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

lmer and glmer

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

Typical datasets/designs in linguistics

- In many of our studies, datasets have the following properties:
  - We collect data from a sample of participants => participant is a random variable (n.b. the mere fact that another researcher would have used another sample qualifies "participant" as a random variable)
  - We collect data from a sample of items => item is a random variable (n.b. the mere fact that another researcher would have used another sample qualifies "item" as a random variable)

Typical datasets/designs in linguistics

- So, participants and items are crossed random variables
  - We collect several measures in a given condition for each participant, on the different items
  - We collect several measures in a given condition for each item, by the different participants

Typical datasets/designs in linguistics

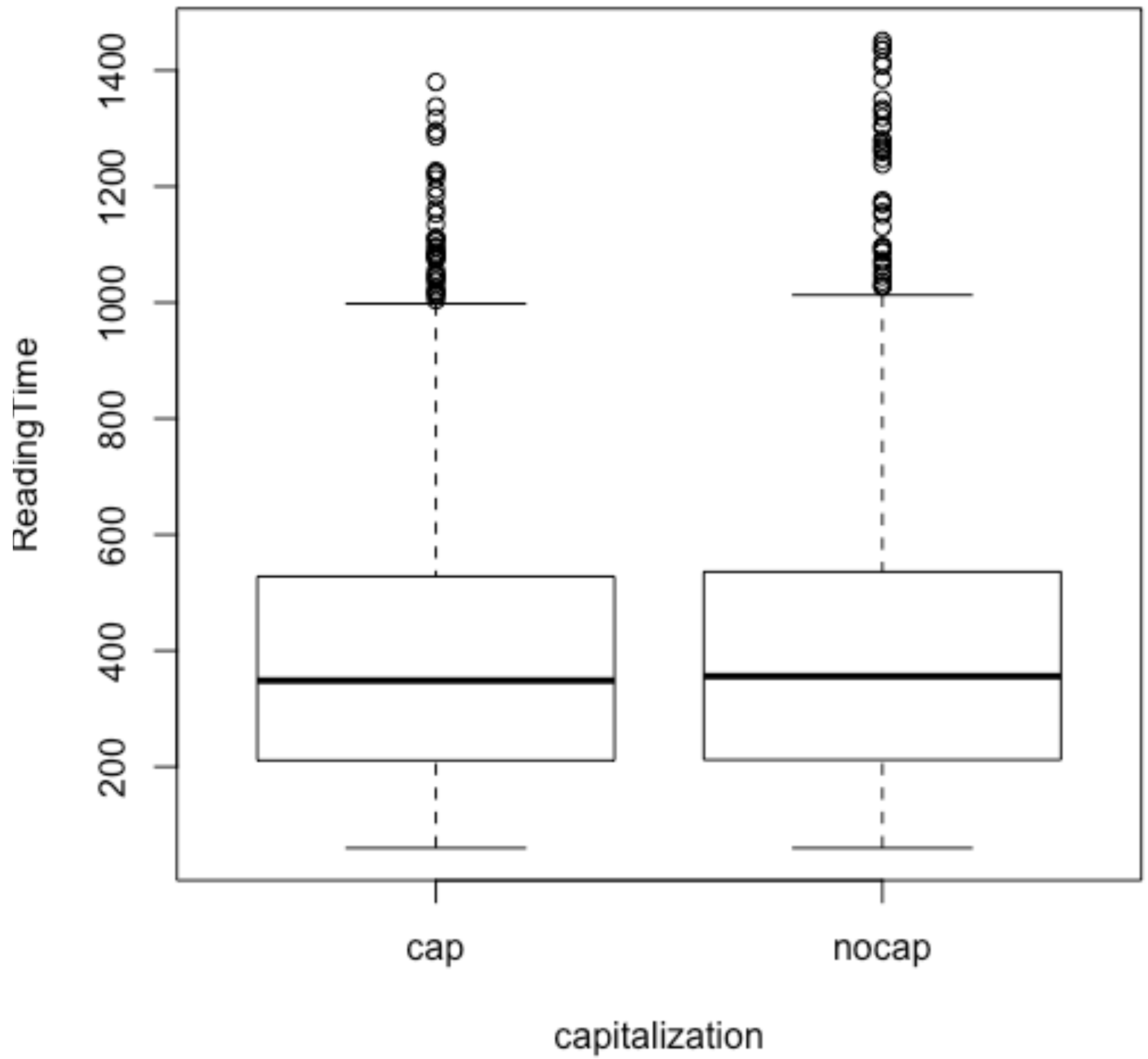
- Two important points:
  - We want to draw conclusions on not just these but other samples from the same population of participants / items
  - There is dependency in the data:
    - datapoints of a given participant are correlated (some participants are faster than others)
    - datapoints for a given item are correlated (some items are easier than others)

Illustration: our psycholinguistic experiment

Properties of the experiment/design:

- 30 participants. Participants selected amongst many => random variable
- Each participant read 120 German sentences (40 of these were fillers, which have already been removed)
- Dependent variable = reading time
- Independent variables = capitalization of the noun, presence of a determiner (categorical variables)
- Items = 20 sentences presented in each of the 4 conditions. Items selected amongst many => random variable
- *Hypothesis*: German nouns without capitalization are read slower.

Some descriptives



> tapply(psycholinguistics_data$ReadingTime,psycholinguistics_data$capitalization, mean)

cap	nocap
407.8897	415.4288

Hypothesis testing

- How can we test our hypothesis?
- Option 1: t-test
  But for this, data points must be independent => take average for cap and nocap items for each participant (This is called a by-participant analysis, also often called F1 analysis)

```
> tapply(psycholinguistics_data$ReadingTime,
         list(psycholinguistics_data$participant, psycholinguistics_data$capitalization), mean))
```

	cap	nocap
as08el22	466.1351	426.3611
au05rd24	415.0000	414.0811
ck06nk23	347.0000	384.8108
…		

Hypothesis testing

- How can we test our hypothesis?
- Option 1: t-test
  But for this, data points must be independent => take average for cap and nocap items for each participant (This is called a by-participant analysis, also often called F1 analysis)

```
> t.test(means.cap.pp[,1],means.cap.pp[,2], paired=TRUE)
```

```
        Paired t-test

data:  means.cap.pp[, 1] and means.cap.pp[, 2]
t = -1.4711, df = 29, p-value = 0.152
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -25.861172   4.222799
sample estimates:
mean of the differences
              -10.81919
```

Problems with this analysis

- By-participant analyses allow concluding that result would replicate with another sample of participants, but not with another sample of items
- The between item variance is not taken into account
- F1 is unconservative (high proportion of type I errors - rejecting the null hypothesis when it's actually true)
- There are missing data, and they are not missing at random
    - T-tests (and ANOVAs) do not deal well with missing data when these are not missing at random (e.g., Lachaud & Renaud, 2011)

Hypothesis testing

- How can we test our hypothesis?
- Option 2: mixed-effects regression
  - A single analysis that treats item and participant as crossed random variables
  - Can include covariates related to participant and item in the same analysis
  - Can include covariates at the single trial level
  - Deals well with missing data not at random

Mixed-effects Regression

- Results can be generalized across subjects and items
- Mixed-effects models are robust to missing data (Baayen, 2008, p. 266)
- Mixed-effects analysis is relatively easy to do and does not require a balanced design (which is generally necessary for repeated-measures ANOVA)
- We can easily test if it is necessary to treat item as a random effect

Mixed-effects Regression

- Lmer: Predict one numerical (dependent) variable on the basis of other, independent, variables (numerical or categorical)
    - (*Logistic* regression is used to predict a categorical dependent variable)
- We can write a regression formula as $y = I + ax_1 + bx_2 + ...$
- E.g., predict the reaction time of a participant on the basis of word frequency (WF), word length (WL) and speaker age (SA):
  $RT = 200 - 5WF + 3WL + 10SA$

Mixed-effects Regression

- Mixed-effects regression modeling distinguishes fixed-effects and random-effects factors
- Fixed-effects factors:
  - Repeatable levels -> depends on experiment design!
  - Small number of levels (e.g., Gender, Word Category, Capitalization)
- Random-effects factors:
  - Levels are a non-repeatable random sample from a larger population
  - Often large number of levels (e.g., Subject, Item)

What are random-effects factors?

- Random-effect factors are factors which are likely to introduce systematic variation
  - Some participants have a slow response (RT), while others are fast
    = Random Intercept for Subject
  - Some words are easy to recognize, others hard
    = Random Intercept for Item
  - The effect of word frequency on RT might be higher for one participant than another: non-native speakers might benefit more from frequent words than native speakers
    = Random Slope for Item Frequency per Subject
  - The effect of speaker age on RT might be different for one word than another: modern words might be recognized easier by younger speakers
    = Random Slope for Subject Age per Item
- Note that it is essential to test for random slopes!

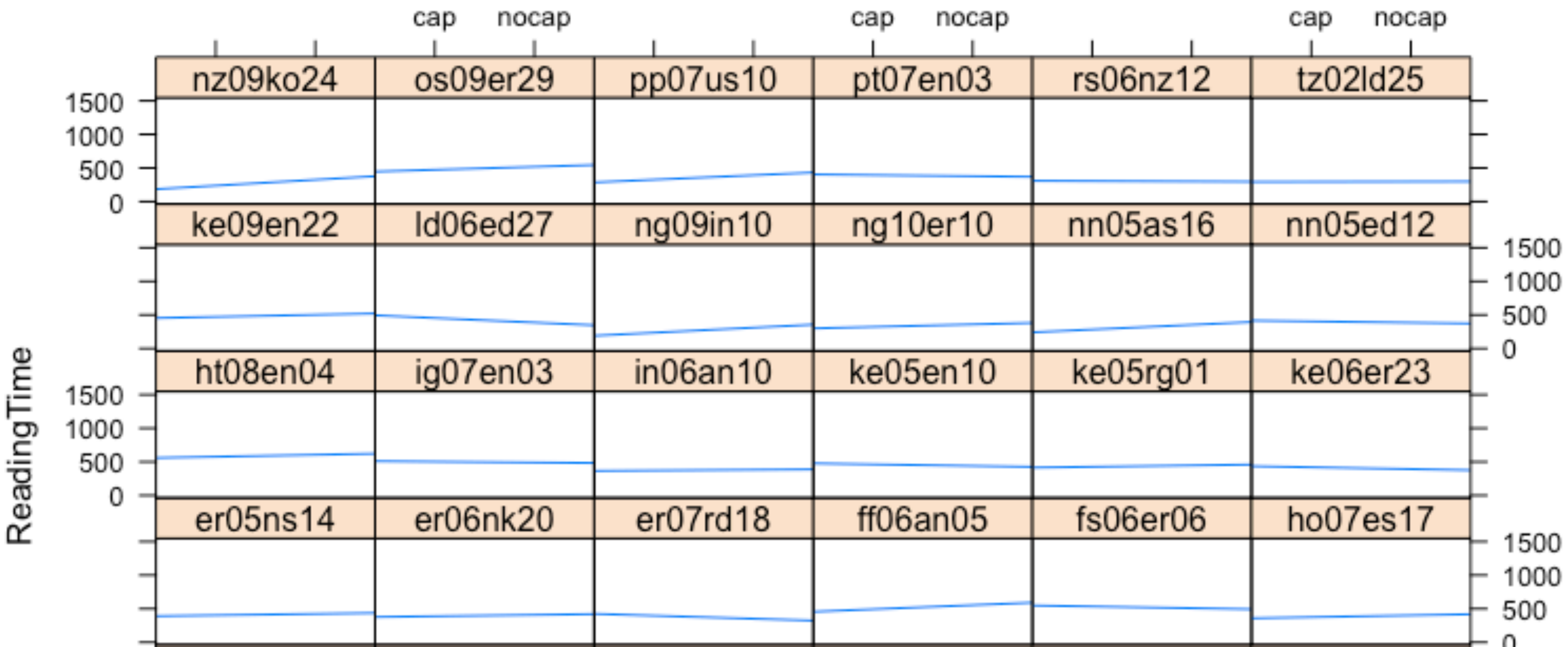
Mixed-effects Regression

- Mixed-effects regression analyses allow us to use random intercepts and slopes (i.e. adjustments to the population intercept and slopes) to make the regression formula as precise as possible for every individual observation in our random effects
  - Likelihood-ratio tests (comparing the goodness of fit of two statistical models) assess whether the inclusion of random intercepts and slopes is warranted
- Note that multiple observations for each level of a random effect are necessary for mixed-effects analysis to be useful (e.g., participants respond to multiple items)

Mixed-effects Regression: our data

- We can visualize the between-participant variance

> print(xyplot(ReadingTime~capitalization|participant, panel=function(x,y,...) {panel.xyplot(x,y,type="r")},psycholinguistics_data)) # function from lattice package



Mixed-effects Regression: our data

- Given these differences between participants, we could use participants as a random effect

```
> m0.lmer <- lmer(log(ReadingTime) ~ capitalization + (1|participant), data = psycholinguistics_data)
> summary(m0.lmer)
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: log(ReadingTime) ~ capitalization + (1 | participant)
   Data: psycholinguistics_data2

REML criterion at convergence: 3759.4

Scaled residuals: 
     Min       1Q   Median       3Q      Max 
-3.09115 -0.74477  0.06081  0.68571  2.50575 

Random effects:
 Groups      Name        Variance Std.Dev.
 participant (Intercept) 0.03092  0.1758  
 Residual                0.35826  0.5985  
Number of obs: 2039, groups:  participant, 30

Fixed effects:
                     Estimate Std. Error        df t value Pr(>|t|)    
(Intercept)         5.816e+00  3.726e-02 3.806e+01 156.112   <2e-16 ***
capitalizationnocap 2.413e-02  2.653e-02 2.009e+03   0.909    0.363    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
captlztnncp -0.361
```

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: log(ReadingTime) ~ capitalization + (1 | participant)
   Data: psycholinguistics_data2

REML criterion at convergence: 3759.4

Scaled residuals:
     Min       1Q   Median       3Q      Max
-3.09115 -0.74477  0.06081  0.68571  2.50575

Random effects:
 Groups      Name        Variance Std.Dev.
 participant (Intercept) 0.03092  0.1758
 Residual                0.35826  0.5985
Number of obs: 2039, groups:  participant, 30

Fixed effects:
                     Estimate Std. Error        df t value Pr(>|t|)
(Intercept)         5.816e+00  3.726e-02 3.806e+01 156.112   <2e-16 ***
capitalizationnocap 2.413e-02  2.653e-02 2.009e+03   0.909    0.363
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
captlztnncp -0.361
```

- We can view the predicted intercepts for each participant

\> ranef(m0.lmer)

```
$participant
          (Intercept)
as08el22  0.051473158
au05rd24 -0.001025549
ck06nk23 -0.085420764
de10ch10 -0.019799498
en04do16  0.288505901
en04el04 -0.044413641
er05ns14 -0.081226263
er06nk20  0.064283699
er07rd18 -0.030783260
ff06an05  0.213551867
fs06er06  0.199338539
```

```
REML criterion at convergence: 28327.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8933 -0.7290 -0.2069  0.4522  4.0571

Random effects:
 Groups      Name        Variance Std.Dev.
 participant (Intercept)  5513     74.25
 Residual                61898    248.79
Number of obs: 2039, groups:  participant, 30

Fixed effects:
                Estimate Std. Error       df t value Pr(>|t|)
(Intercept)      410.325     14.641   28.828  28.027   <2e-16 ***
capitalization1    5.050      5.514 2008.772   0.916     0.36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
capitalztn1 -0.005
```

Specific 'model' for every observation

- $RT = 410 - 5cap$ (general model)
  - The intercepts and slopes may vary (according to the estimated standard variation for each parameter) and this influences the item- and participant-specific values
- $RT = 300 - 5cap$ (subject: fast)

Some notes/issues

- All models are wrong
  - Some models are better than others
  - The correct model can never be known with certainty
  - The simpler the model, the better it is

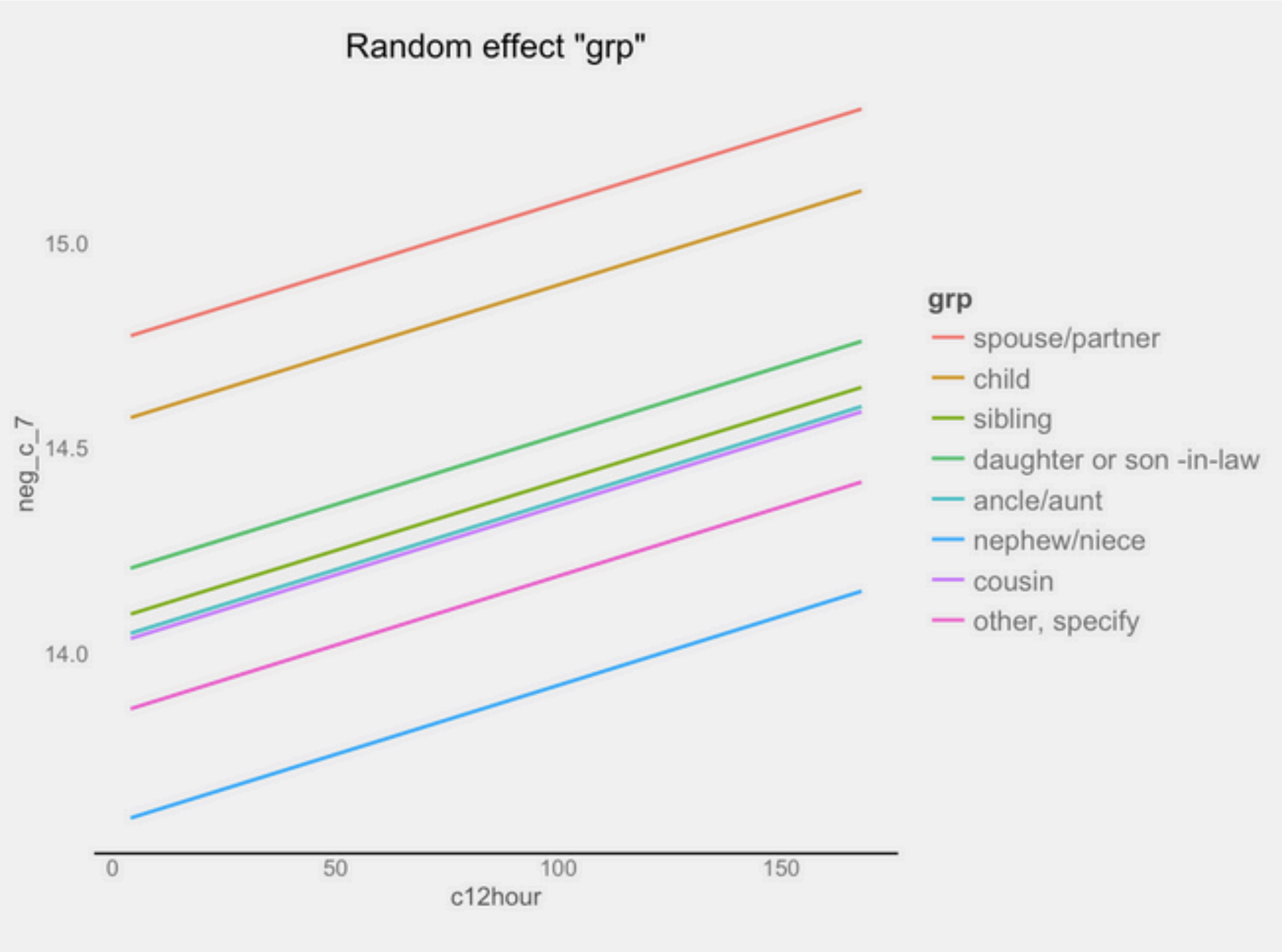
- Assumptions about the predictors
  - We assume linearity

- Assumptions about the residuals
  - The errors should follow a normal distribution
  - Residuals should be independent
  - Check the distribution of residuals: if not normally distributed then transform dependent variable
  - Important: no *a priori* exclusion of outliers without a clear reason
    - A good reason is not necessarily that the value is over 2.5 SD above the mean
    - A good reason (e.g.) may be that the response is faster than possible

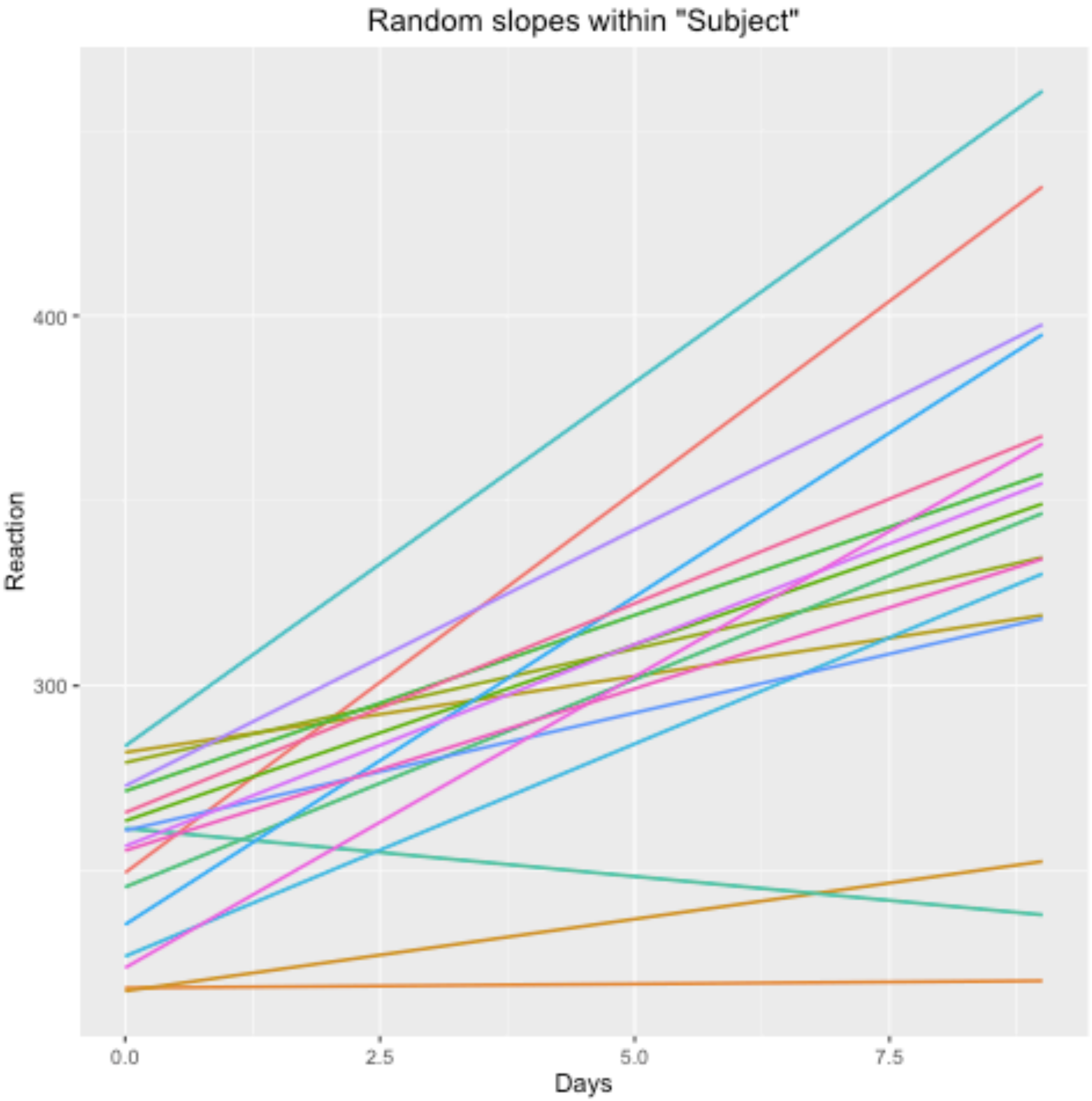
Let’s expand our model

- We can (and often should) complexify the random part of the model
- We can have lmer fit different intercepts AND slopes for each participant

```
> m1.lmer <- lmer(log(ReadingTime) ~ capitalization + (capitalization +
                  1|participant), data = psycholinguistics_data)
> summary(m1.lmer)
```



Random effect "grp"
15.0
14.5
14.0
neg_c_7
0
50
100
150
c12hour
grp
spouse/partner
child
sibling
daughter or son -in-law
ancle/aunt
nephew/niece
cousin
other, specify


Random slopes within "Subject"
400
300
Reaction
0.0
2.5
5.0
7.5
Days

Let’s expand our model

> m1.lmer <- lmer(log(ReadingTime) ~ capitalization + (capitalization + 1|participant), data = psycholinguistics_data)
> summary(m1.lmer)

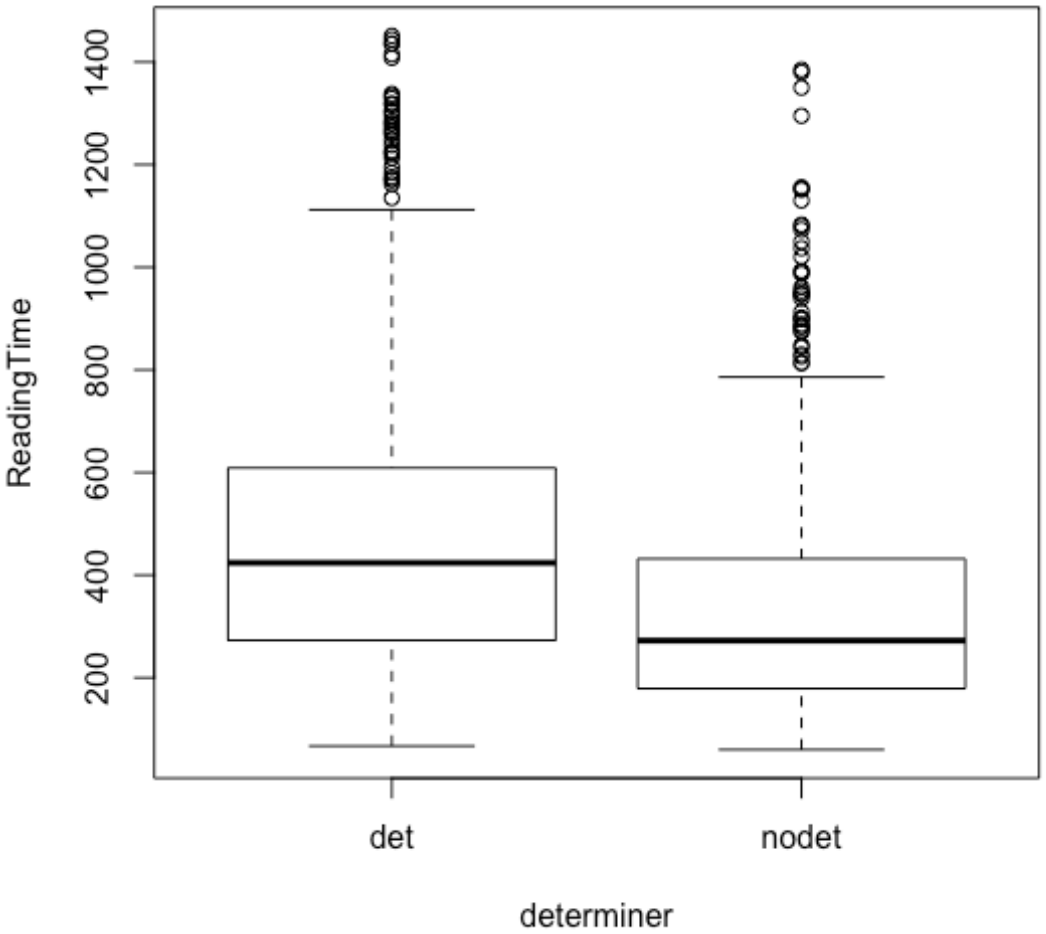
```
Scaled residuals:
     Min       1Q   Median       3Q      Max
-3.09329 -0.74046  0.05996  0.69040  2.58075

Random effects:
 Groups      Name            Variance Std.Dev. Corr
 participant (Intercept)     0.031141 0.1765
             capitalization1 0.000625 0.0250   -1.00
 Residual                    0.357613 0.5980
Number of obs: 2039, groups:  participant, 30

Fixed effects:
                 Estimate Std. Error        df t value Pr(>|t|)
(Intercept)       5.82798    0.03485  28.82516 167.222   <2e-16 ***
capitalization1   0.01268    0.01402 123.03953   0.905    0.367
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
capitalztn1 -0.306
```

- As can be noticed in the output of the model, we have now a value for:
- The SD of the varying participant intercept
- The SD of the varying participant slope
- A correlation parameter. This is the correlation between the participant intercept and slope.
- We can remove it with the following syntax: 1 + capitalization || participant (you will learn about how to select the optimal random effect structure in later lectures)

```
Scaled residuals:
     Min       1Q   Median       3Q      Max 
-3.09329 -0.74046  0.05996  0.69040  2.58075 

Random effects:
 Groups      Name            Variance Std.Dev. Corr 
 participant (Intercept)     0.031141 0.1765        
             capitalization1 0.000625 0.0250   -1.00
 Residual                    0.357613 0.5980        
Number of obs: 2039, groups:  participant, 30

Fixed effects:
                 Estimate Std. Error        df t value Pr(>|t|)    
(Intercept)       5.82798    0.03485  28.82516 167.222   <2e-16 ***
capitalization1   0.01268    0.01402 123.03953   0.905    0.367    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
            (Intr)
capitalztn1 -0.306
```

Let’s expand our model

- Presence of a determiner was also manipulated in the experiment



> m3.lmer <- lmer(log(ReadingTime) ~ capitalization * determiner + (1 + capitalization || participant), data = psycholinguistics_data)

> summary(m3.lmer)

```
Fixed effects:
                             Estimate Std. Error         df t value Pr(>|t|)    
(Intercept)                   5.81144    0.03622   28.07018 160.457   <2e-16 ***
capitalization1               0.01607    0.01323   32.57075   1.215    0.233    
determiner1                  -0.19566    0.01260 1996.53365 -15.533   <2e-16 ***
capitalization1:determiner1  -0.01598    0.01257 2002.41700  -1.272    0.204    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- We now have a model with an effect of determiner
- Note that we need to be careful about how to interpret this effect, it depends on what the reference level/baseline is!

```
Fixed effects:
                              Estimate Std. Error         df t value Pr(>|t|)    
(Intercept)                    5.81144    0.03622   28.07018 160.457   <2e-16 ***
capitalization1                0.01607    0.01323   32.57075   1.215    0.233    
determiner1                   -0.19566    0.01260 1996.53365 -15.533   <2e-16 ***
capitalization1:determiner1   -0.01598    0.01257 2002.41700  -1.272    0.204    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tips and tricks

- Package sjPlot, function tab_model(): creates HTML-table of the model output, including R^2

	ReadingTime		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	410.32	381.63 – 439.02	**<0.001**
capitalization1	5.05	-5.76 – 15.86	0.360
Random Effects			
σ^2	61897.96		
$\tau_{00\ participant}$	5513.01		
ICC	0.08		
$N_{participant}$	30		
Observations	2039		
Marginal R^2 / Conditional R^2	0.000 / 0.082		

Tips and tricks

- PCA of random-effects covariance matrix

> summary(rePCA(m1.lmer))

```
$participant
Importance of components:
                        [,1] [,2]
Standard deviation     0.298    0
Proportion of Variance 1.000    0
Cumulative Proportion  1.000    1
```

- View random effects without viewing fixed effects

> VarCorr(m1.lmer)

```
Groups      Name            Std.Dev. Corr
participant (Intercept)     0.17647
            capitalization1 0.02500  -1.000
Residual                    0.59801
```

Tips and tricks

- If you're only interested in the effect of one variable within levels of the other variable, you could use a nested model (/ syntax)

> m4.lmer <- lmer(log(ReadingTime) ~ capitalization / determiner + (1 + capitalization || participant), data = psycholinguistics_data)

> summary(m4.lmer)

```
Fixed effects:
                                 Estimate Std. Error         df t value Pr(>|t|)
(Intercept)                       6.00710    0.03798   33.92594 158.169   <2e-16 ***
capitalization1                   0.03205    0.01749   98.40922   1.833   0.0699 .
capitalizationnocap:determinernodet -0.42329  0.03526 1995.72317 -12.005   <2e-16 ***
capitalizationcap:determinernodet -0.35935    0.03592 2003.40444 -10.004   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tips and tricks

- You can select an optimizer for your model, which may help it to converge. You can just add this to the end of your model

  m5.lmer <- lmer(log(ReadingTime) ~ capitalization * determiner + (1 + capitalization || participant), data = psycholinguistics_data, **control = glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 25000**)))

Some more notes/issues

- Very dependent on user input
- Various approaches being used (keeping it maximal, minimal)
- Results may change according to used package; difficult to see what happens ’behind the scenes’
- The lme4 package is still under development. Results with newer versions may differ slightly

glmer

- Used for fitting Generalized Linear Mixed-Effects Models
- Similar in lmer when it comes to syntax
- You can set the ’family’ (distribution), which can be binomial (when predicting a factor), but also gaussian, poisson etc.

Conclusion

- Mixed-effects regression is more flexible than using ANOVAs
- Testing for inclusion of random intercepts and slopes is essential when you have multiple responses per subject or item
- Mixed-effects regression is relatively easy with (g)lmer in R

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