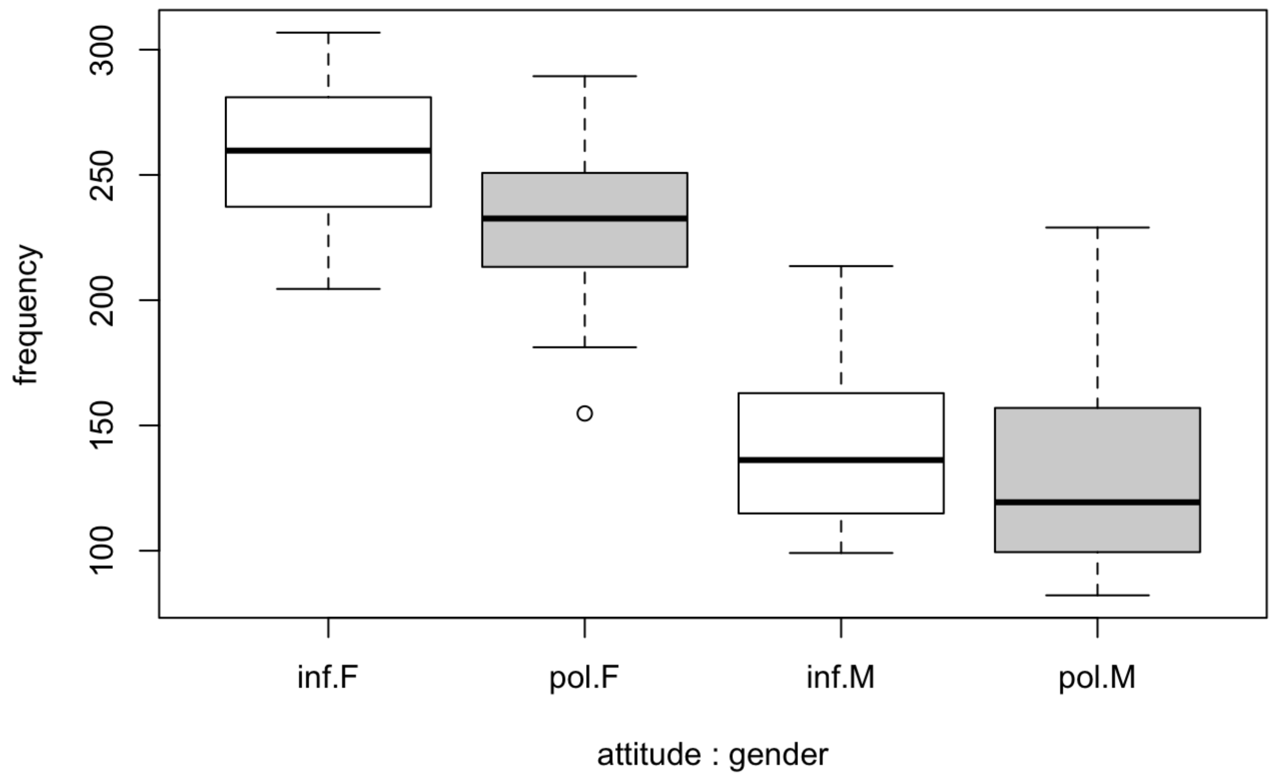
- The effect of attitude1 is positive and can be considered as very small and significant (beta = 9.85, SE = 2.77, 95% CI [4.41, 15.28], std. beta = 0.15, p < .001).

The interpretation above shows high explanatory power of the model (85%), where the fixed effects (attitude), why still significant, explain only 2%.

## Multiple MEM: adding another predictor

The visualization below reveals that there are clear differences between man and woman (gender is the only remaining variable in our dataset). Thus, we might add gender to try to explain more of the variance in our data:

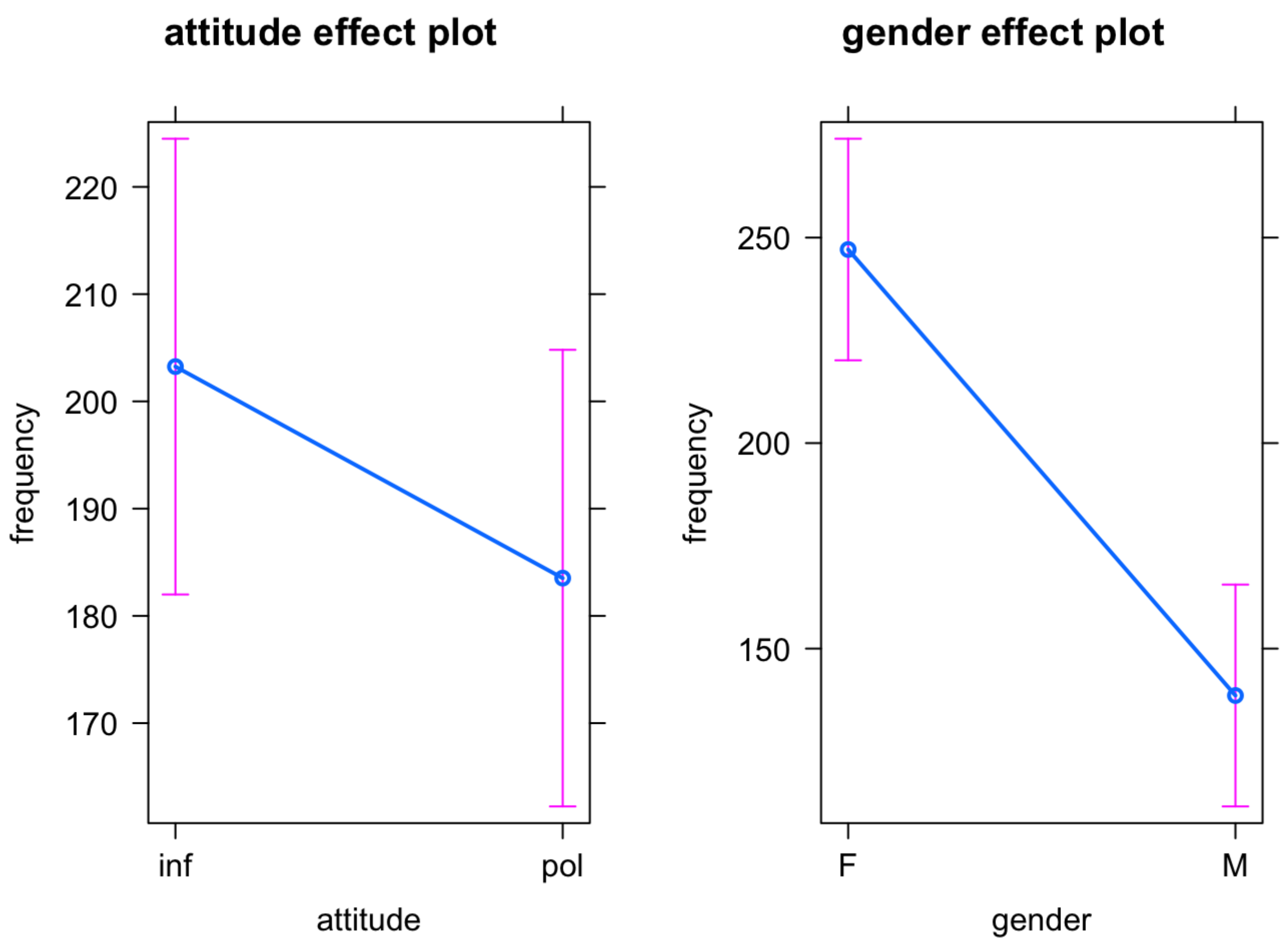
```
boxplot(frequency ~ attitude*gender,
col=c("white","lightgray"),politeness)
```



Moreover, the intercept of 202.59 from the model without gender ("m2" above) completely misses this difference, but rather averages the effects of gender. 202.59 is actually the average of our data for only *informal* condition, because *informal* is alphabetically before *polite*.

Ignoring the *gender* variable reminds me on measuring average length of mans and woman hair, which makes no sense at all. Thus, let's include gender as an additional fixed effect (because it has only two fixed categories):

```
m2 <- lmer(frequency ~ attitude + gender + (1|subject) + (1|scenario),
           data=politeness, REML = F)
plot(allEffects(m2))
```



```
report(m2)
```

```
## We fitted a linear mixed model (estimated using ML and nloptwrap optimizer) to predict frequency with a
##
##   - The effect of attitudepol is negative and can be considered as small and significant (beta = -19.72
##   - The effect of genderM is negative and can be considered as very large and significant (beta = -108.
```

The overall model predicting frequency (formula = frequency ~ attitude + gender + (1 | subject) + (1 | scenario)) has an total explanatory power (conditional R2) of 85.79%, in which the fixed effects explain 67.42% of the variance (marginal R2). The model's intercept is at 256.85 (SE = 16.12, 95% CI [226.08, 287.61]). Within this model:

- The effect of attitudepol is significant (beta = -19.72, SE = 5.58, 95% CI [-30.74, -8.70], t(70) = -3.53, p < .001) and can be considered as small (std. beta = -0.30, std. SE = 0.09).
- The effect of genderM is significant (beta = -108.52, SE = 21.01, 95% CI [-149.34, -67.68], t(4) = -5.16, p < .01) and can be considered as large (std. beta = -1.66, std. SE = 0.32).

**Wow, the explanatory power of fixed effects increased dramatically: from 1.97% to 67.42%.** And oppositely to the variation of the subject and scenario as random effects, which was hidden in the fixed effect of attitude, the variation between sexes as a fixed effect was hidden in the random effects of subject and scenario. However, the explanatory power of the model decreased a bit, which makes me wonder, whether the model with *gender* in it is better then without:

```
m1 <- lmer(frequency ~ attitude + (1|subject) + (1|scenario), data=politeness, REML = F)
m2 <- lmer(frequency ~ attitude + gender + (1|subject) + (1|scenario), data=politeness, REML = F)
anova(m2, m1)
```

```
## Data: politeness
## Models:
## m1: frequency ~ attitude + (1 | subject) + (1 | scenario)
## m2: frequency ~ attitude + gender + (1 | subject) + (1 | scenario)
##    npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## m1    5 817.04 829.13 -403.52   807.04
## m2    6 807.10 821.61 -397.55   795.10 11.938  1    0.00055 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It actually is, which means that the fixed effect of *gender* is significant. The model summary below corroborates with this conclusion:

```
tab_model(m2, p.style = "both")
```

| | frequency |
|---|---|

| *Predictors* | *Estimates* | *CI* | *p* |
|---|---|---|---|
| (Intercept) | 256.85 $^{***}$ | 229.75 – 283.95 | **<0.001** |
| attitude [pol] | -19.72 $^{***}$ | -30.59 – -8.85 | **<0.001** |
| gender [M] | -108.52 $^{***}$ | -142.96 – -74.08 | **<0.001** |
| **Random Effects** | | | |
| $\sigma^2$ | 637.40 | | |
| $\tau_{00\ \text{scenario}}$ | 205.23 | | |
| $\tau_{00\ \text{subject}}$ | 416.98 | | |
| ICC | 0.49 | | |
| N $_{\text{subject}}$ | 6 | | |
| N $_{\text{scenario}}$ | 7 | | |
| Observations | 83 | | |
| Marginal $R^2$ / Conditional $R^2$ | 0.709 / 0.853 | | |

- *p<0.05   ** p<0.01   *** p<0.001*

## Explanatory vs. predictive power of the model

But how can a model with a lower explanatory power be significantly better? Well, the *explanatory power* shows the *quality of fit*, while AIC, being an indicator of the *predictive power*, shows *the quality of the model*. Thus, the model can be *overfit* to a particular dataset and have very high explanatory power, but be miserable at prediction (extrapolation or interpolation) if provided with new data.

Different intercepts for different subjects and different intercepts for different scenarios solved the non-independence problem in out data. This also allowed us to exclude the inside-subject and inside-scenarios variation from the fixed effects, which was caused by multiple responses per subject and per scenario. The residuals (645.9) is our **error term** which remains unexplained by the model. The coefficient “attitudepol” is the slope for the categorical effect of politeness. Minus 19.695 means that to go from “informal” to “polite”, you have to lower the voice by -19.695 Hz.

The order of the factors in the model is alphabetical, thus, “informal” comes before “polite”, so the slope represents the change from “inf” to “pol”. Changing the order from “pol” to “inf” would only change the sign of the coefficient -19.695 to positive. Standard errors, significance etc. would remain the same.

We added *gender* as a **fixed effect** because the relationship between sex and voice pitch is clear and predictable (i.e., we expect females to have higher voice). This is different from the random effects subject and scenario, where the relationship between these and voice is much more unpredictable and therefore *random*.

Compared to the model without gender, the variation that’s associated with the random effect *subject* dropped considerably from 3367.7 to 615 and the variation of the *scenario* dropped just a little, from 216.8 to 205. Thus, as mentioned above, gender variation was indeed hidden mostly within the *subject*, and a bit in the *scenario* variables. Adding the effect of gender allowed us to explain more variation from the model via transforming some random variation to the fixed one.

The Intercept in the gender-consisting model (257) makes now much more sense. If you compare it to the very next plot above, where we split the attitude data by gender, you’ll find that the Intercept corresponds to the median of females, and the males coefficient (genderM) of 257-109 = 148 is very much where the males subplot is. The coefficient for the fixed effect of attitude remained almost the same.

All right, what can we do, if we included all the predictors into the model? Well, we can add an interaction term and do a post-hoc analysis of factors, which compares the estimated means of levels to each other.

## Interaction

In case, you did not know, a star (*) between predictors adds the interaction term into the model.

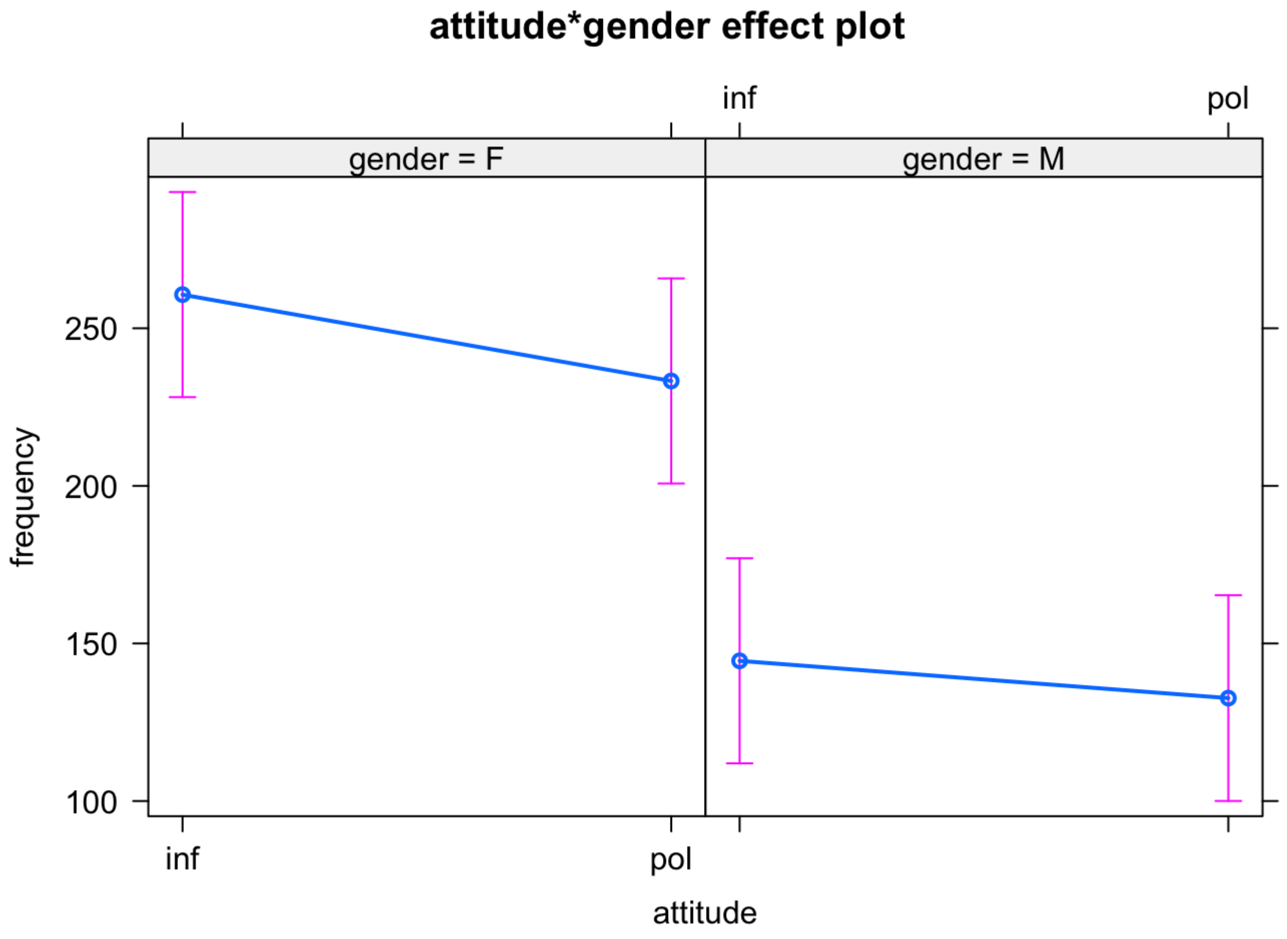
```
m2 <- lmer(frequency ~ attitude * gender + (1|subject) + (1|scenario), data=politeness)
summary(m2)
```

```
## Linear mixed model fit by REML. t-tests use Satterthwaite's method [
## lmerModLmerTest]
## Formula: frequency ~ attitude * gender + (1 | subject) + (1 | scenario)
##    Data: politeness
## 
## REML criterion at convergence: 766.8
## 
## Scaled residuals: 
##     Min      1Q  Median      3Q     Max 
## -2.1191 -0.5604 -0.0768  0.5111  3.3352 
## 
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  scenario (Intercept) 218.3    14.77   
##  subject  (Intercept) 617.1    24.84   
##  Residual             637.4    25.25   
## Number of obs: 83, groups:  scenario, 7; subject, 6
## 
## Fixed effects:
##                     Estimate Std. Error       df t value Pr(>|t|)    
## (Intercept)          260.686     16.348    5.737  15.946  5.7e-06 ***
## attitudepol          -27.400      7.791   69.017  -3.517 0.000777 ***
## genderM             -116.195     21.728    4.566  -5.348 0.004023 ** 
## attitudepol:genderM   15.572     11.095   69.056   1.403 0.164958    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Correlation of Fixed Effects:
##             (Intr) atttdp gendrM
## attitudepol -0.238              
## genderM     -0.665  0.179       
## atttdpl:gnM  0.167 -0.702 -0.252
```

```
report(m2, CI = 95)
```

```
## We fitted a linear mixed model (estimated using REML and nloptwrap optimizer) to predict frequency with
## 
##   - The effect of attitudepol is negative and can be considered as medium and significant (beta = -27.4
##   - The effect of genderM is negative and can be considered as very large and significant (beta = -116.
##   - The effect of attitudepol:genderM is positive and can be considered as small and not significant (b
```

```
plot(allEffects(m2))
```



Scroll to the right to read the whole interpretation.

```
report(m2) %>% table_long()
```

```
## Parameter           | Coefficient |    SE |  CI_low | CI_high |     t | df_error |    p | Std_Coefficie
## ----------------------------------------------------------------------------------------------------------
## (Intercept)         |      260.69 | 16.35 |  228.65 |  292.73 | 15.95 |       76 | 0.00 |            1.
## attitudepol         |      -27.40 |  7.79 |  -42.67 |  -12.13 | -3.52 |       76 | 0.00 |           -0.
## genderM             |     -116.20 | 21.73 | -158.78 |  -73.61 | -5.35 |       76 | 0.00 |           -1.
## attitudepol:genderM |       15.57 | 11.10 |   -6.17 |   37.32 |  1.40 |       76 | 0.16 |            0.
##                     |             |       |         |         |       |          |      |
## AIC                 |             |       |         |         |       |          |      |
## BIC                 |             |       |         |         |       |          |      |
## R2 (conditional)    |             |       |         |         |       |          |      |
## R2 (marginal)       |             |       |         |         |       |          |      |
## ICC                 |             |       |         |         |       |          |      |
## RMSE                |             |       |         |         |       |          |      |
```

## Post-hoc / Contrast Analysis

```
emmeans(m2, pairwise ~ gender * attitude, adjust = "bonferroni")$contrasts %>%
  tidy() %>%
  mutate_if(is.numeric, ~round(., 4)) %>%
  kable()
```

| level1 | level2 | estimate | std.error | df | statistic | p.value |
|---|---|---|---|---|---|---|
| F,inf | M,inf | 116.1952 | 21.7283 | 4.5585 | 5.3476 | 0.0243 |
| F,inf | F,pol | 27.4000 | 7.7913 | 69.0004 | 3.5168 | 0.0047 |
| F,inf | M,pol | 128.0232 | 21.7680 | 4.5911 | 5.8813 | 0.0161 |
| M,inf | F,pol | -88.7952 | 21.7283 | 4.5585 | -4.0866 | 0.0689 |
| M,inf | M,pol | 11.8280 | 7.9013 | 69.0766 | 1.4970 | 0.8337 |
| F,pol | M,pol | 100.6232 | 21.7680 | 4.5911 | 4.6225 | 0.0424 |

## Pick up the final best model

```
m0 <- lmer(frequency ~ attitude * gender + (1|subject) + (1|scenario), data=politeness, REML = F)
m1 <- lmer(frequency ~ attitude + gender + (1|subject) + (1|scenario), data=politeness, REML = F)
m2 <- lmer(frequency ~ attitude + gender + (1|subject),  data=politeness, REML = F)
m3 <- lmer(frequency ~ attitude + gender + (1|scenario), data=politeness, REML = F)
m4 <- lmer(frequency ~ attitude + (1|subject) + (1|scenario), data=politeness, REML = F)
m5 <- lmer(frequency ~ gender   + (1|subject) + (1|scenario), data=politeness, REML = F)
m6 <- lmer(frequency ~ attitude + (1|subject), data=politeness, REML = F)
m7 <- lmer(frequency ~ attitude + (1|scenario), data=politeness, REML = F)
m8 <- lmer(frequency ~ gender   + (1|subject),  data=politeness, REML = F)
m9 <- lmer(frequency ~ gender   + (1|scenario), data=politeness, REML = F)
m10 <- lm(frequency ~ attitude, data=politeness, REML = F)
m11 <- lm(frequency ~ gender,   data=politeness, REML = F)

anova(m0, m1, m2, m3, m4, m5, m6, m7, m8, m9, m10, m11)
```

```
## Data: politeness
## Models:
## m10: frequency ~ attitude
## m11: frequency ~ gender
## m6: frequency ~ attitude + (1 | subject)
## m7: frequency ~ attitude + (1 | scenario)
## m8: frequency ~ gender + (1 | subject)
## m9: frequency ~ gender + (1 | scenario)
## m2: frequency ~ attitude + gender + (1 | subject)
## m3: frequency ~ attitude + gender + (1 | scenario)
## m4: frequency ~ attitude + (1 | subject) + (1 | scenario)
## m5: frequency ~ gender + (1 | subject) + (1 | scenario)
## m1: frequency ~ attitude + gender + (1 | subject) + (1 | scenario)
## m0: frequency ~ attitude * gender + (1 | subject) + (1 | scenario)
##     npar    AIC    BIC  logLik deviance    Chisq Df Pr(>Chisq)    
## m10    3 933.22 940.48 -463.61   927.22                          
## m11    3 838.08 845.33 -416.04   832.08  95.1444  0  < 2.2e-16 ***
## m6     4 826.51 836.18 -409.25   818.51  13.5702  1  0.0002298 ***
## m7     4 935.22 944.90 -463.61   927.22   0.0000  0  1.0000000    
## m8     4 823.13 832.80 -407.56   815.13 112.0936  0  < 2.2e-16 ***
## m9     4 837.19 846.87 -414.60   829.19   0.0000  0  1.0000000    
## m2     5 816.34 828.43 -403.17   806.34  22.8569  1  1.745e-06 ***
## m3     5 832.05 844.14 -411.02   822.05   0.0000  0  1.0000000    
## m4     5 817.04 829.13 -403.52   807.04  15.0062  0  < 2.2e-16 ***
## m5     5 816.72 828.81 -403.36   806.72   0.3202  0  < 2.2e-16 ***
## m1     6 807.10 821.61 -397.55   795.10  11.6178  1  0.0006532 ***
## m0     7 807.11 824.04 -396.55   793.11   1.9963  1  0.1576796    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show that “m1” is the best model, following by a non-significantly different “m0”. The last shows that the interaction in our model is redundant.

## A final interpretation of our best model

“We used R (R Core Team, 2012) and lme4 (Bates, Maechler & Bolker, 2012) to perform a linear mixed effects analysis of the relationship between pitch and politeness. As fixed effects, we entered politeness and gender (without interaction term) into the model. As random effects, we had intercepts for subjects and items, as well as by-subject and by-item random slopes for the effect of politeness (in the next article). Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. P-values were obtained by likelihood ratio tests of the full model with the effect in question against the model without the effect in question.”

**Despite the fact that our model already became very complex, we are only one step away from unleashing the real power of MEMs - random slopes. Thus, hang on! It’s worth it!**

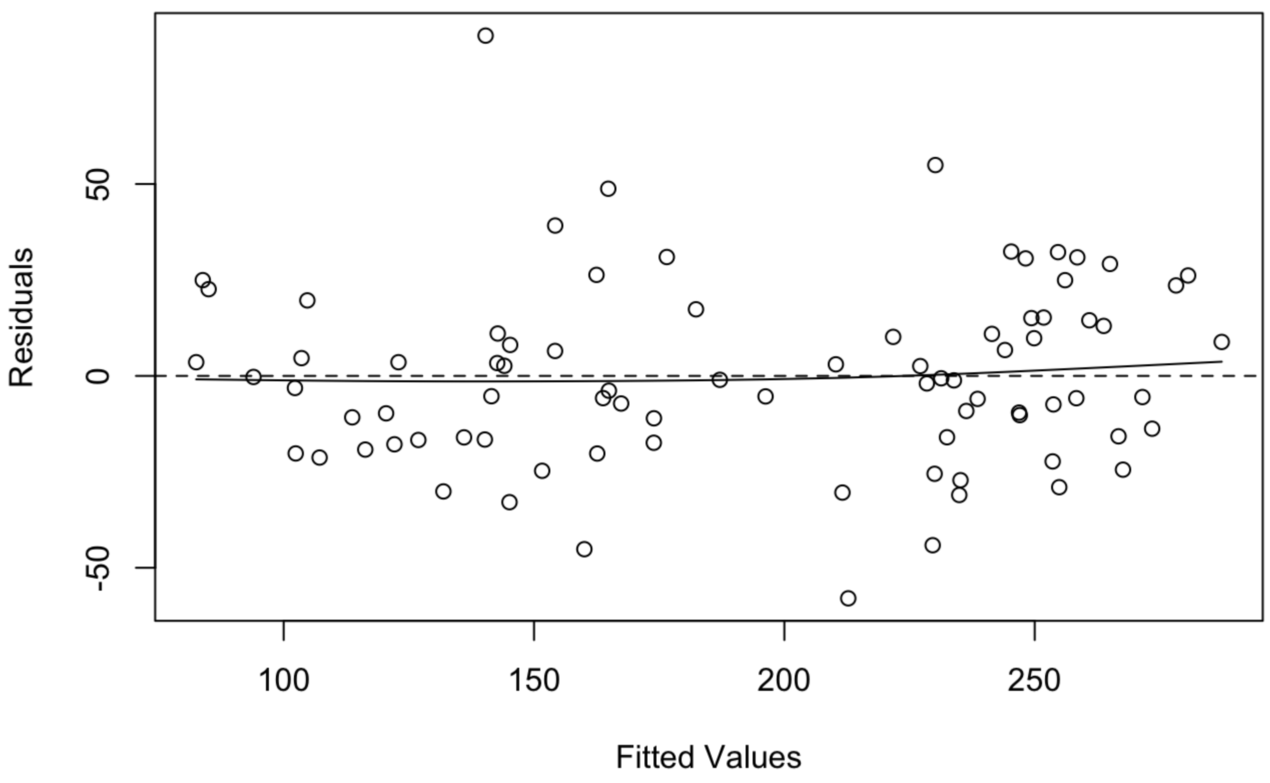
# When NOT to use Mixed Effects Models

- in simple cases where the data is perfectly balanced and the assumptions of repeated measures ANOVA hold
- in cases where simple linear regression have a better quality (lower AIC)
- if you are interested in the influence of a categorical variable, which can potentially be a random effect, on the response variable

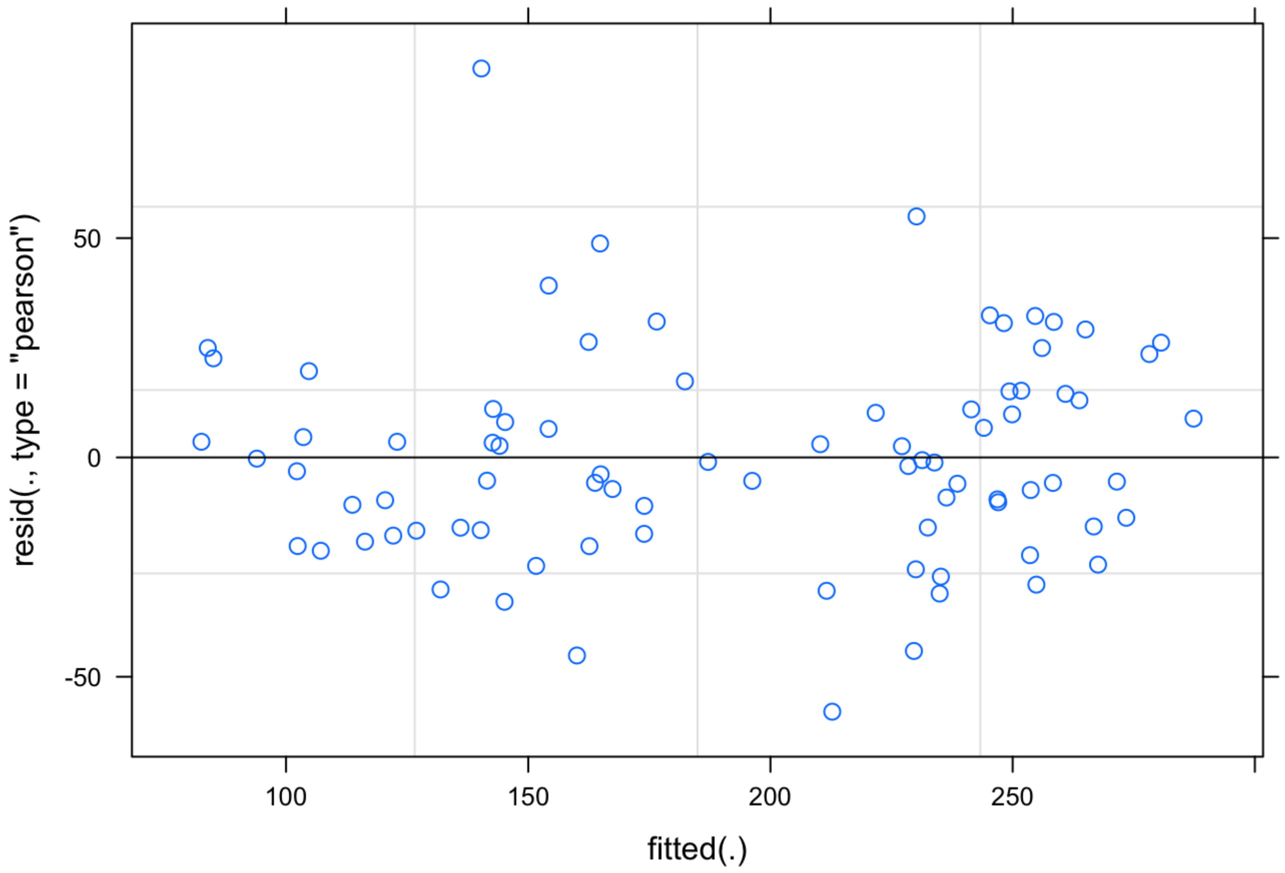
## Assumptions

It is important to check assumptions. These assumptions are the same as those of a linear model, except of the *independence of observations* which we have relaxed by using MEM. So, there are 5 assumptions left: linearity, normality, homoscedasticity, collinearity and influential data points. If assumptions are violated, the log-transformation of the data often helps.

```
# check for linearity, normality and homoscedasticity
plot(fitted(m1), residuals(m1), xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, lty = 2)
lines(smooth.spline(fitted(m1), residuals(m1)))
```



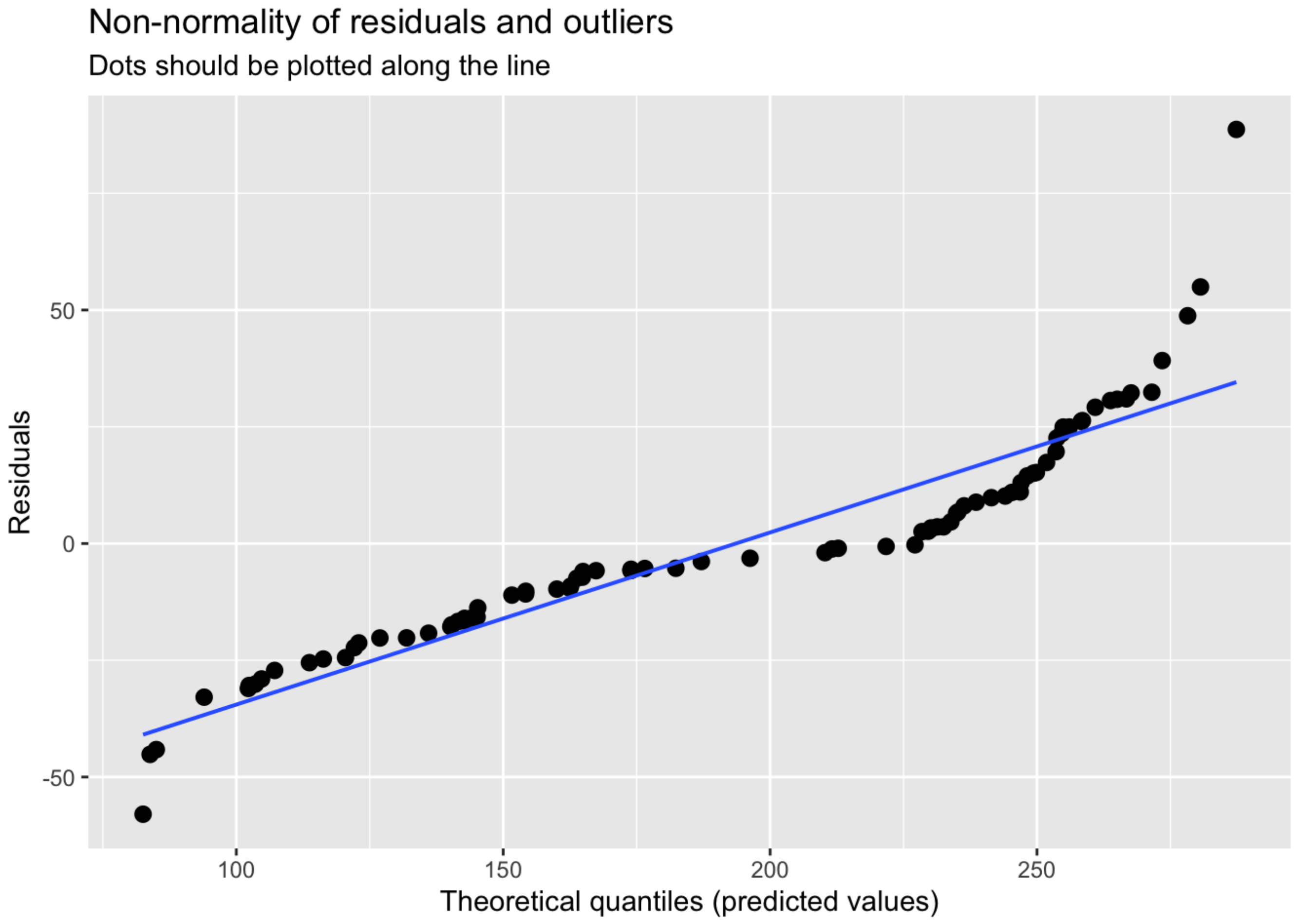
```
# or simply
plot(m1)
```



```
# or a fancy one ;)
plot_model(m1, type = "diag")
```

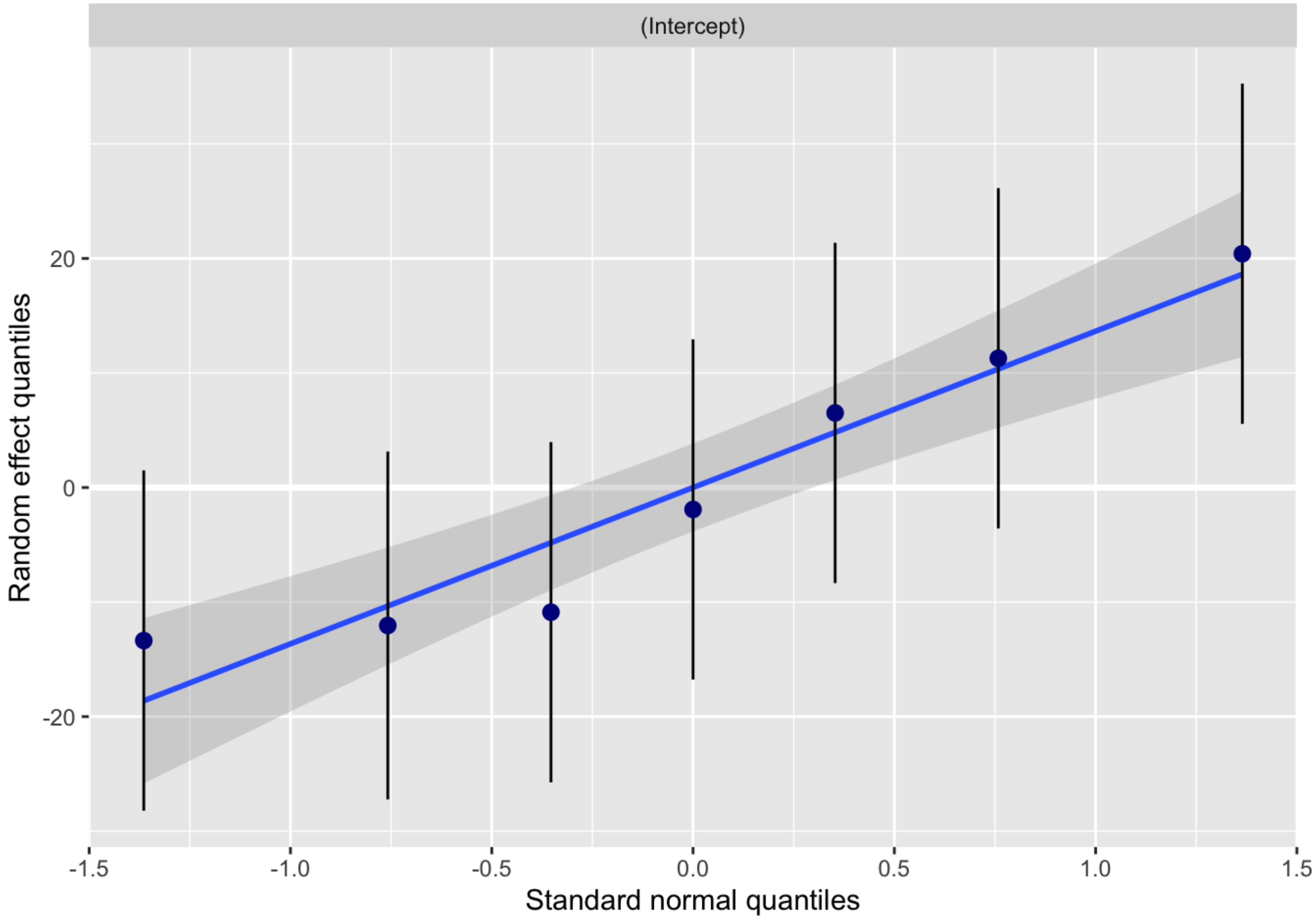
```
## [[1]]
```

```
## `geom_smooth()` using formula 'y ~ x'
```



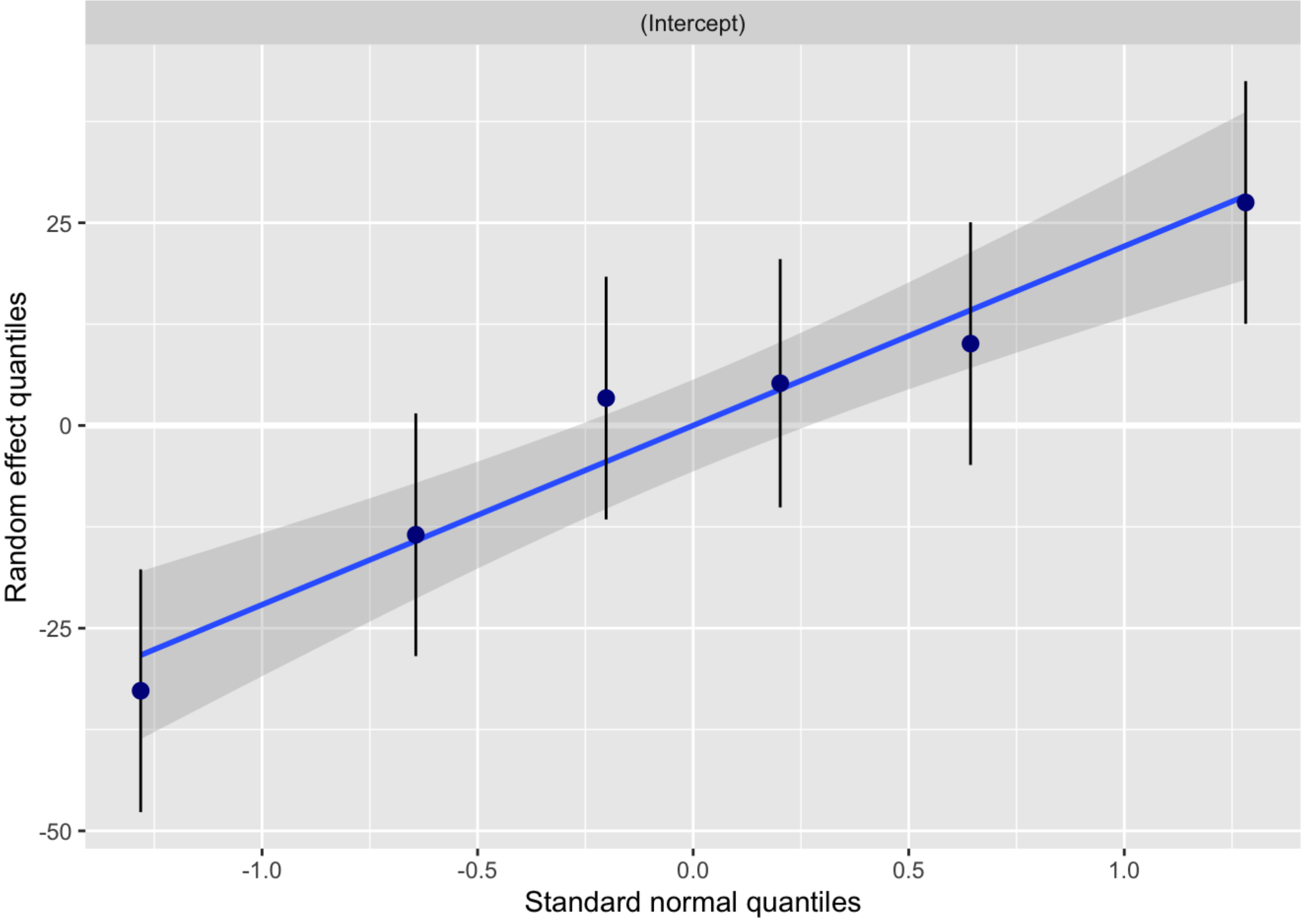
```
##
## [[2]]
## [[2]]$scenario
```

```
## `geom_smooth()` using formula 'y ~ x'
```

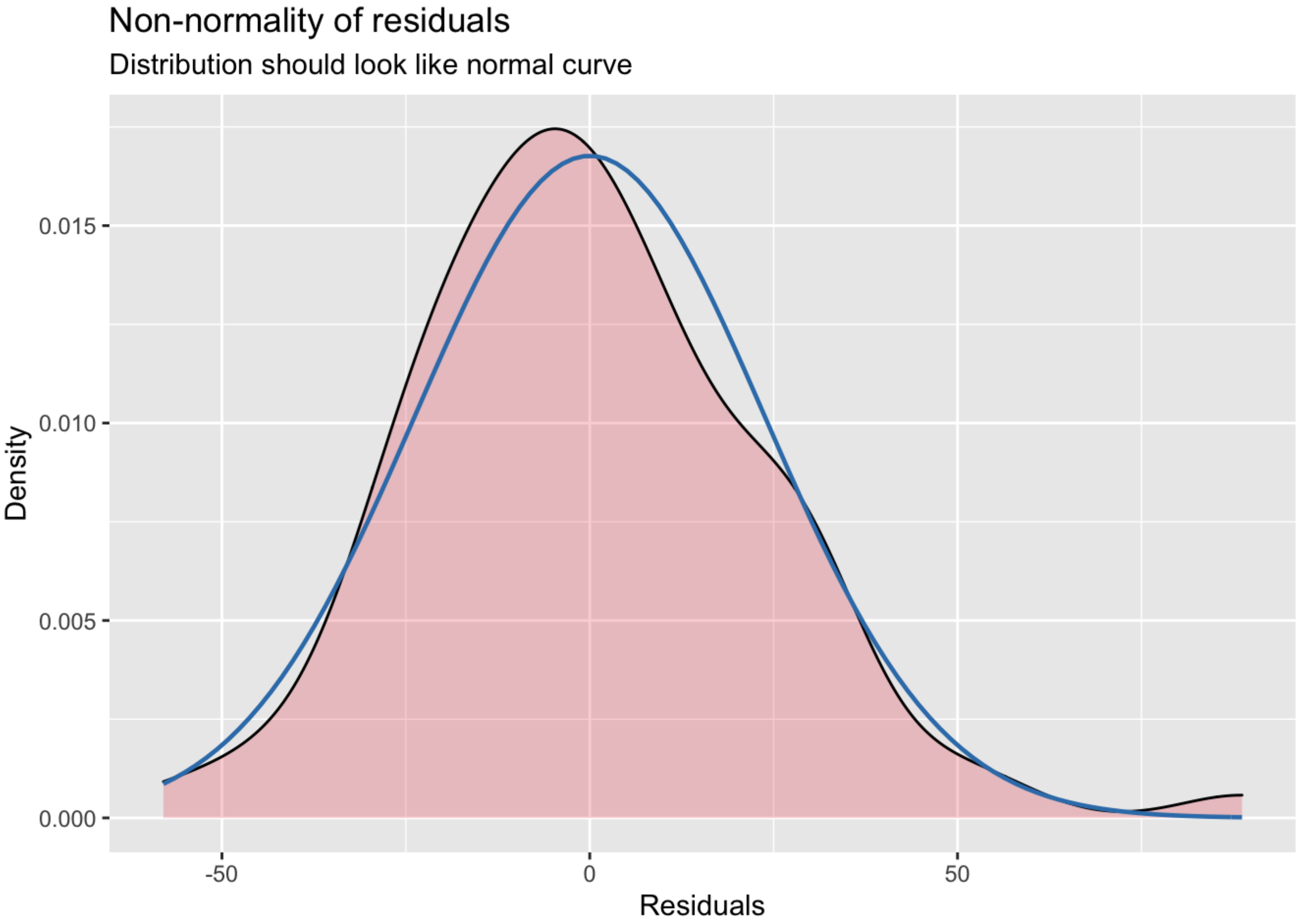


```
##
## [[2]]$subject
```

```
## `geom_smooth()` using formula 'y ~ x'
```

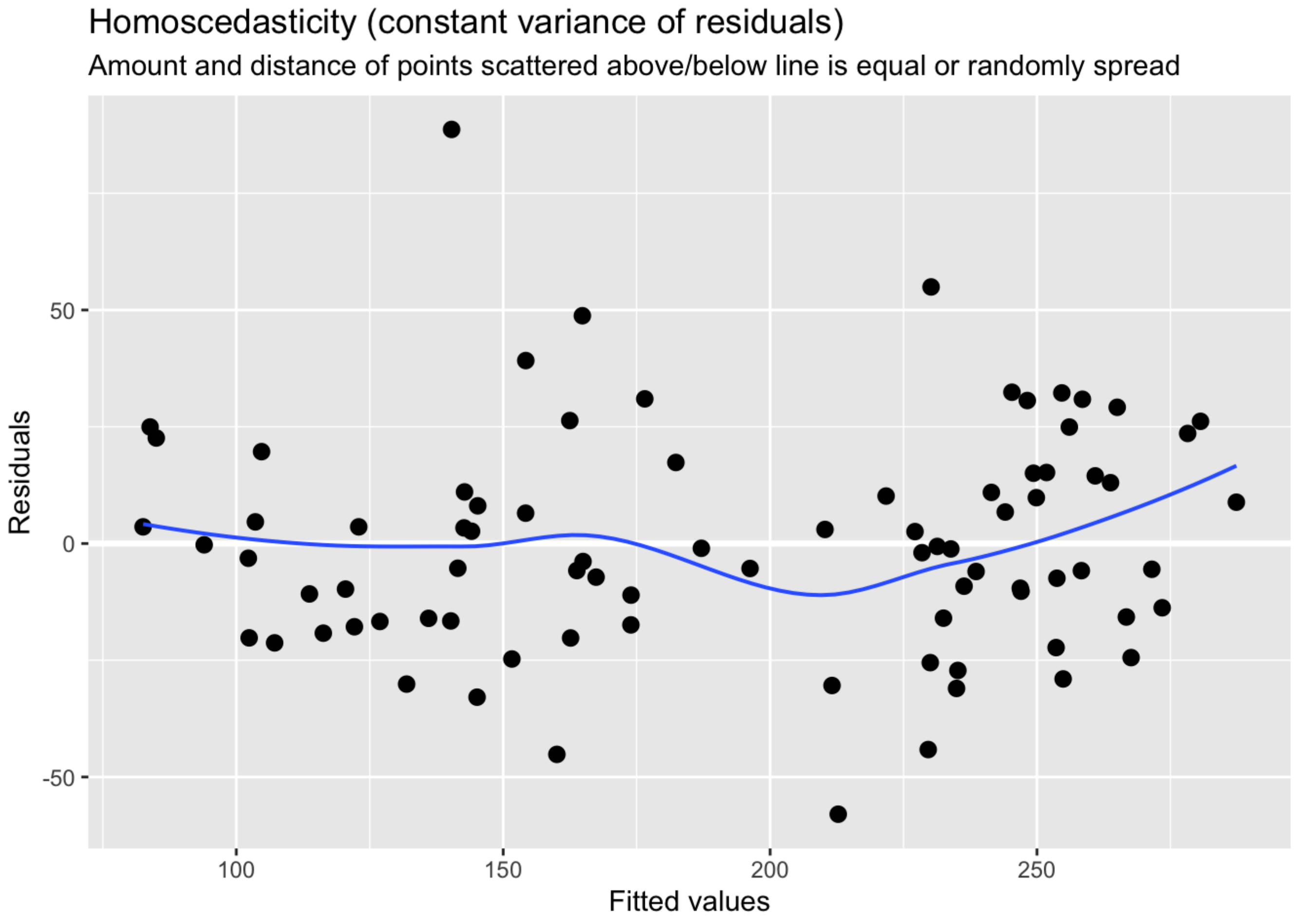


```
##
##
## [[3]]
```



```
##
## [[4]]
```

```
## `geom_smooth()` using formula 'y ~ x'
```

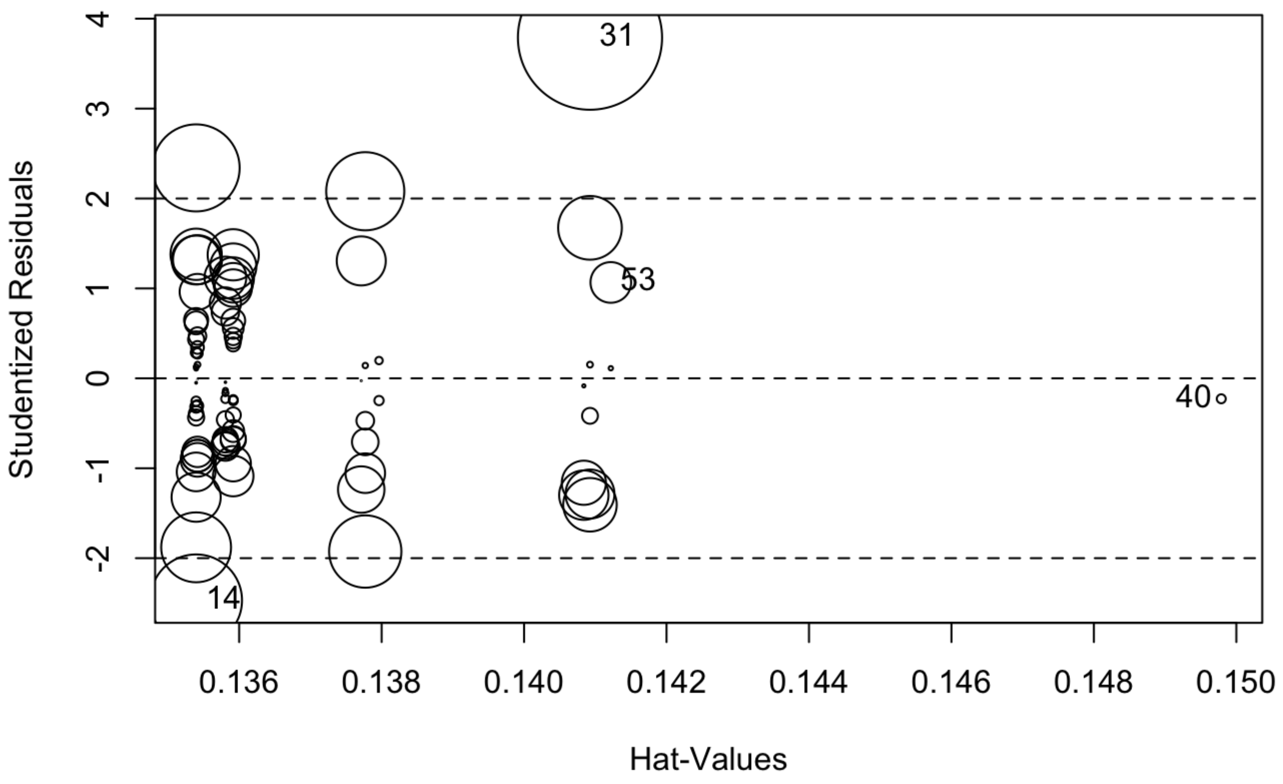


```
# check for multicollinearity
car::vif(m1)
```

```
## Registered S3 methods overwritten by 'car':
##   method                          from
##   influence.merMod                lme4
##   cooks.distance.influence.merMod lme4
##   dfbeta.influence.merMod         lme4
##   dfbetas.influence.merMod        lme4
```

```
## attitude   gender
## 1.000018 1.000018
```

```
# check for influential points
car::influencePlot(m1)
```



```
##       StudRes       Hat       CookD
## 14 -2.4690169 0.1353955 0.318209936
## 31  3.7892423 0.1409253 0.785129716
## 40 -0.2274527 0.1497880 0.003038162
## 53  1.0666339 0.1412176 0.062361289
```

# What’s next

- Crossed vs. nested random effects
- MEM with random slopes

# Further readings and references

- Amazing tutorial and the source of the “politeness” data: http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf or https://arxiv.org/pdf/1308.5499.pdf

- http://coltekin.net/cagri/R/r-exercisesse12.html#x18-5300012

- https://ourcodingclub.github.io/2017/03/15/mixed-models.html

- a bit more advanced article for `afex` package: https://cran.r-project.org/web/packages/afex/vignettes/afex_mixed_example.html

linear regression statistics repeated measures

**Yury Zablotski**

Data Scientist at LMU Munich, Faculty of Veterinary Medicine

Passion for applying Biostatistics and Machine Learning to Life Science Data

## Related

- Mixed Effects Models 3: Random Slopes
- Statistical tests vs. linear regression
- Model diagnostics
- Multiple linear regression
- Constraints for linear regression