Statistics for Linguists
08 July 2022

10:00	Workshop introduction
10:15	Loading and exploring datasets
10:45	Data transformation and coding
11:15	Practical exercise
12:15	Review of practical
12:30 - 13:30	LUNCH BREAK
13:30	lmer and glmer
14:30	Post-hoc analysis and model visualization
15:00	Practical exercise
16:00	Review of practical
16:15	Model building
17:00	End of workshop

Statistics for Linguists

Post-hoc analysis and model visualization

Margreet Vogelzang – mv498@cam.ac.uk

Learning objectives

- You will learn to load/import data
- Explore a dataset and create descriptive statistics
- Transform a dataset (if needed)
- Code your factors
- Build a mixed model
- **Perform post-hoc statistics**
- **Visualize your data and your model**

What does this mean?

> m5.lmer <- lmer(log(ReadingTime) ~ capitalization * determiner + (1 | participant), data = psycholinguistics_data)

> summary(m5.lmer)

```
 Groups      Name        Variance Std.Dev.
 participant (Intercept) 0.03543  0.1882
 Residual                0.31941  0.5652
Number of obs: 2039, groups:  participant, 30

Fixed effects:
                             Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                   5.81156    0.03661   28.85073 158.753   <2e-16 ***
capitalization1               0.01593    0.01258 2006.66649   1.266    0.205
determiner1                  -0.19560    0.01260 2009.09620 -15.520   <2e-16 ***
capitalization1:determiner1  -0.01616    0.01257 2006.50155  -1.285    0.199
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

What does this mean?

> m6.lmer <- lmer(log(ReadingTime) ~ condition + (1 | participant), data = psycholinguistics_data)

> summary(m6.lmer)

```
 Groups      Name        Variance Std.Dev.
 participant (Intercept) 0.03543  0.1882
 Residual                0.31941  0.5652
Number of obs: 2039, groups:  participant, 30

Fixed effects:
                 Estimate Std. Error         df t value Pr(>|t|)
(Intercept)       5.97508    0.04188   49.36933 142.665   <2e-16 ***
condition+C/-D   -0.35888    0.03593 2008.31218  -9.988   <2e-16 ***
condition-C/+D    0.06418    0.03402 2006.24503   1.887   0.0594 .
condition-C/-D   -0.35935    0.03511 2007.16928 -10.235   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Post-hoc analyses

- Post-hoc analyses might be necessary because lmers don't provide a comparison of every level against each other level of a factor

- For continuous variables, a post-hoc analysis may involve correlation testing

- For factors, we can use special packages such as *emmeans*

emmeans

- Computes estimated marginal means (EMMs) for specified factors or factor combinations in a linear model; and optionally, comparisons or contrasts among them

```
> m5.lmer <- lmer(log(ReadingTime) ~ capitalization * determiner + (1 | participant), data = psycholinguistics_data)

> emmeans(m5.lmer, pairwise ~ capitalization * determiner, adjust = "Tukey")
```

emmeans

> emmeans(m5.lmer, pairwise ~ capitalization * determiner, adjust = "Tukey")

```
$emmeans
 capitalization determiner emmean     SE   df lower.CL upper.CL
 nocap          nodet        5.62 0.0429 54.7     5.53     5.70
 cap            nodet        5.62 0.0436 58.1     5.53     5.70
 nocap          det          6.04 0.0420 50.4     5.95     6.12
 cap            det          5.98 0.0419 49.7     5.89     6.06

Degrees-of-freedom method: kenward-roger
Results are given on the log (not the response) scale.
Confidence level used: 0.95

$contrasts
 contrast                   estimate     SE   df t.ratio p.value
 nocap nodet - cap nodet -0.000471 0.0371 2007  -0.013  1.0000
 nocap nodet - nocap det -0.423531 0.0353 2007 -12.006  <.0001
 nocap nodet - cap det   -0.359354 0.0351 2007 -10.235  <.0001
 cap nodet - nocap det   -0.423061 0.0361 2009 -11.719  <.0001
 cap nodet - cap det     -0.358884 0.0359 2008  -9.987  <.0001
 nocap det - cap det      0.064177 0.0340 2006   1.887  0.2341
```

emmeans

> emmeans(m6.lmer, pairwise ~ condition, adjust = "Tukey")

```
$emmeans
 condition emmean     SE   df lower.CL upper.CL
 +C/+D       5.98 0.0419 49.7     5.89     6.06
 +C/-D       5.62 0.0436 58.1     5.53     5.70
 -C/+D       6.04 0.0420 50.4     5.95     6.12
 -C/-D       5.62 0.0429 54.7     5.53     5.70

Degrees-of-freedom method: kenward-roger
Results are given on the log (not the response) scale.
Confidence level used: 0.95

$contrasts
 contrast            estimate     SE   df t.ratio p.value
 (+C/+D) - (+C/-D)  0.358884 0.0359 2008   9.987  <.0001
 (+C/+D) - (-C/+D) -0.064177 0.0340 2006  -1.887  0.2341
 (+C/+D) - (-C/-D)  0.359354 0.0351 2007  10.235  <.0001
 (+C/-D) - (-C/+D) -0.423061 0.0361 2009 -11.719  <.0001
 (+C/-D) - (-C/-D)  0.000471 0.0371 2007   0.013  1.0000
 (-C/+D) - (-C/-D)  0.423531 0.0353 2007  12.006  <.0001
```

Post-hoc analyses

- For continuous variables, this makes little sense.

```
> emmeans(mn.lmer, pairwise ~ sequential_trial, adjust="Tukey")
```

- How could we further investigate the relation between sequential_trial and reading time?
  - Group the independent variable (often done in cognitive domains, e.g. high vs. low working memory)
  - Correlations (with different subsets of the data, when needed)
  - Visually presenting the data or model output

Post-hoc analyses

- Correlation

> cor.test(x,y)

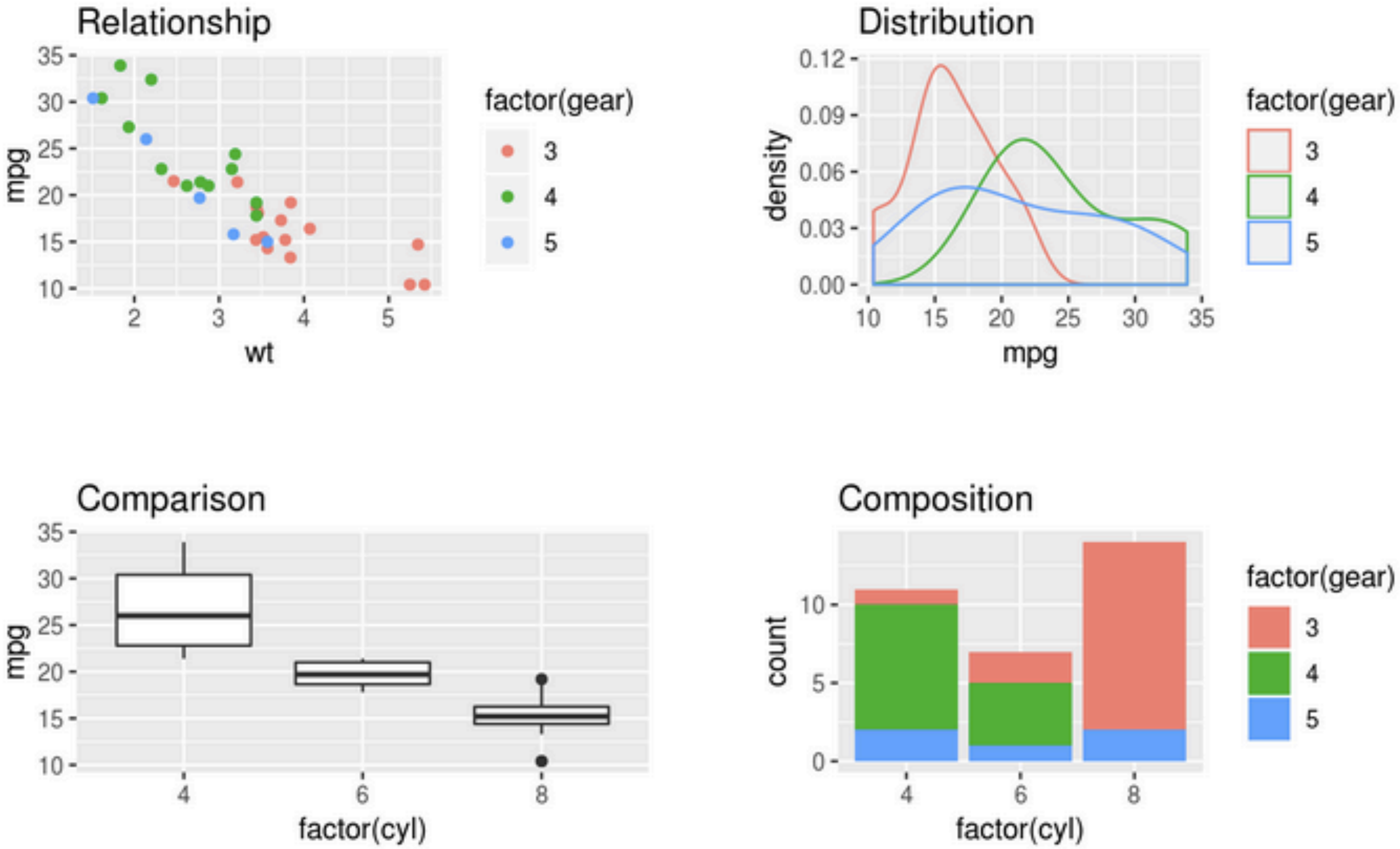
- Multiple types possible (pearson's, spearman's,…)
- However, this again ignores the issue of variability between participants and items

```
        Pearson's product-moment correlation

data:  psycholinguistics_data2$sequential_trial and psycholinguistics_data2$ReadingTime
t = 0.4664, df = 2037, p-value = 0.641
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.03309113  0.05371890
sample estimates:
       cor
0.01033335
```

Model visualization

- Data plots are easiest with the ggplot package



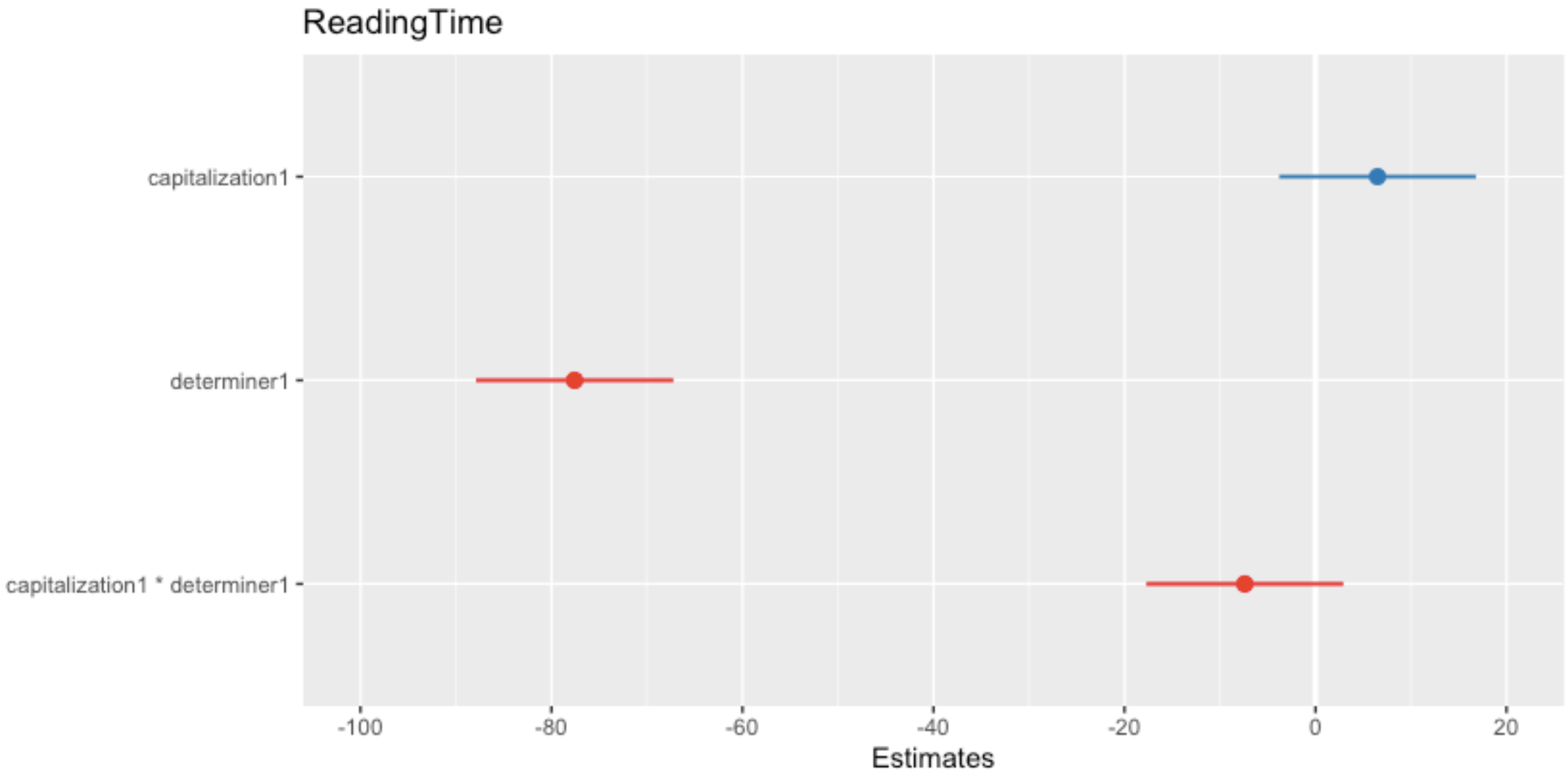
Model visualization

- However, *model* plots display the model effects, not the original data
- The can be used for presentation (in articles etc.) or for better understanding your results

```
> plot_model(m5.lmer) # from sjPlot package
```

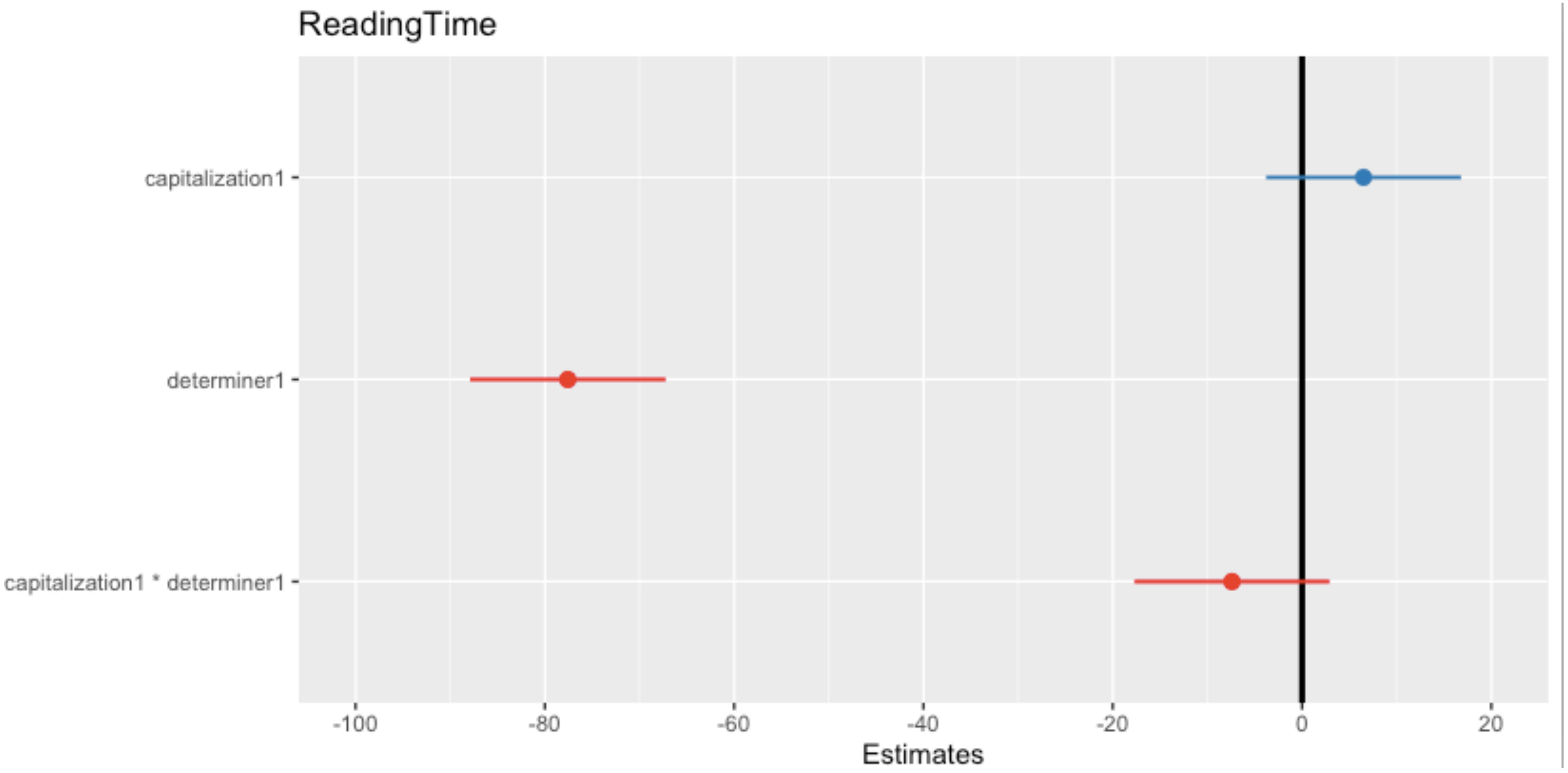
Model visualization

> plot_model(m5.lmer) # from sjPlot package



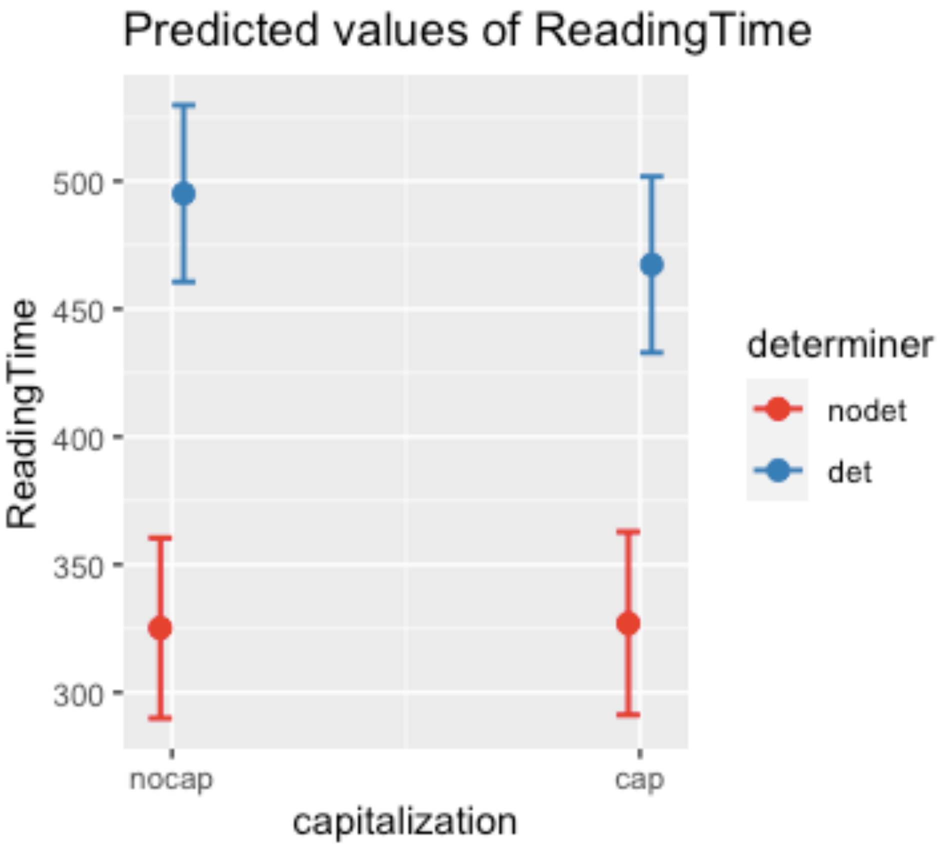
Model visualization

> plot_model(m5.lmer) # from sjPlot package



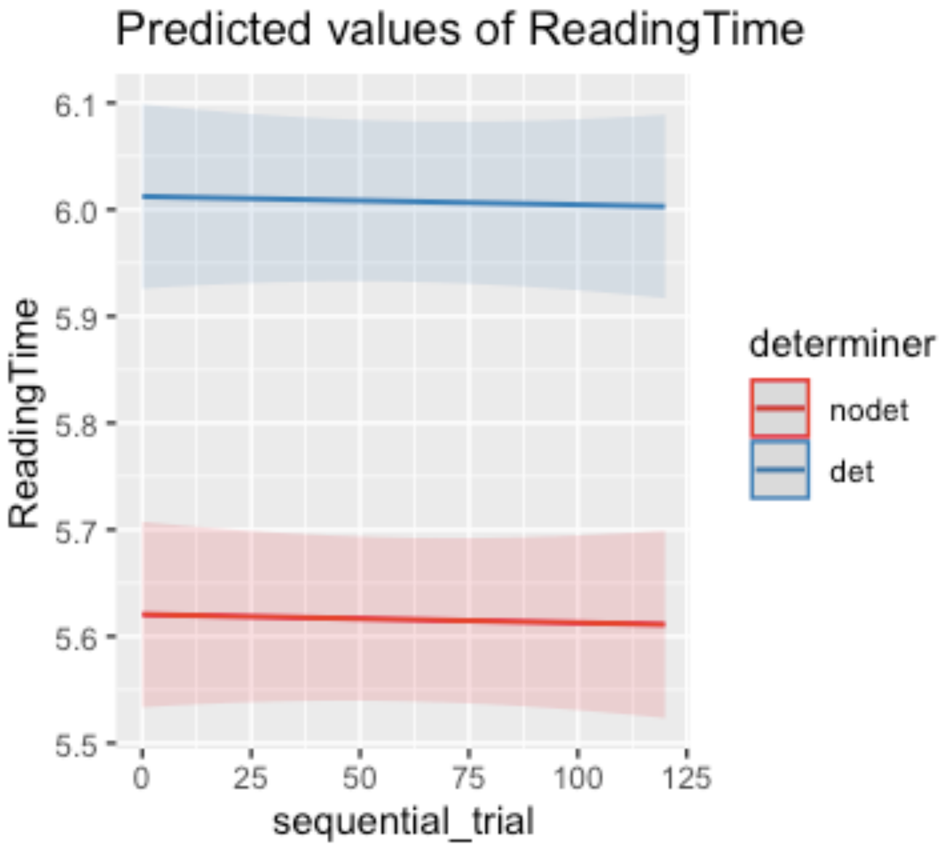
Model visualization

- > plot_model(m5.lmer, type = "eff", terms = c("capitalization", "determiner"))



Model visualization

- Also with one continuous variable
- > plot_model(m7.lmer, type = "eff", terms = c("sequential_trial", "determiner"))



Effect sizes

- Effect size is the size of the effect
- In our data, the effect of capitalization was 430-419 = 11 ms (dependent on outlier removal technique)
- Whether 11 ms is a lot or not depends on the variation in the data
- This is reflected in a relative or standardized effect size
- Popular measure of standardized effect size for comparisons between two means: Cohen's d

Effect sizes

- Popular measure of standardized effect size for comparisons between two means: Cohen's d

```
> lme.dscore(m5.lmer, psycholinguistics_data, type="lme4")
```

	t	df	d
capitalization1	1.5909857	2025.474	0.07070221
determiner1	-14.3795260	2027.232	-0.63873819
capitalization1:determiner1	-0.8180792	2025.402	-0.03635547

- This value will then have to be interpreted
  - Typically d = 0.2 represents a 'small' effect size, 0.5 represents a 'medium' effect size and 0.8 a 'large' effect size

Statistics for Linguists
08 July 2022

10:00	Workshop introduction
10:15	Loading and exploring datasets
10:45	Data transformation and coding
11:15	Practical exercise
12:15	Review of practical
12:30 - 13:30	LUNCH BREAK
13:30	lmer and glmer
14:30	Post-hoc analysis and model visualization
15:00	Practical exercise
16:00	Review of practical
16:15	Model building
17:00	End of workshop