Multilevel modelling

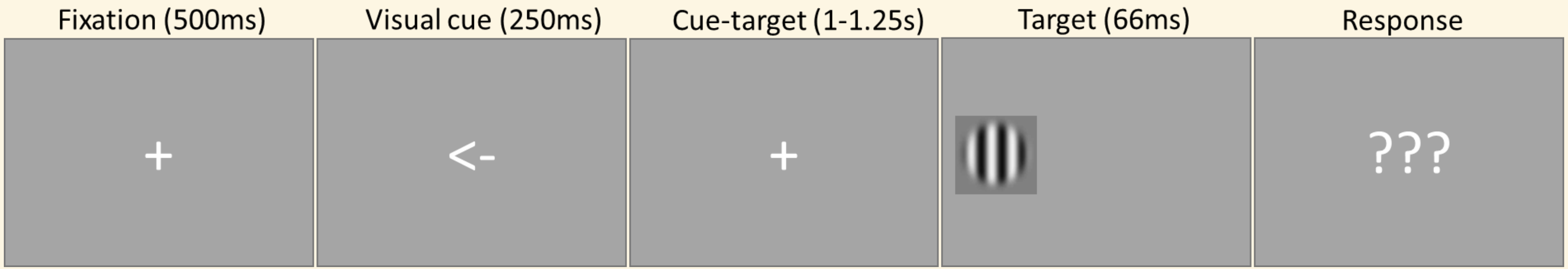
2021/04/20

Multilevel data

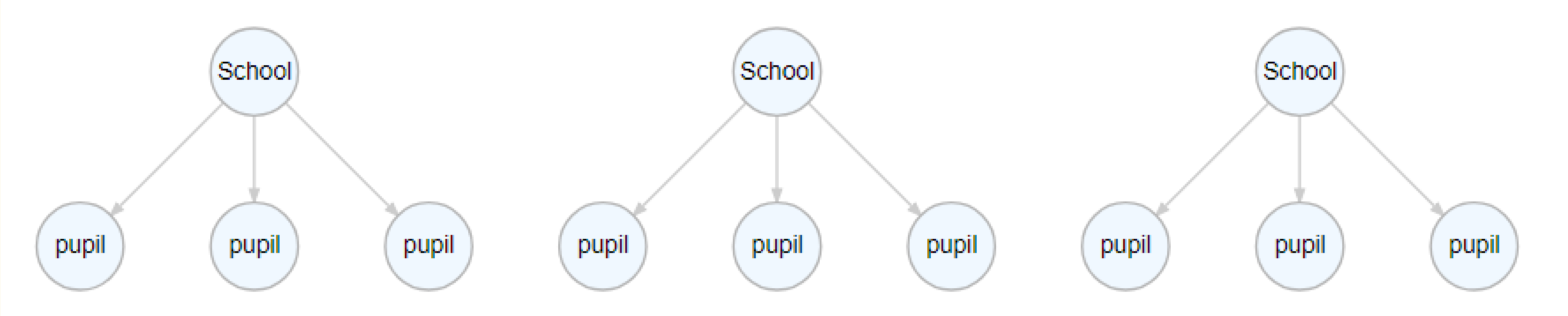
There are *many* situations in psychology where we have *nested* data.

Intervention studies are typically longitudinal - the same participants are tested multiple times on the same outcome measure.

Typical cognitive experiments show participants many repeats of similar trials.



Multilevel data

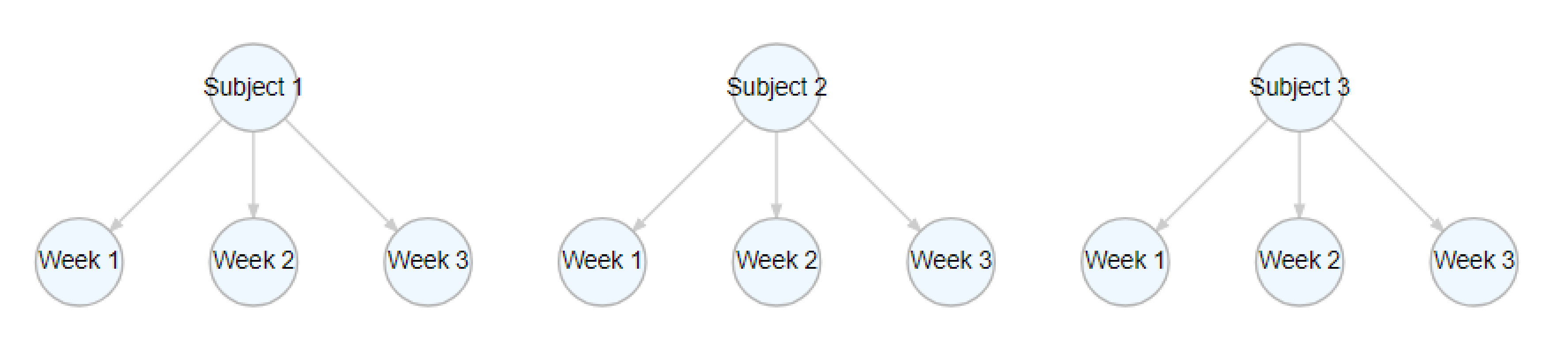


In this example, each pupil is a unit of observation.

But these pupils are not fully independent from each other - pupils who attend one school tend to be more similar to each other than they are to pupils who attend other schools.

Thus, *pupils* (Level 1) are nested in *schools* (Level 2).

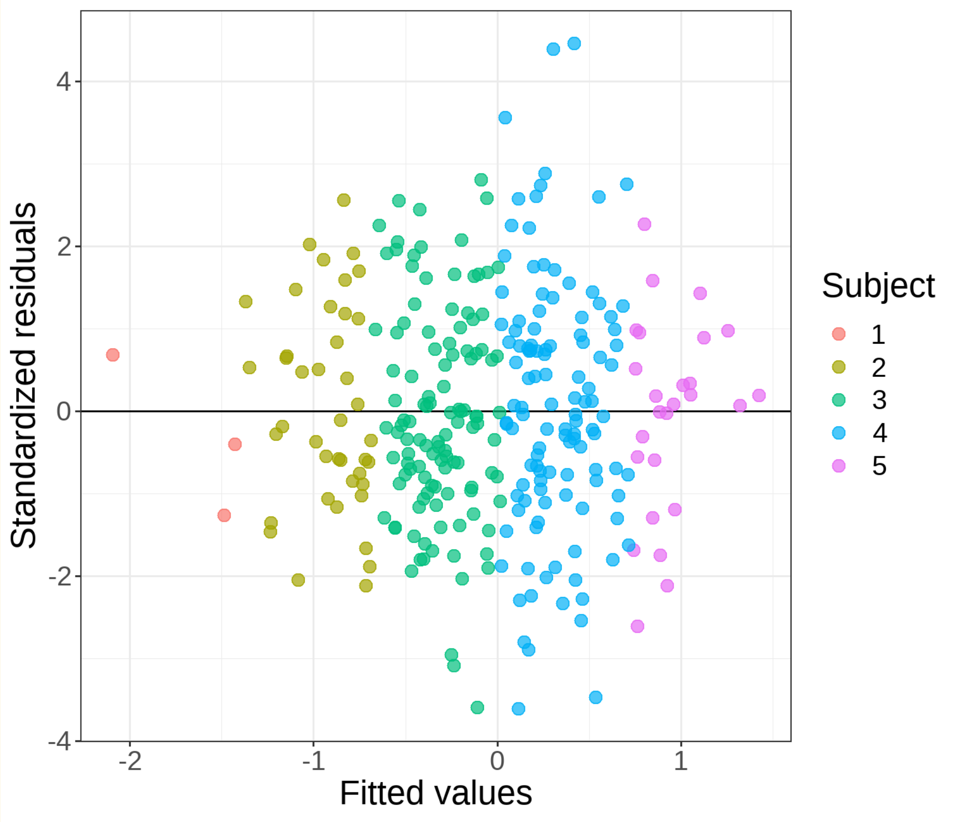
Multilevel data



Other data may be *longitudinal*. For example, you may measure outcomes such as, for example, performance or attitudes on repeated occasions to see how they vary over time.

The measurements each week are the main unit of observation, but they are nested within subjects.

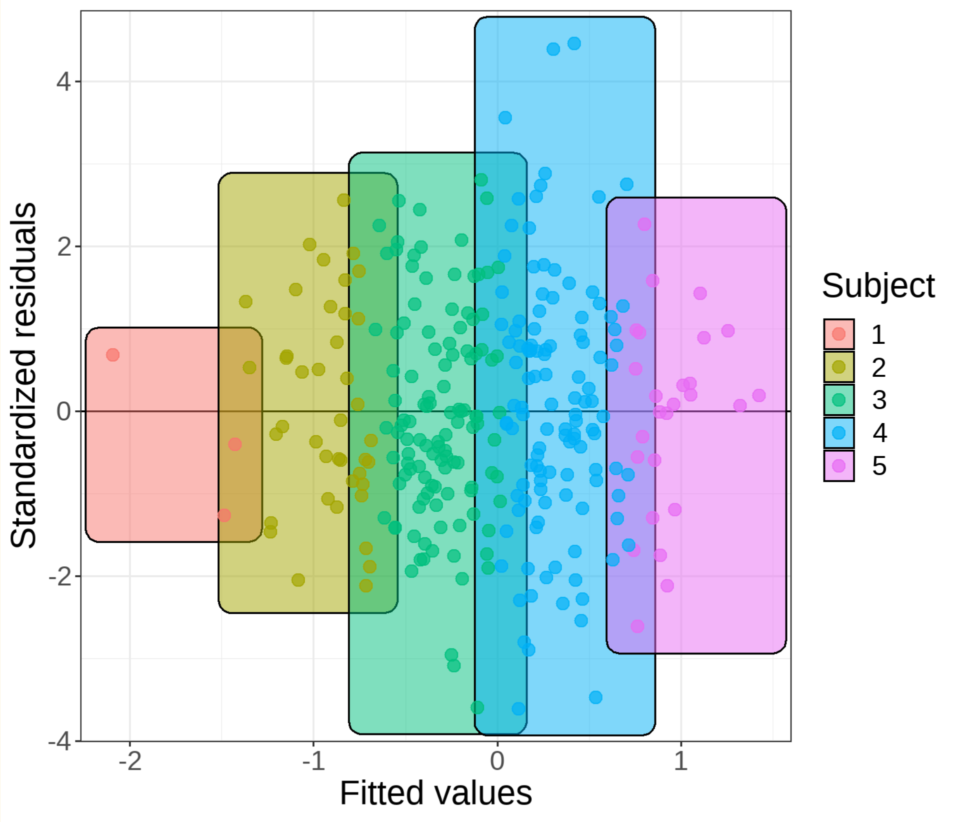
Clustered data



Data from nested designs like those we have just seen often have *clusters* of correlated observations.

Different people have different reaction speeds, or baseline attitudes; different schools have different teachers and different general environments.

Clustered data



Data from nested designs like those we have just seen often have *clusters* of correlated observations.

Different people have different reaction speeds, or baseline attitudes; different schools have different teachers and different general environments.

The problem with nesting

sleepstudy

```r
head(sleepstudy, 12)
```

```
##    Reaction Days Subject
## 1  249.5600    0     308
## 2  258.7047    1     308
## 3  250.8006    2     308
## 4  321.4398    3     308
## 5  356.8519    4     308
## 6  414.6901    5     308
## 7  382.2038    6     308
## 8  290.1486    7     308
## 9  430.5853    8     308
## 10 466.3535    9     308
## 11 222.7339    0     309
## 12 205.2658    1     309
```

The *sleepstudy* dataset contains data from a sleep deprivation experiment.

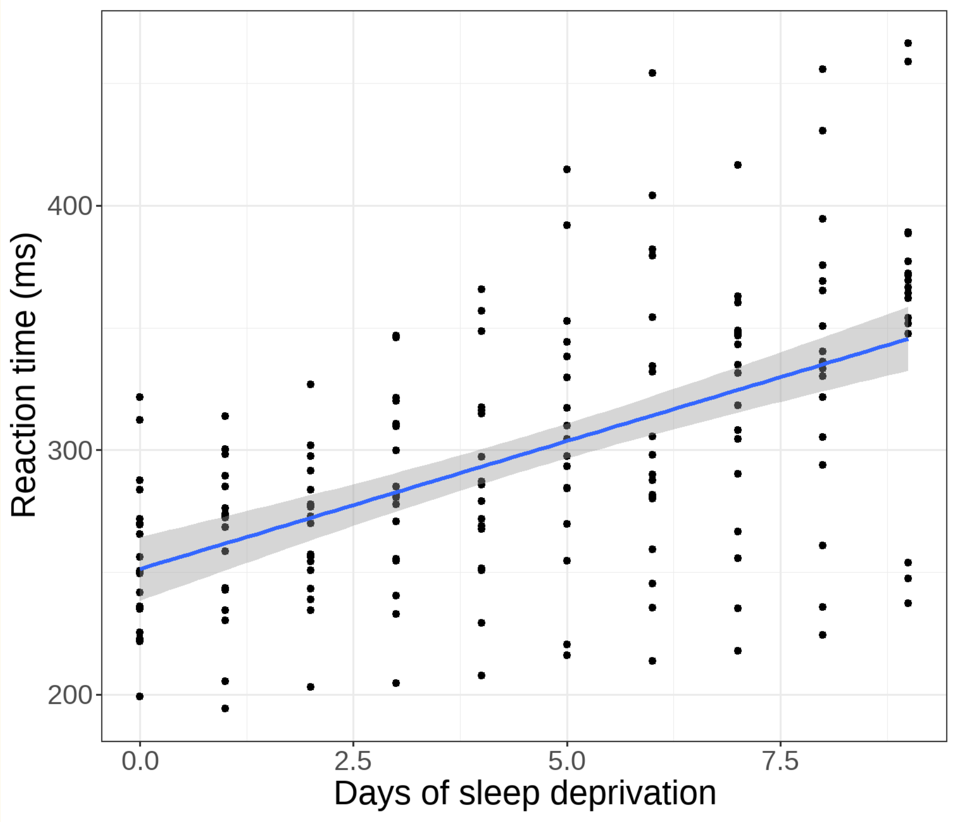
Over the course of ten days, subjects were only allowed to sleep for 3 hours each night.

Each day their reaction times on a variety of cognitive tasks were recorded.

This is a *nested*, multilevel design.

Each observation - average RT on a given day - is nested within a *subject*.

sleepstudy

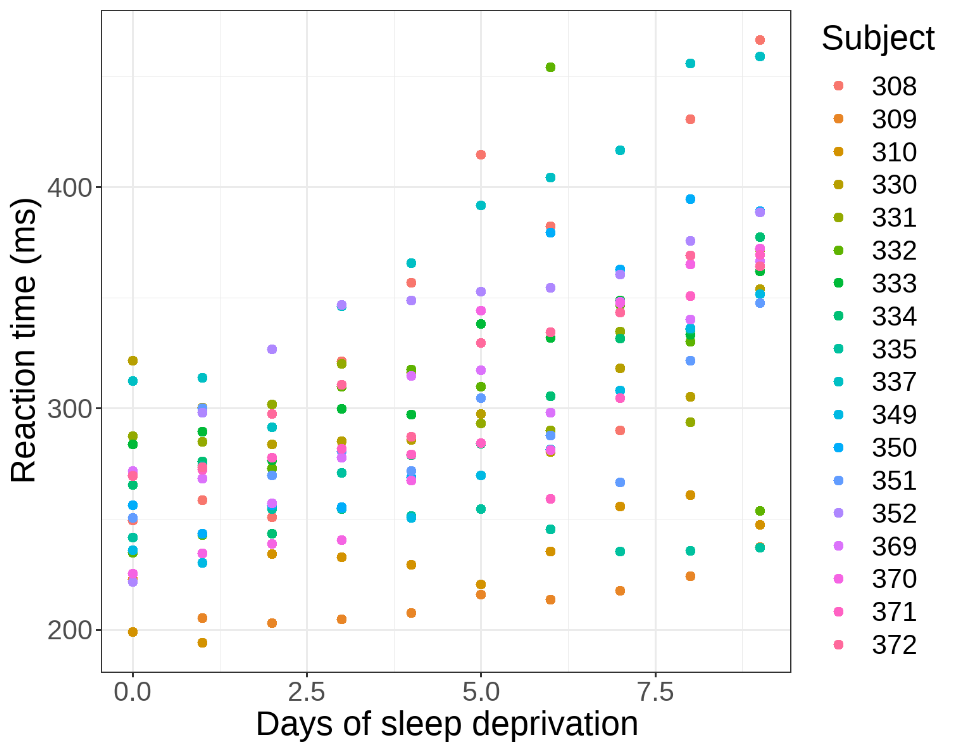


We could simply fit a linear model to the whole dataset.

```
basic_lm <- lm(Reaction ~ Days,
               data = sleepstudy)
basic_lm
```

```
##
## Call:
## lm(formula = Reaction ~ Days, data = sleepstudy)
##
## Coefficients:
## (Intercept)         Days
##      251.41        10.47
```

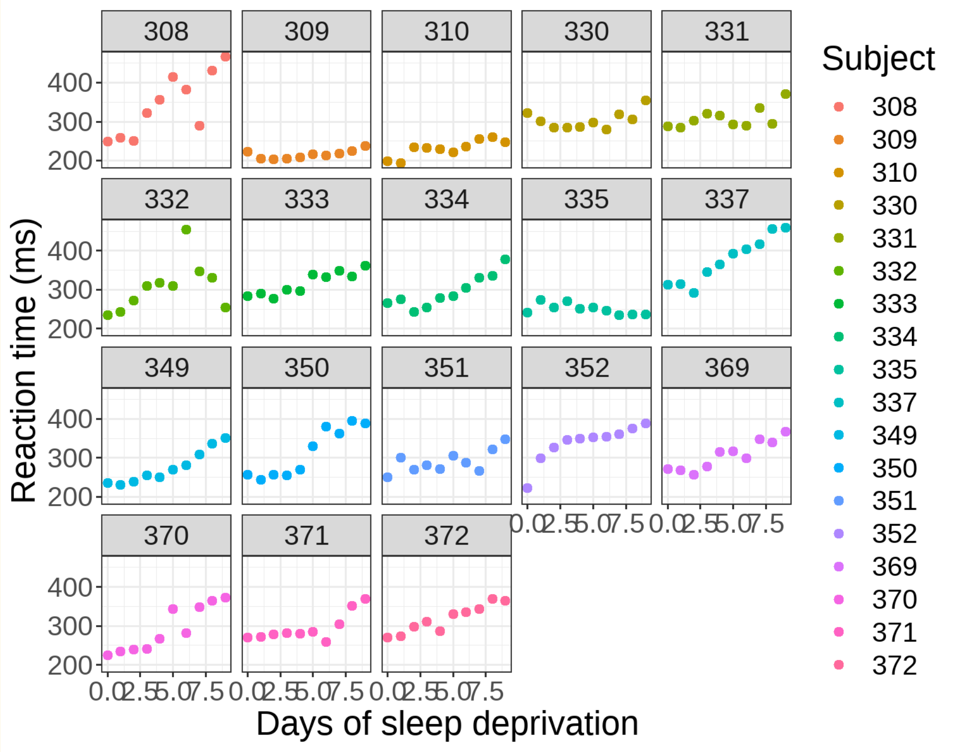
Joint graph | Split panels | All the things



But the data clearly has more structure than that!

Here each dot is coloured to show which participant contributed which data points.

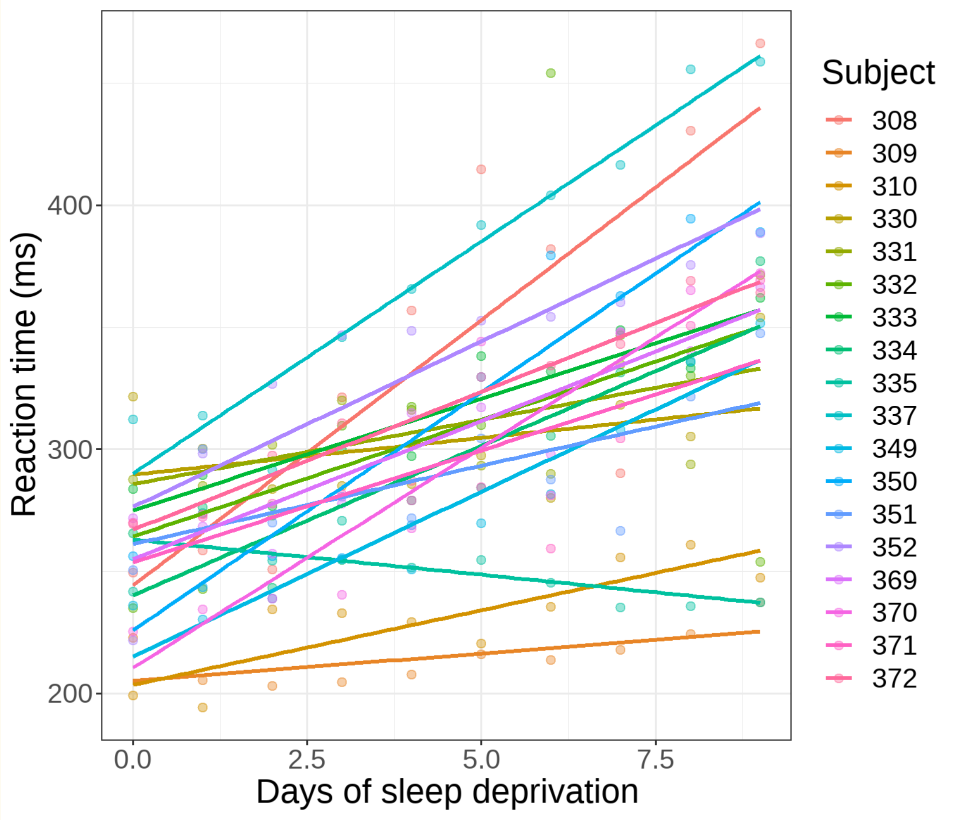
Joint graph | **Split panels** | All the things



If we split the plot up to show each subject separately, we get more sense of the variability.

For example, Subject 308 shows a very strong effect of sleep deprivation on reaction time, while Subject 309 shows very little effect of sleep deprivation.

Joint graph | Split panels | All the things

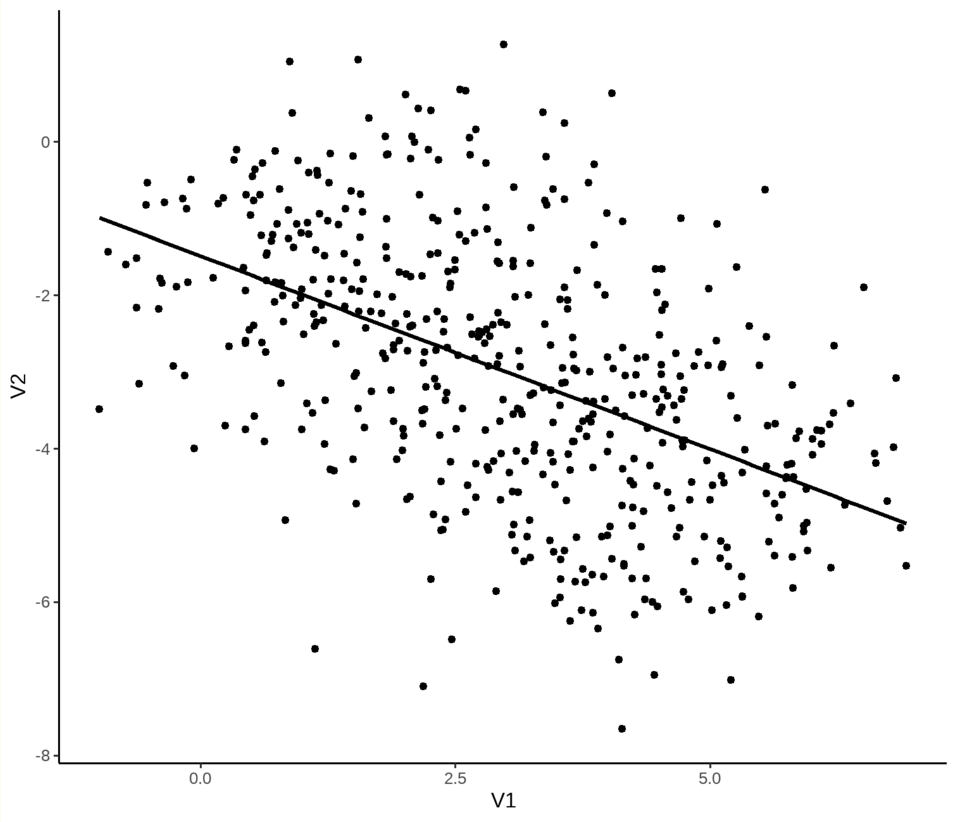


Our simple linear model ignores the fact that many of our observations are repeated measurements from each participant.

It assumes the effect is the same for everyone.

There are 18 participants in this study. Some of them are generally faster or slower than others; some of them show more effect of sleep deprivation than others.

Simpson's paradox

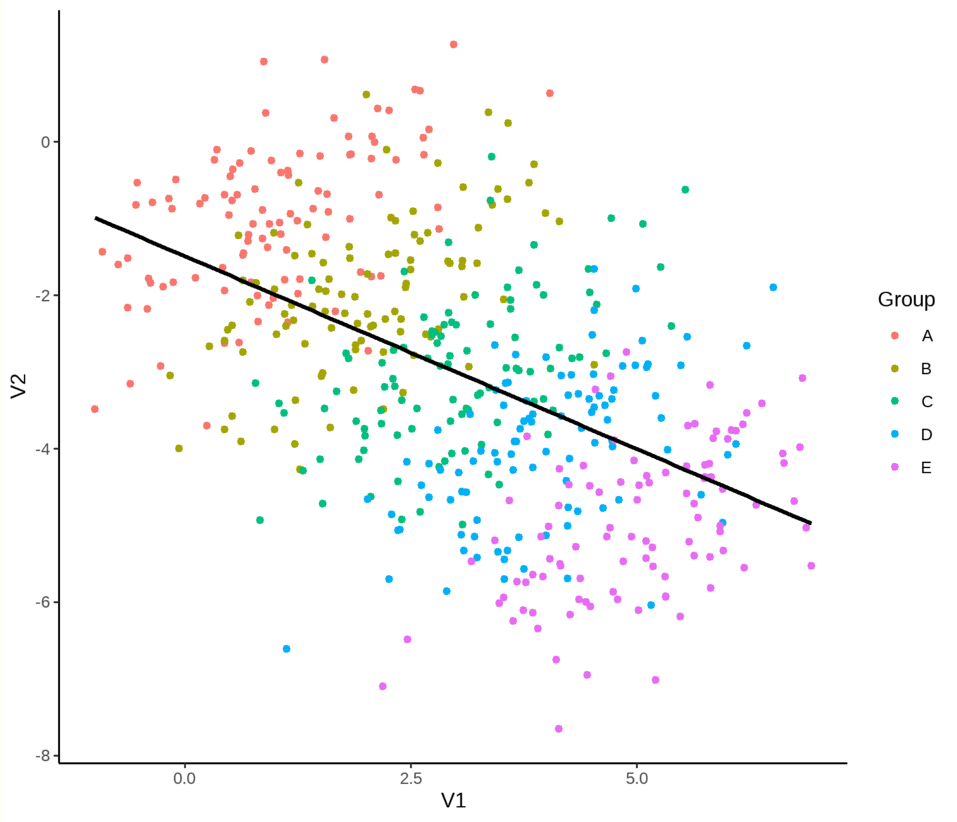


This data has a correlation coefficient of

-0.5

As V1 increases, V2 decreases!

Simpson's paradox

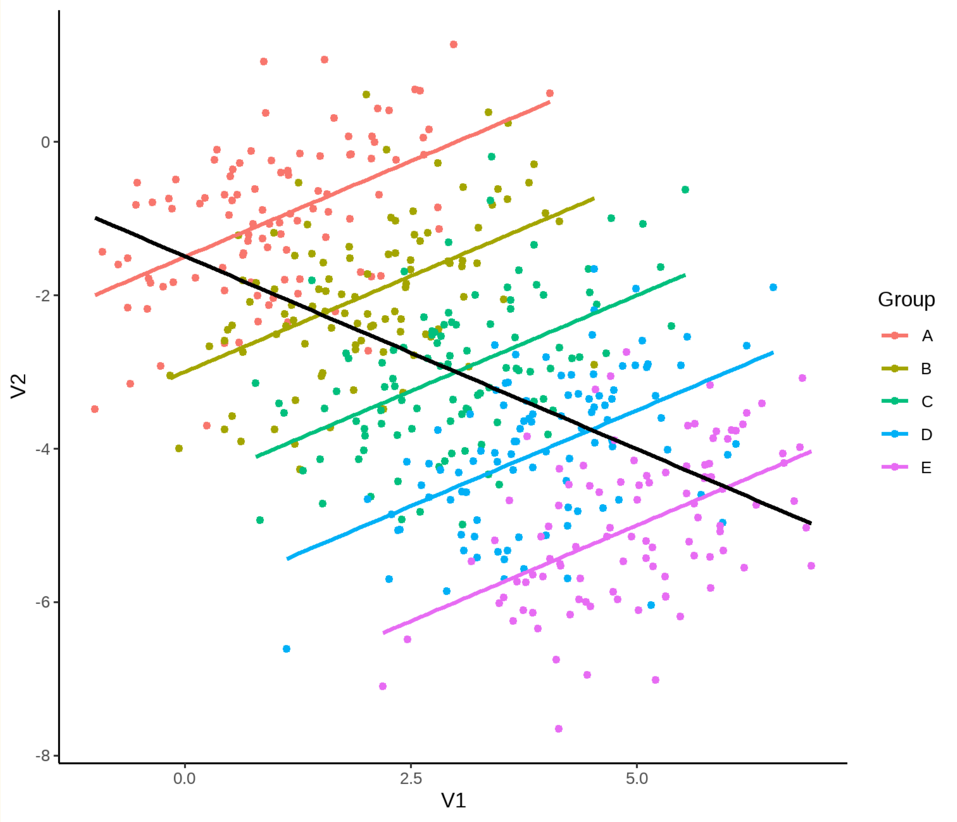


But wait!

What is this?

There are five different groups of people?

Simpson's paradox



Within each group, the correlation is the other way round - as V1 increase, V2 also increases!

This is known as **Simpson's paradox**, or the *ecological fallacy*.

The effect if grouping is ignored is the *reverse* of the effect in each individual group.

Estimating multilevel models

Multilevel models

Multilevel models allow us to account for the nested, correlated nature of the data, and explicitly model the variability between people.

You may also see them called:

- Hierarchical models
- Mixed-effects models
- Random-effects models
- Mixed models

Multilevel models using lme4

The most important library for fitting this type of model is `lme4`.

A multilevel model can be fitted with the `lmer()` function.

```
library(lme4)
multilev <-
  lmer(Reaction ~ 1 + Days + (1 + Days | Subject),
       data = sleepstudy)
```

lmer(Reaction ~ 1 + Days + (1 + Days| Subject), data = sleepstudy)

Fixed effects are highlighted in blue.

Random effects are highlighted in red.

Fixed and random effects

Fixed effects are the *population-average* effect: e.g. the *average* effect of days of sleep deprivation on reaction time.

Random effects are those that vary across the *sampling units*. e.g. the variation in average reaction time across people

They are *random* because the *sampling units* are randomly drawn from a wider *population*. e.g. the specific participants in an experiment are usually a random subset of all possible participants

A basic linear model

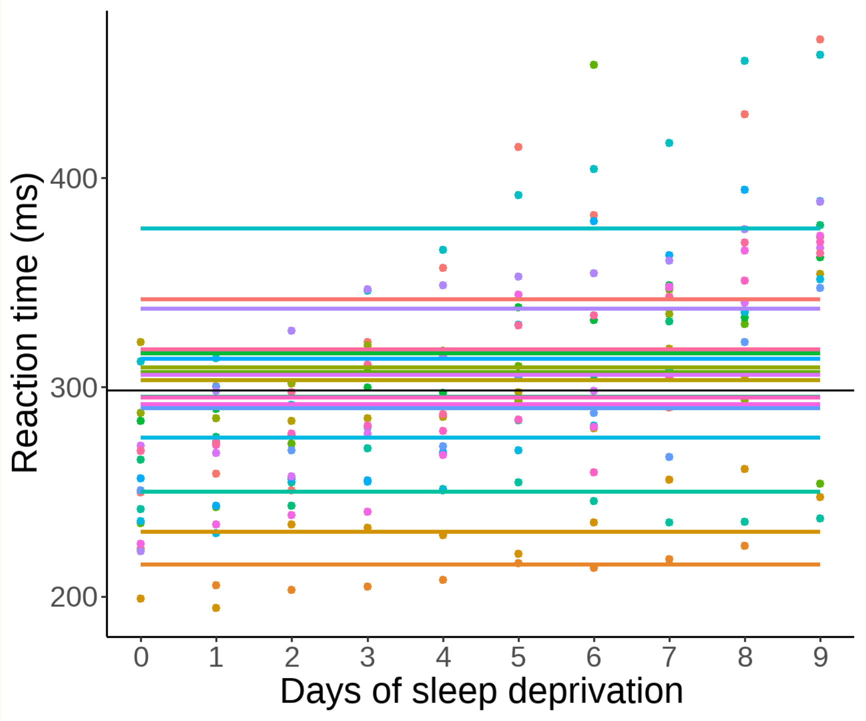
```
basic_lm <- lm(Reaction ~ 1 + Days, data = sleepstudy)
summary(basic_lm)
```

```
##
## Call:
## lm(formula = Reaction ~ 1 + Days, data = sleepstudy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.848  -27.483    1.546   26.142  139.953
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  251.405      6.610  38.033  < 2e-16 ***
## Days          10.467      1.238   8.454 9.89e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 47.71 on 178 degrees of freedom
## Multiple R-squared:  0.2865,	Adjusted R-squared:  0.2825
## F-statistic: 71.46 on 1 and 178 DF,  p-value: 9.894e-15
```

Random intercepts

Individual intercepts

Individual intercepts | Split by subject

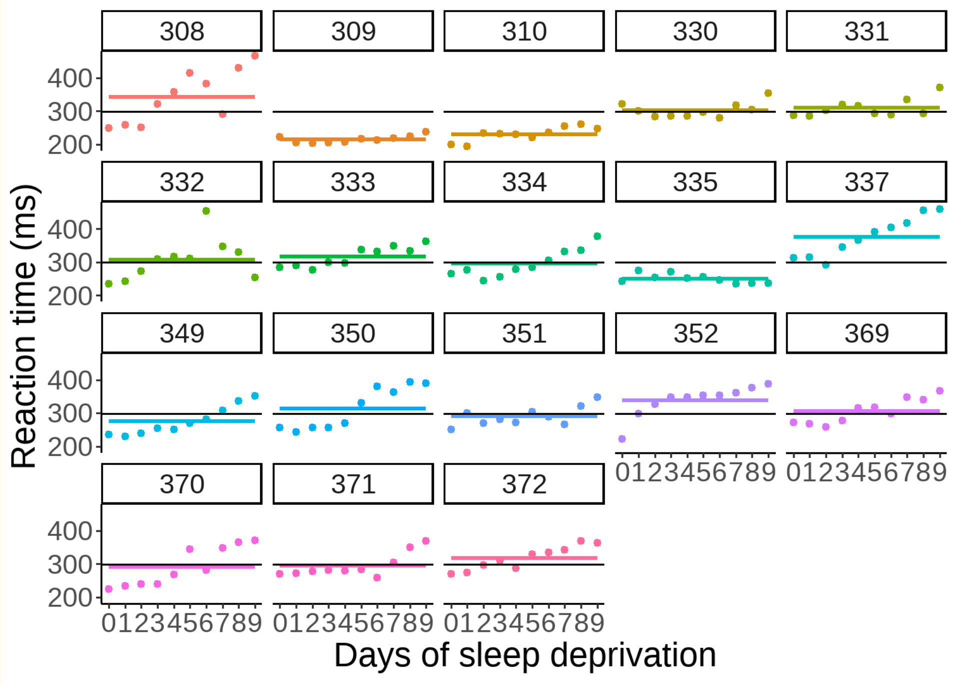


The black line on this plot shows the overall **mean** reaction time. This is the *intercept* of the basic model.

Each coloured line on this plot shows an individual participant's **mean** reaction time.

Individual intercepts

Individual intercepts | Split by subject



If we look at the plots individually for each subject, we can see a little the individual intercepts a little more clearly.

Some people are faster on average than the overall mean, while others are slower.

A *random-intercept* model models that variability!

Modelling random intercepts

Remember that in our basic model, the *intercept* represents the mean reaction time.

We can model the variability of the intercept better by including a *random effect* term - *(1 | Subject)*.

```
int_only <-
  lmer(Reaction ~ 1 + Days + (1 | Subject),
       data = sleepstudy) # Random intercept
```

This model is a *random-intercept* model - it captures how mean reaction times vary across subjects.

```r
summary(int_only)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ 1 + Days + (1 | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1786.5
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.2257 -0.5529  0.0109  0.5188  4.2506
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Subject  (Intercept) 1378.2   37.12
##  Residual              960.5   30.99
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 251.4051     9.7467   25.79
## Days         10.4673     0.8042   13.02
##
## Correlation of Fixed Effects:
##      (Intr)
## Days -0.371
```

```r
tab_model(basic_lm, int_only, dv.labels = c("Reaction time (ms)", "Reaction time (ms)"))
```

	Reaction time (ms)			Reaction time (ms)		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	251.41	238.36 – 264.45	**<0.001**	251.41	232.30 – 270.51	**<0.001**
Days	10.47	8.02 – 12.91	**<0.001**	10.47	8.89 – 12.04	**<0.001**
Random Effects						
σ^2				960.46		
τ_{00}				1378.18 $_{Subject}$		
ICC				0.59		
N				18 $_{Subject}$		
Observations	180			180		
R^2 / R^2 adjusted	0.286 / 0.282			0.280 / 0.704		

Standard linear model

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 251.40510   6.610154 38.033169 2.156888e-87
## Days         10.46729   1.238195  8.453663 9.894096e-15
```

Intercept only mixed-model

```
##               Estimate Std. Error  t value
## (Intercept) 251.40510  9.7467163 25.79383
## Days         10.46729  0.8042214 13.01543
```

The *standard errors* differ, which means the *t-values* differ.

The *intercept* variability increased, while the `Days` variability decreased!

Random effects

The *fixed* effects give us a measure of average performance and the overall effect of Days of sleep deprivation on RT.

```
fixef(int_only)
```

```
## (Intercept)        Days
##   251.40510    10.46729
```

The *random* effects tell us how much variability there is *between-participants*. In this case, we only estimated participant-specific intercepts.

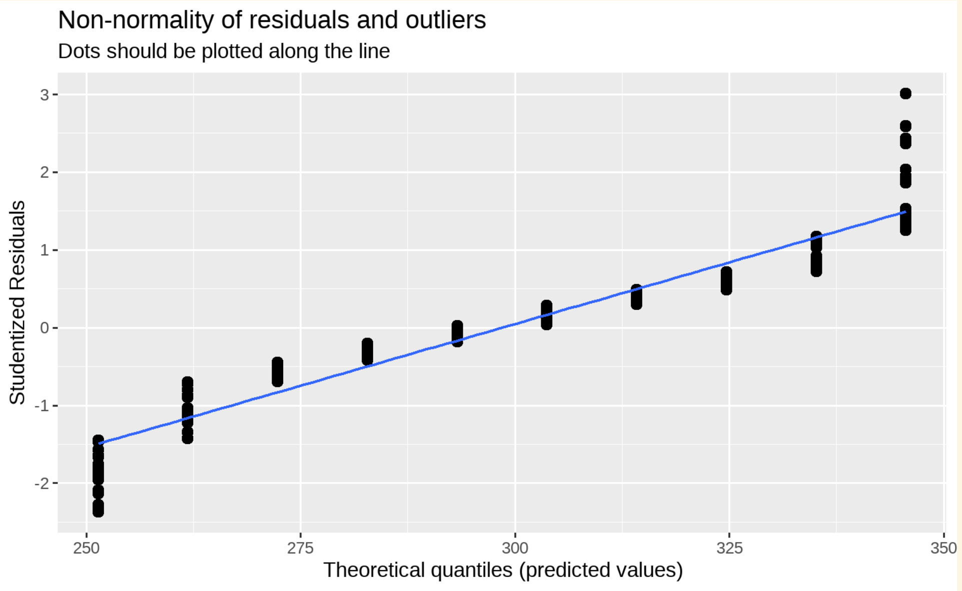
```
summary(int_only)$varcor
```

```
##  Groups   Name        Std.Dev.
##  Subject  (Intercept) 37.124
##  Residual             30.991
```

A quick look at the residuals

Standard model | Mixed model

```r
library(sjPlot)
plot_model(basic_lm, type = "diag")[[1]]
```



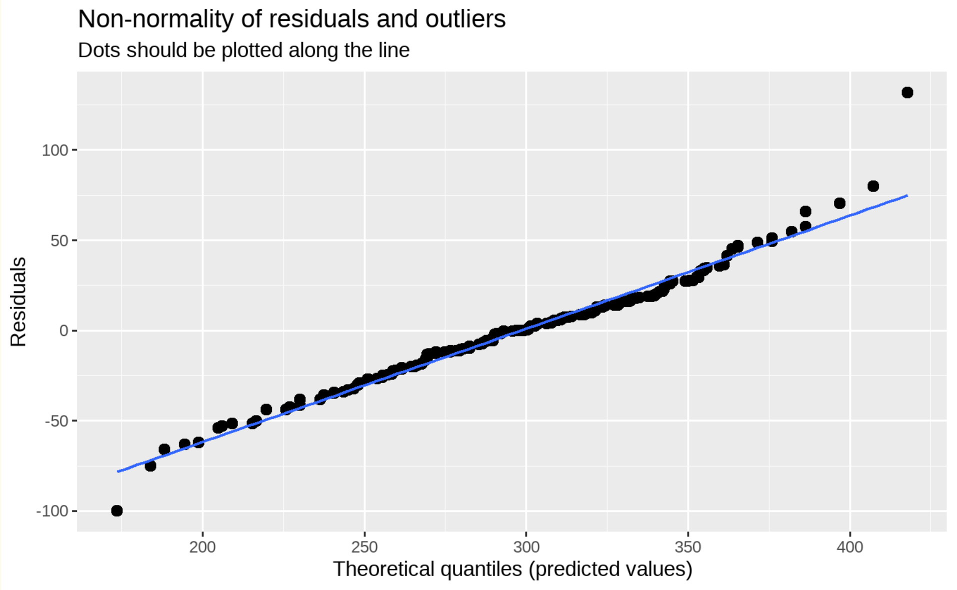
These residuals don't look great - the dots seems to show a slight curve, and our predictions at each end are also poor.

This suggests there's some structure not being captured by the model.

A quick look at the residuals

Standard model | Mixed model

```
plot_model(int_only, type = "diag")[[1]]
```



This model - the *random intercept* model - is doing a *much much* better joib than our basic linear model.

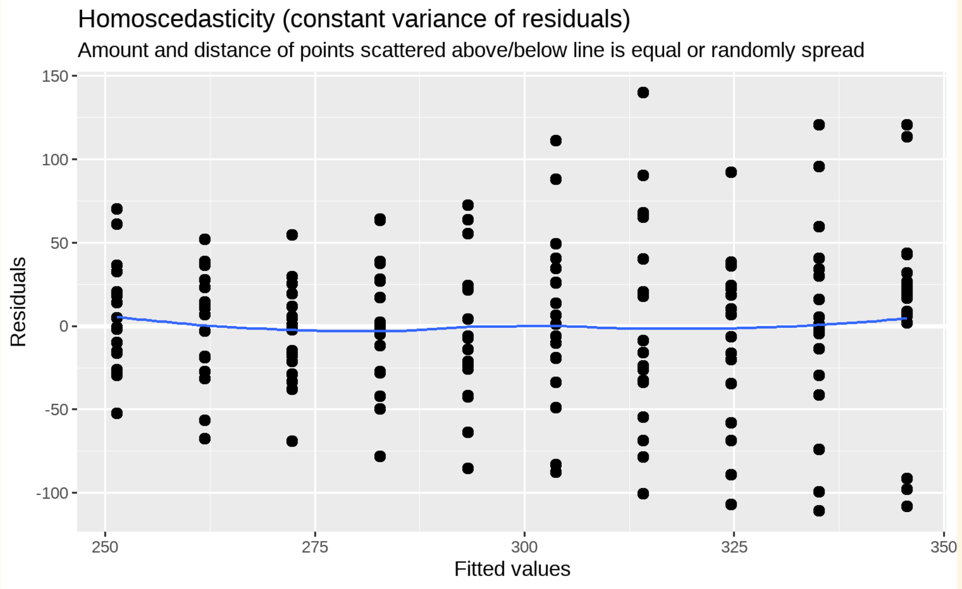
The points now lie almost entirely along the line.

This indicates a better correspondence between the model predictions and the actual data!

A quick look at the residuals

Standard model | Mixed model

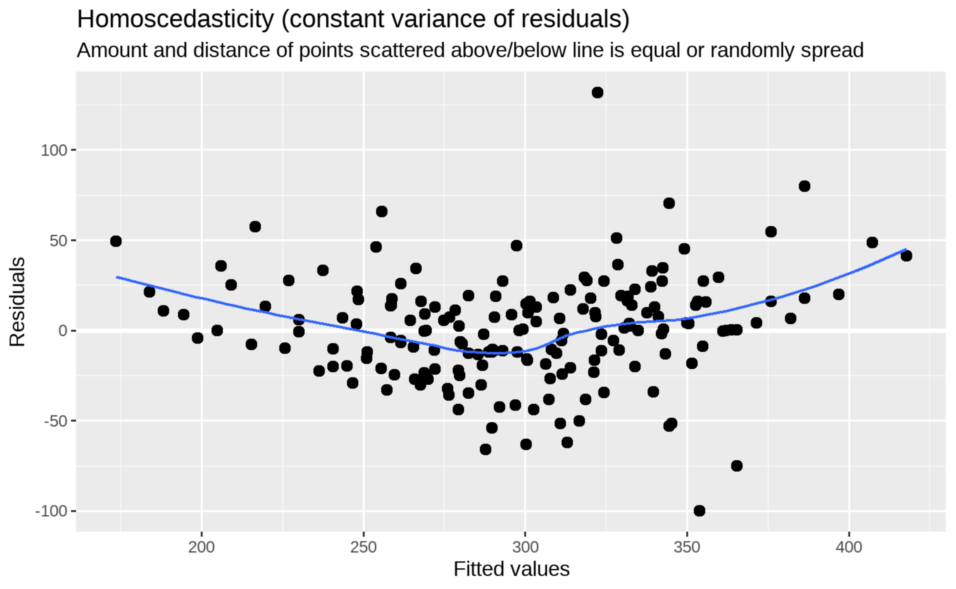
```
plot_model(basic_lm, type = "diag")[[3]]
```



A quick look at the residuals

Standard model | **Mixed model**

```r
plot_model(int_only, type = "diag")[[4]]
```



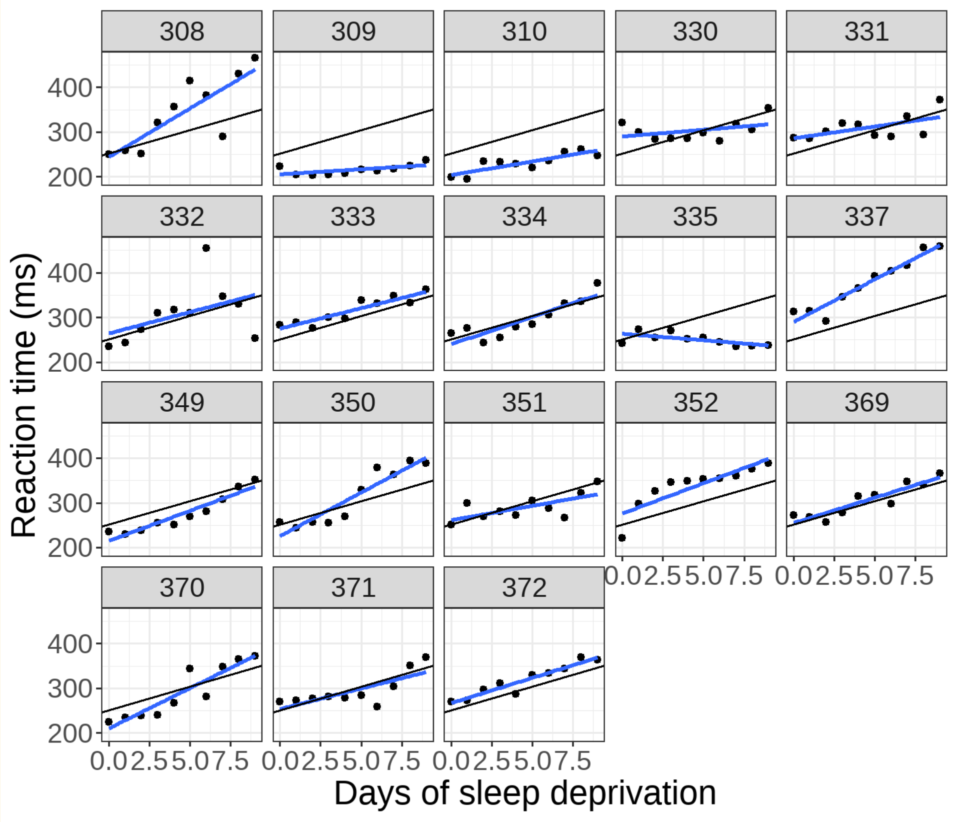
But this plot shows evidence of heteroskedasticity - non-constant variance.

The dots seem to *curve* somewhat.

This suggests there is still something not quite right in our model.

Random slopes

Individual slopes



This plot now show individual plots for each participant with the individual effect of `Days` added.

The general trend is consistent, but it's clear that some participants have stronger effects than others.

And it looks a little like people who are generally fast responders show *less* effect of `Days` of sleep deprivation.

Modelling random slopes

We can model how much the effect of `Days` varies between participants by adding *random slopes* to our model - `(Days | Subject)`.

```
random_slope <- lmer(Reaction ~ 1 + Days + (1 + Days | Subject),
                     data = sleepstudy)
```

Note that `Days` now appears twice.

The first time models the *population-average* effect of `Days`.

The second time models the *individual* effect of `Days`.

```r
summary(random_slope)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Reaction ~ 1 + Days + (1 + Days | Subject)
##    Data: sleepstudy
##
## REML criterion at convergence: 1743.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.9536 -0.4634  0.0231  0.4634  5.1793
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  Subject  (Intercept) 612.10   24.741
##           Days         35.07    5.922   0.07
##  Residual             654.94   25.592
## Number of obs: 180, groups:  Subject, 18
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)  251.405      6.825  36.838
## Days          10.467      1.546   6.771
##
## Correlation of Fixed Effects:
##      (Intr)
```

	Reaction times (ms)			Reaction times (ms)		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	251.41	232.30 – 270.51	**<0.001**	251.41	238.03 – 264.78	**<0.001**
Days	10.47	8.89 – 12.04	**<0.001**	10.47	7.44 – 13.50	**<0.001**
Random Effects						
σ^2	960.46			654.94		
τ_{00}	1378.18 $_{\text{Subject}}$			612.10 $_{\text{Subject}}$		
τ_{11}				35.07 $_{\text{Subject.Days}}$		
ρ_{01}				0.07 $_{\text{Subject}}$		
ICC	0.59			0.72		
N	18 $_{\text{Subject}}$			18 $_{\text{Subject}}$		
Observations	180			180		
Marginal R^2 / Conditional R^2	0.280 / 0.704			0.279 / 0.799		

Model comparisons

Is this model an improvement? Use `anova()` to check!

```
anova(int_only, random_slope)
```

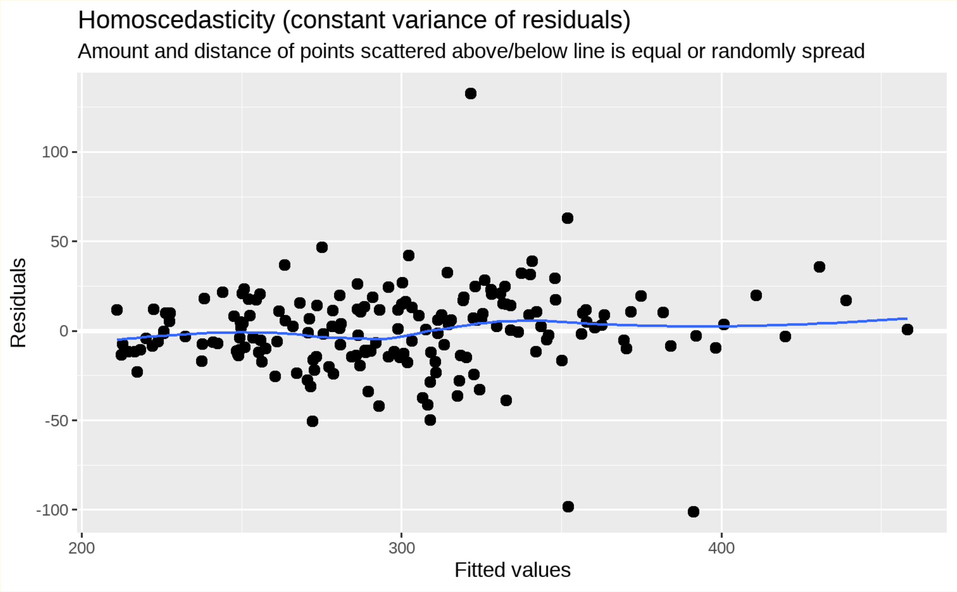
```
## refitting model(s) with ML (instead of REML)

## Data: sleepstudy
## Models:
## int_only: Reaction ~ 1 + Days + (1 | Subject)
## random_slope: Reaction ~ 1 + Days + (1 + Days | Subject)
##              npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
## int_only        4 1802.1 1814.8 -897.04   1794.1
## random_slope    6 1763.9 1783.1 -875.97   1751.9 42.139  2  7.072e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(it's significant, so yes!)

A quick look at the residuals

```
plot_model(random_slope, type = "diag")[[4]]
```



These residuals are the best of all so far.

A few points look suspiciously like outliers, but overall, there's little to suggest any particular problems with this model!

Multiple random effects

The "language as fixed-effect" fallacy

A common circumstance in psychological research is that we have more than one random effect.

For example, in language experiments, subject often need to read a many different words; these may be words from different categories, or vary in other ways.

These words themselves are random samples, but many researchers treat them as being *fixed*.

Clark, 1973

The `politeness` study

Winter and Grawunder (2012) looked at the relationship between vocal pitch and the level of politeness of a sentence.

Participants were asked to imagine how they would respond to a variety of scenarios when talking politely or informally.

```
politeness <- read_csv("data/politeness_data.csv")
head(politeness)
```

```
## # A tibble: 6 x 5
##   subject gender scenario attitude frequency
##   <chr>   <chr>     <dbl> <chr>        <dbl>
## 1 F1      F             1 pol           213.
## 2 F1      F             1 inf           204.
## 3 F1      F             2 pol           285.
## 4 F1      F             2 inf           260.
## 5 F1      F             3 pol           204.
## 6 F1      F             3 inf           287.
```

The `politeness` study

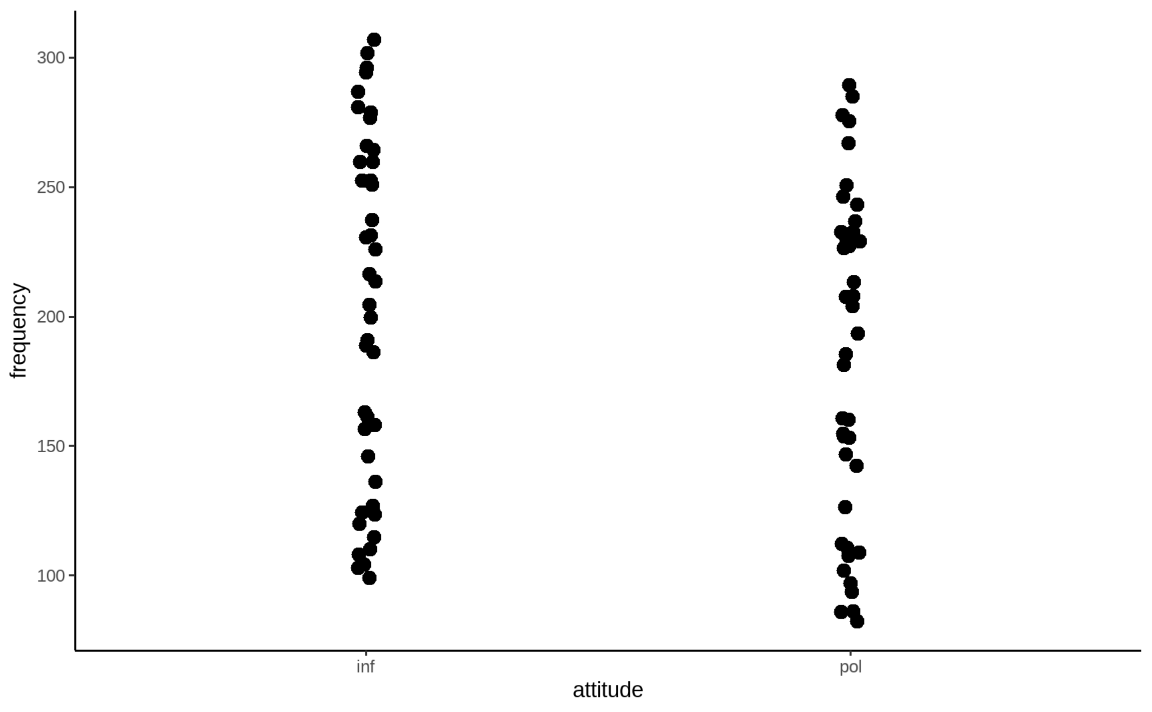
In the `politeness` study, there are *two* distinct groupings:

1) Subjects repeat the same task (imaging a scenario) over and over again

2) Individual scenarios are repeated by different subjects

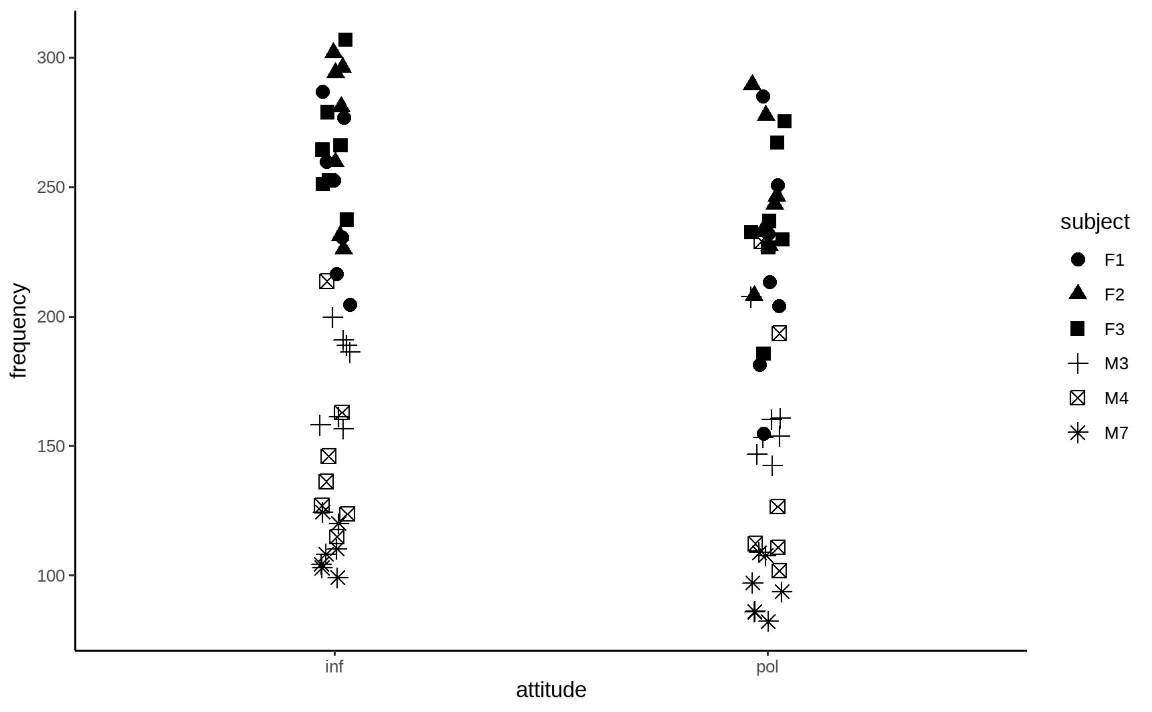
Thus there are *two* possible sources of correlated data - we'd expect responses to particular scenarios to be fairly consistent across subjects, and responses by individual subjects to be fairly consistent across items, All data | Subject | Scenario | Both

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



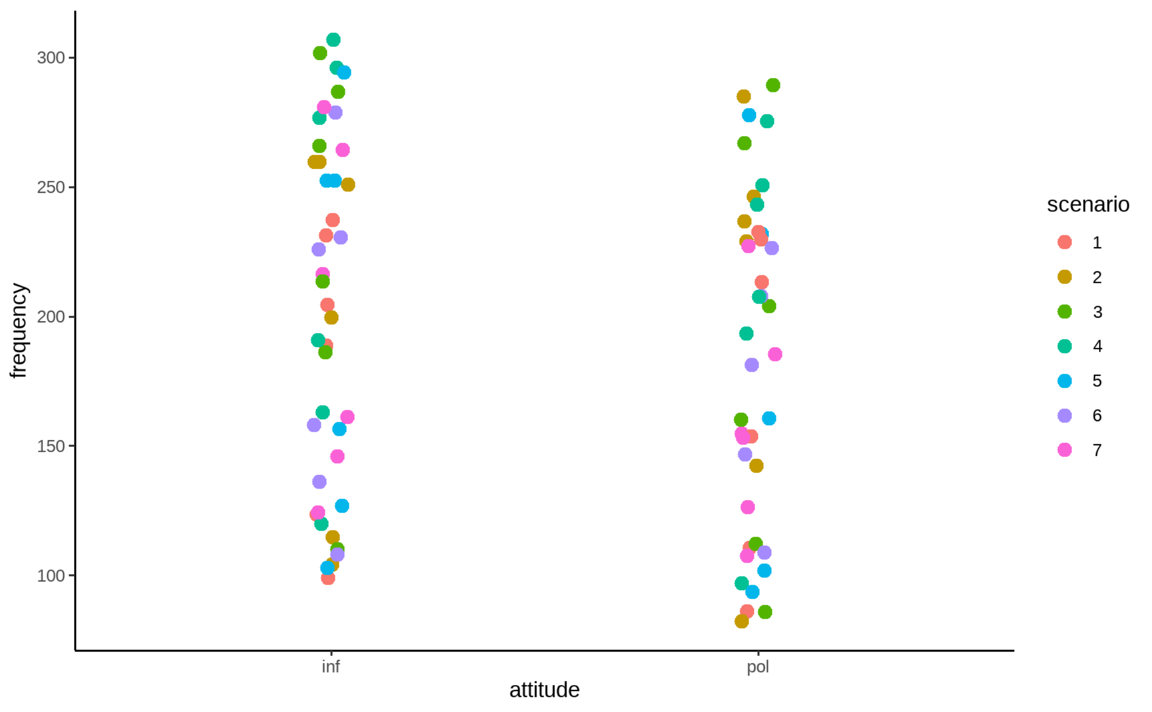
All data | **Subject** | Scenario | Both

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



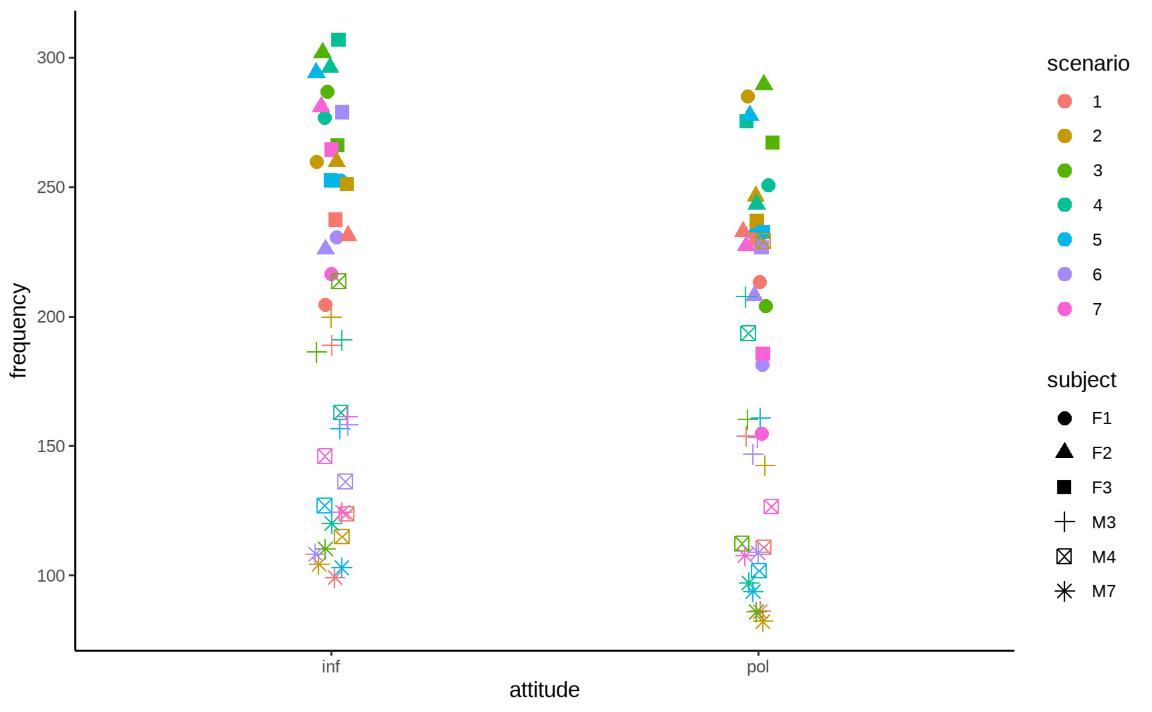
All data | Subject | **Scenario** | Both

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



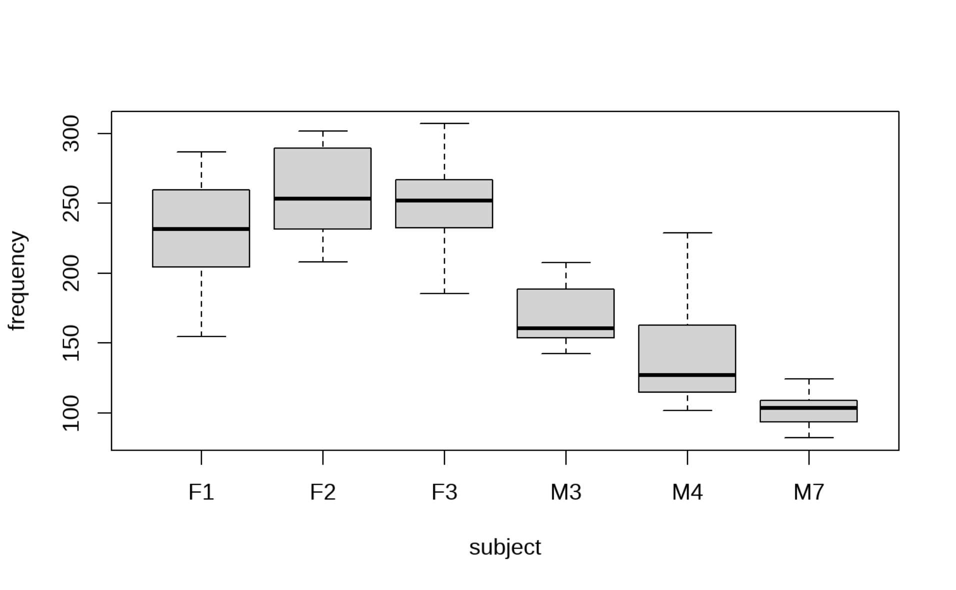
All data Subject Scenario **Both**

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Variability between subjects

```
boxplot(frequency ~ subject,
        data = politeness)
```

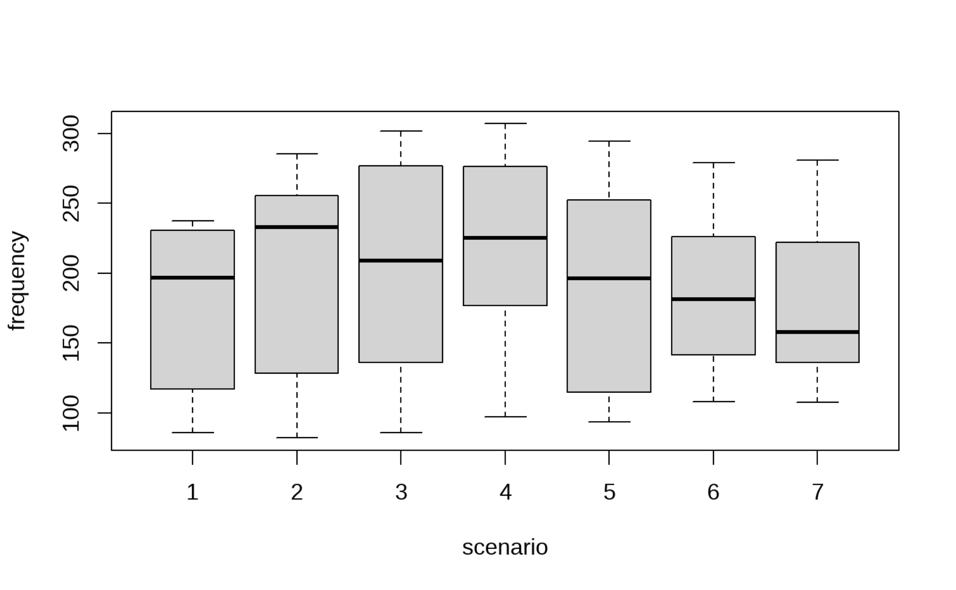


Individual participants vary in their baseline vocal frequency.

Male participants typically have lower frequency voices than female participants.

Variability between scenarios

```
boxplot(frequency ~ scenario,
        data = politeness)
```



There seems to be some variability across scenarios.

Scenario 7 seems consistently lower than scenario 4, for example.

But there does seem to be less variability than across participants.

Multiple random effects

We can model *both* of these sources of variability simultaneously by adding *multiple random effects*.

```
full_mod <- lmer(frequency ~ attitude + (1|subject) + (1|scenario),
                 data = politeness)
```

Whereas before we only added `(1|subject)`, here we also add `(1|scenario)`.

This models separate intercepts for each subject and each scenario, allowing for, for example, high-pitched individuals or scenarios that typically elicit low-pitched responses.

```
## 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
```

	Frequency (Hz)		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	202.59	150.15 – 255.02	**<0.001**
attitude [pol]	-19.69	-30.64 – -8.75	**<0.001**
Random Effects			
σ^2	646.02		
τ_{00} scenario	218.98		
τ_{00} subject	4014.54		
ICC	0.87		
$N_{subject}$	6		
$N_{scenario}$	7		
Observations	83		
Marginal R^2 / Conditional R^2	0.020 / 0.870		

Some final words and references

Generalized linear mixed effects models

As discussed last week, there are many types of data for which a linear model is *inappropriate*.

Fortunately, we can fit **generalized linear mixed effects models** too!

```
glmer(DV ~ IV1 + IV2 + (IV1 | random_factor), family = binomial(),
data = your_data)
```

Additional reading

Complete vs Partial vs no pooling

An introduction to mixed models

Keep it Maximal

Generalizing over encounters: statistical and theoretical considerations

Understanding mixed-effects models through data simulation