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

Model building

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

Model building

- Various approaches being used (keeping it maximal, minimal)

- Build up models one factor at a time
  - lmer(ReadingTime ~ capitalization)
  - Interactions + main effects: determiner * capitalization
  - Interaction only (no main effects): determiner : capitalization
  - intercept: (1 | randEf)
  - Intercept + slope: (1 + fixedEf| randEf)
- Compare models based on the AIC

Model building

One common approach:

- Fit maximal model

```
> mm <- lmer(DV ~ Factor + (Factor | Subj) + (Factor | Item), data=data,
REML=FALSE)
```

```
> print(summary(lmm), corr=FALSE)
```

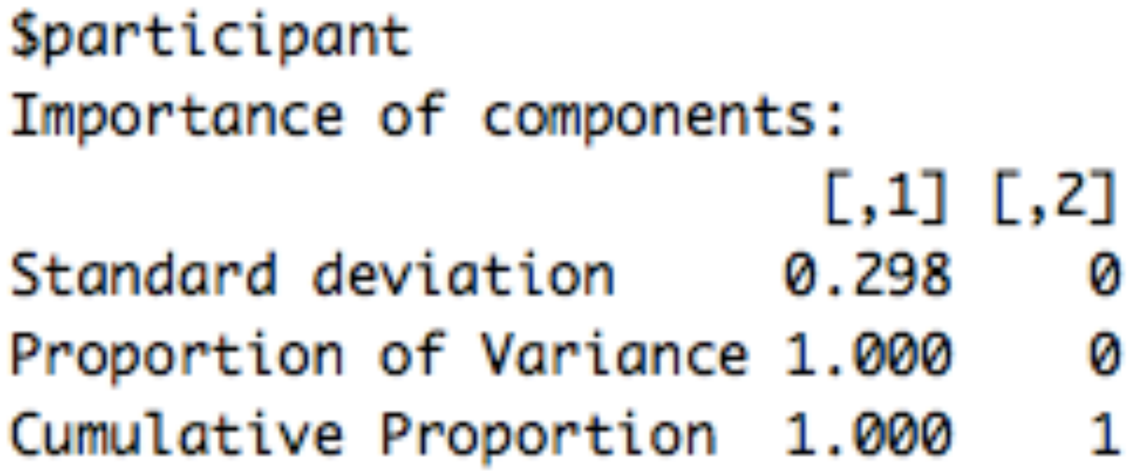
- Check random effects structure

```
> summary(rePCA(lmm))
```

Model building

- Check random effects structure

> summary(rePCA(lmm))



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

Model building

- Check random effects structure

```
> summary(rePCA(lmm))
```

- If proportion of variance explained is non-zero for all principal components (PCs), both for subject-related and for item-related PCs, you can likely keep them.
- If some are zero, remove them
- Convergence problems mean that the model is not supported by the data

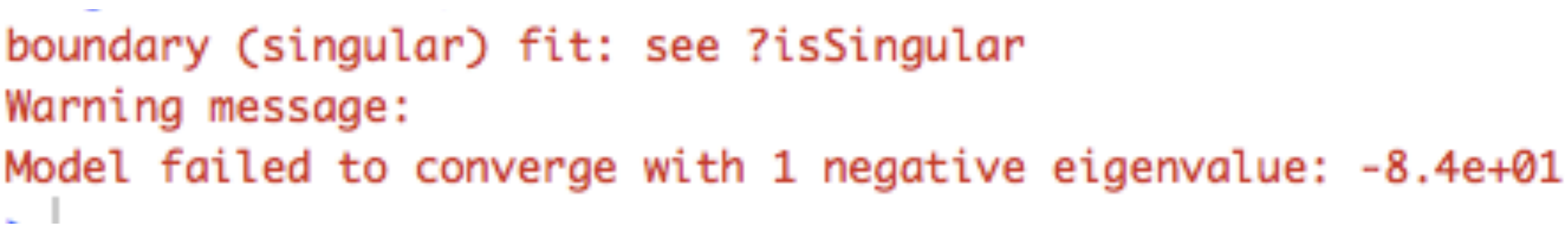
Model building

- Convergence problems mean that the model is not supported by the data
- When you obtain a singular fit, this is often indicating that the model is overfitted

Model building

- Convergence problems mean that the model is not supported by the data
- When you obtain a singular fit, this is often indicating that the model is overfitted

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

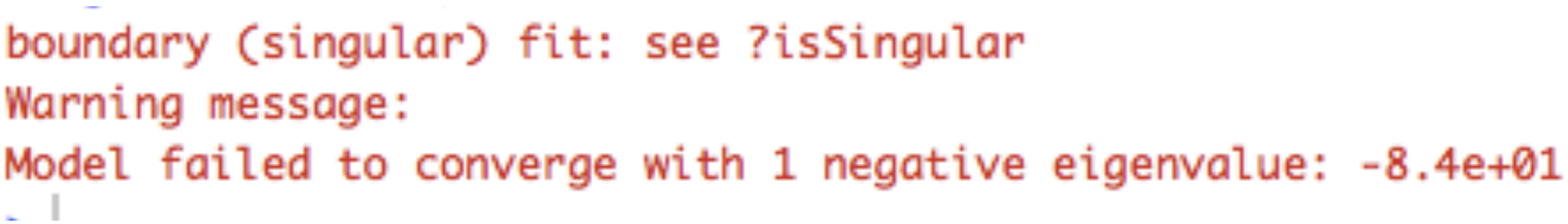


```
boundary (singular) fit: see ?isSingular
Warning message:
Model failed to converge with 1 negative eigenvalue: -8.4e+01
```

Model building

- Convergence problems mean that the model is not supported by the data
- When you obtain a singular fit, this is often indicating that the model is overfitted

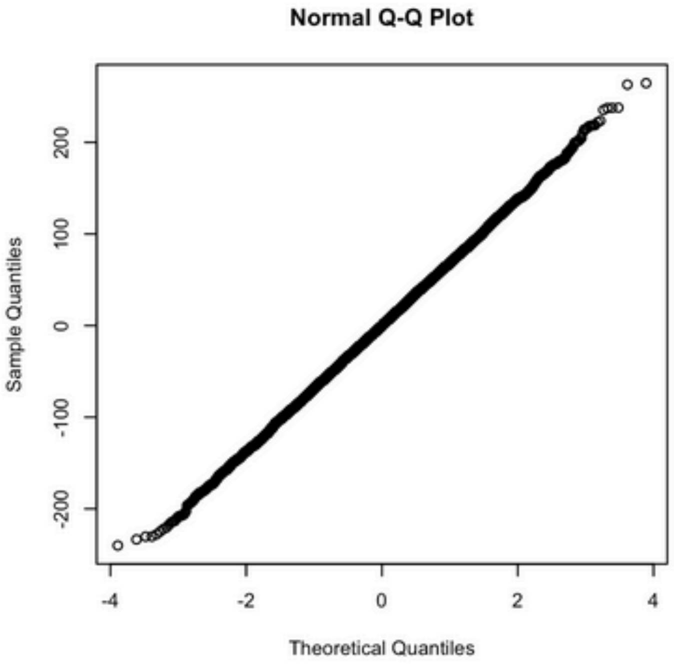
```
> m3.lmer <- lmer(log(ReadingTime) ~ capitalization * determiner + (1 + capitalization | participant), data = psycholinguistics_data)
```

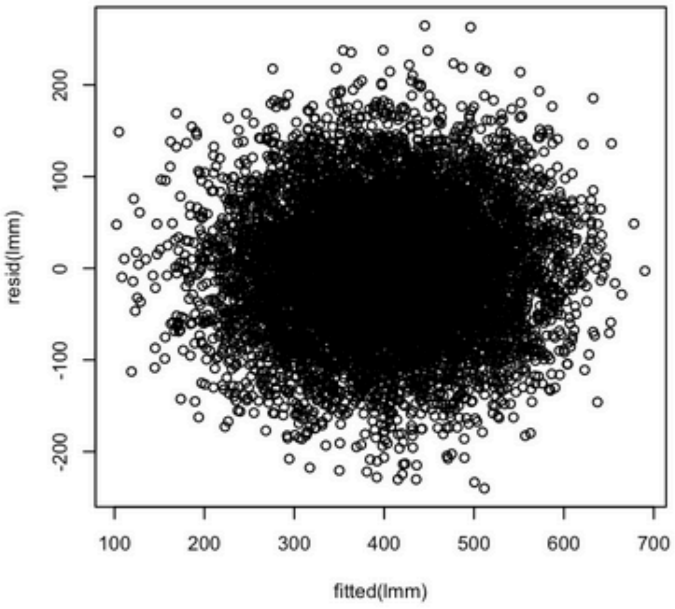


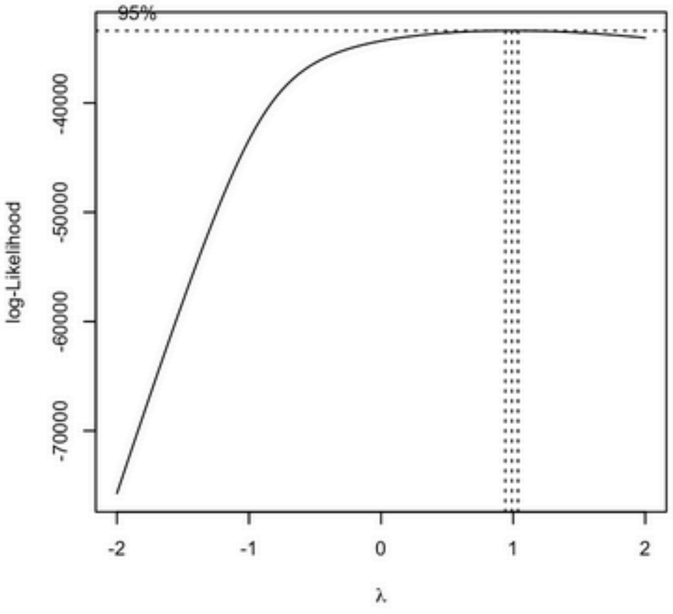
```
boundary (singular) fit: see ?isSingular
Warning message:
Model failed to converge with 1 negative eigenvalue: -8.4e+01
```

Check residuals

```
> qqnorm(resid(lmm))
> plot(fitted(lmm), resid(lmm))
> boxcox(DV ~ Factor, data=data)
```

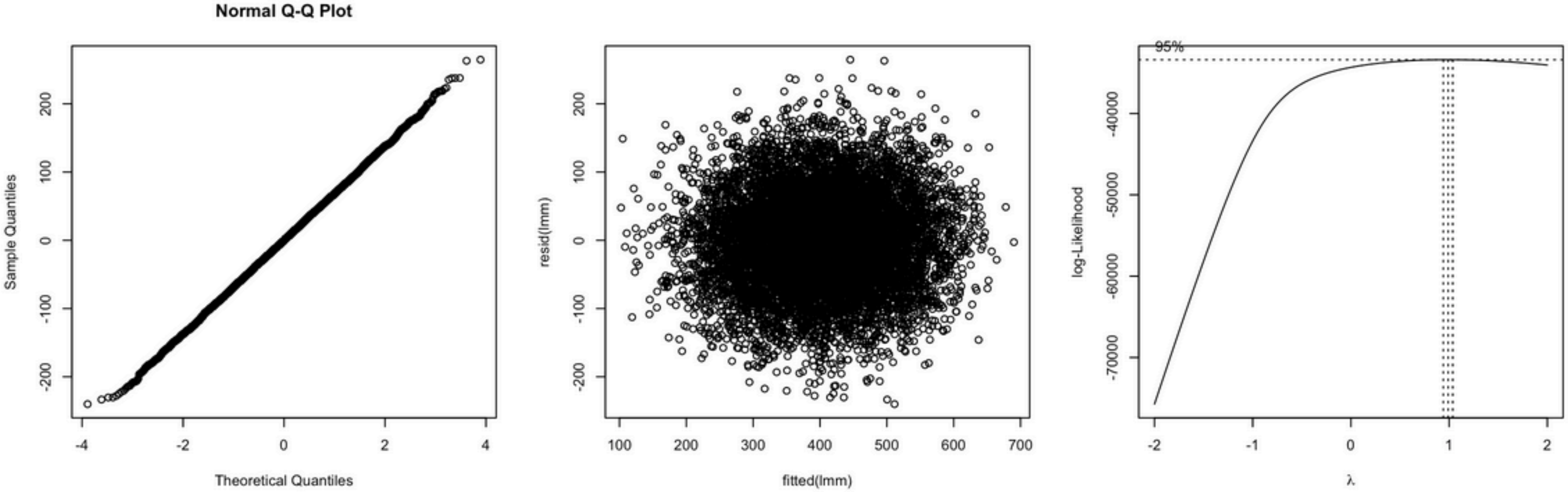






Check residuals

- The figure shows that the residuals do not strongly deviate from a normal distribution (left panel) and that the variance of the residuals is similar across different fitted values (homoscedasticity, middle panel). Power transformations (right panel) further show no transformation is needed to better approximate a normal distribution.



Model building

- When you’ve determined the maximal random effects structure, you can start building your fixed effects, step by step
    - You can build up: start with an empty model, add factors and test if they contribute
    - Or build down: start with a full model and remove factors that don’t contribute
    - And/Or keep some fixed effects based on your hypotheses

Model building

- Your model building choices may depend on your hypothesis and the amount of variables you have

- Example 1: we want to know the effect of *condition.* We can add it and see if it reaches significance. If not, we leave it in and report it
  - Advantage: there is a p-value to report, as the factor is still in the model
  - Disadvantage: leaving factors that don’t contribute increases chances of overfitting

Model building

- Your model building choices may depend on your hypothesis and the amount of variables you have

- Example 2: we want to which of 10 cognitive abilities or socio—economic background factors, and language history measures contribute to reading time
    - Adding all factors and interactions is too much for the model; it does not converge
    - We could add factors stepwise
    - Advantage: probably the only feasible approach, leads to a well-fitting model
    - Disadvantage: no estimates to report for factors that didn't make the cut

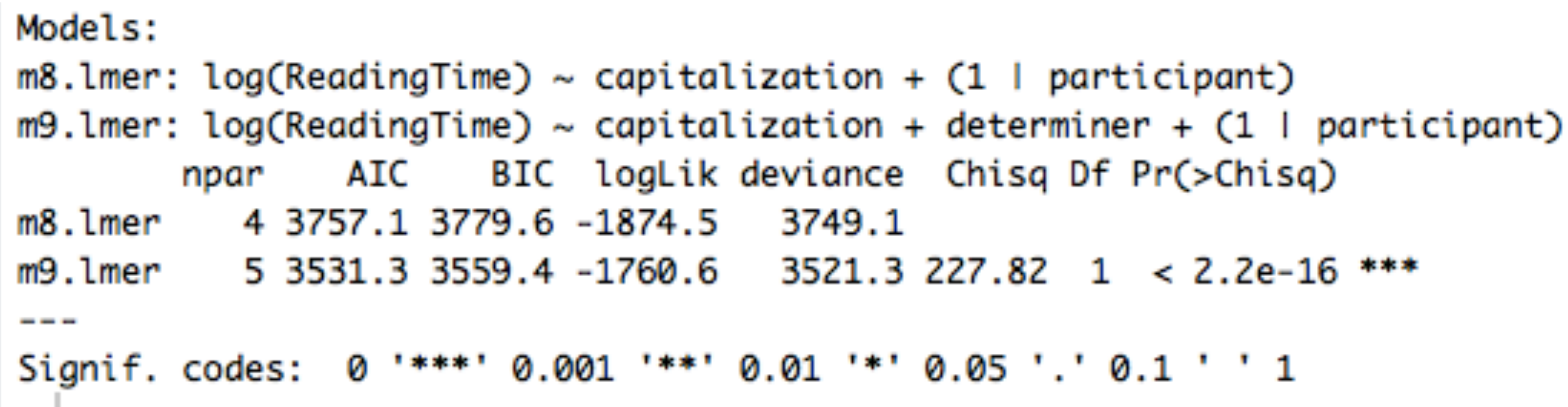
Model comparisons

- How do we determine whether a factor ’contributes’?
- This is not based on significance p-value
- Should be based on model comparison. An easy-ish way is with anova()

Model comparisons

> m8.lmer <- lmer(log(ReadingTime) ~ capitalization + (1 | participant),
>   data = psycholinguistics_data2)
> m9.lmer <- lmer(log(ReadingTime) ~ capitalization + determiner + (1 |
>   participant), data = psycholinguistics_data2)

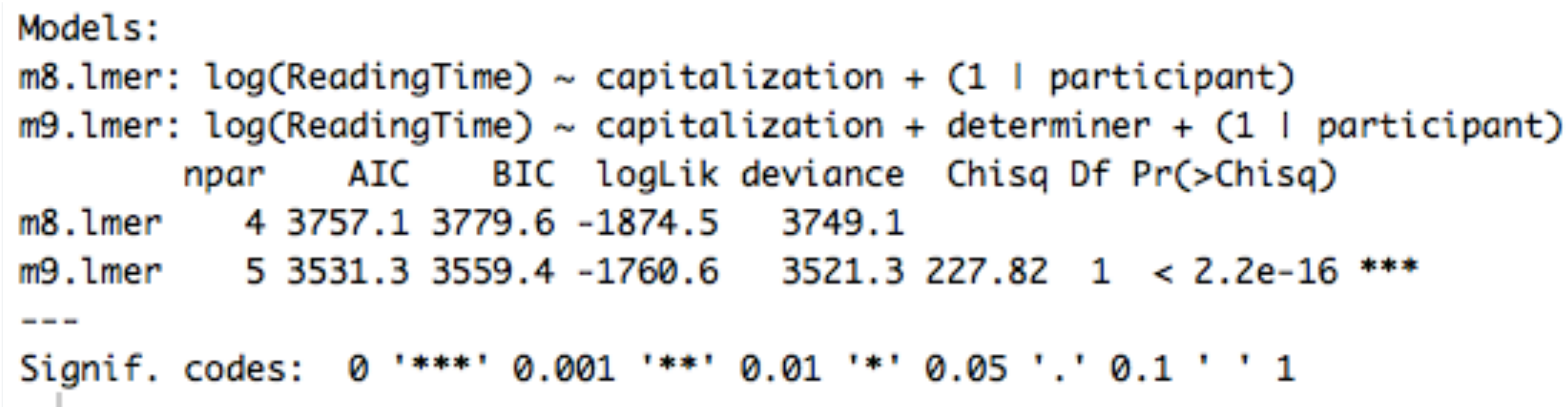
> anova(m8.lmer, m9.lmer)



```
Models:
m8.lmer: log(ReadingTime) ~ capitalization + (1 | participant)
m9.lmer: log(ReadingTime) ~ capitalization + determiner + (1 | participant)
        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
m8.lmer    4 3757.1 3779.6 -1874.5   3749.1
m9.lmer    5 3531.3 3559.4 -1760.6   3521.3 227.82  1  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model comparisons

- Be careful to only add one factor/ take on step at a time
- Look at the AIC or BIC values
  - Lower is better
  - Rule of thumb: a difference of 2 points makes a better model
  - BIC corrects more strictly for the number of parameters than AIC

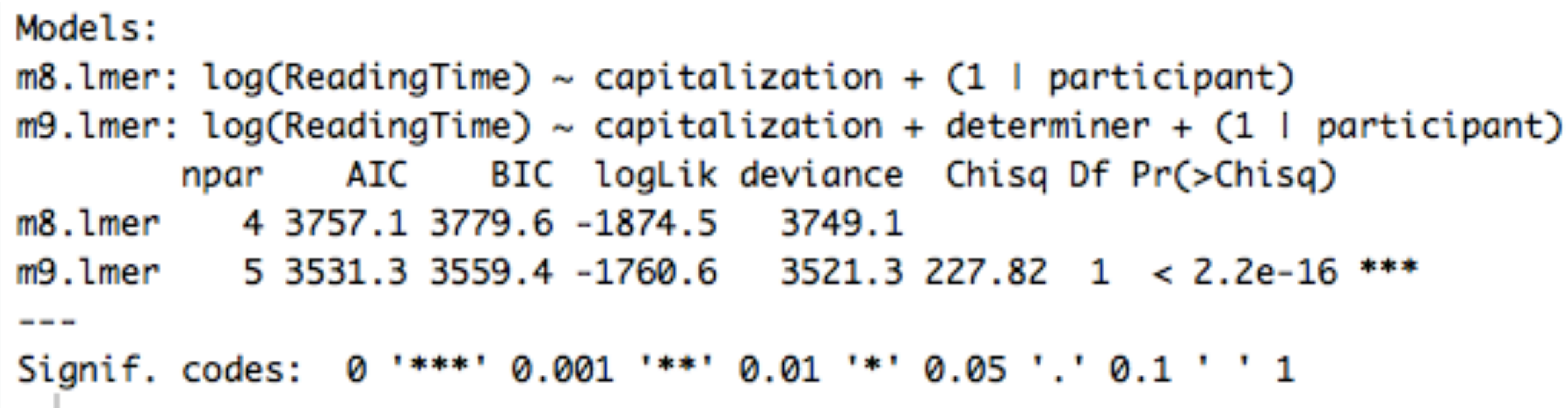


```
Models:
m8.lmer: log(ReadingTime) ~ capitalization + (1 | participant)
m9.lmer: log(ReadingTime) ~ capitalization + determiner + (1 | participant)
        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
m8.lmer    4 3757.1 3779.6 -1874.5   3749.1
m9.lmer    5 3531.3 3559.4 -1760.6   3521.3 227.82  1  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model comparisons

- You can add as many models as you like to anova()

> anova(m8.lmer, m9.lmer, m10.lmer, m11.lmer, …)



```
Models:
m8.lmer: log(ReadingTime) ~ capitalization + (1 | participant)
m9.lmer: log(ReadingTime) ~ capitalization + determiner + (1 | participant)
        npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
m8.lmer    4 3757.1 3779.6 -1874.5   3749.1
m9.lmer    5 3531.3 3559.4 -1760.6   3521.3 227.82  1  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Strategy for determining a parsimonious model

Here is sketch of recommendation about how to proceed to determine a parsimonious model. Note that theoretical expectations about parameters may lead to different sequences

1. Perform a PCA on the maximal model: summary(rePCA(max_lmm)). If the maximal model doesn't converge, then proceed to step (3).
2. If all principle component(s) explain non-zero variance, use the maximal model
3. If one (or more) PC explain zero variance, fit a zero correlation parameter model
4. Remove predictor with smallest variance of random slopes. E.g.: lmm <- lmer(DV ~ 1 + c1 + c3 + (1 + c1 + c3 || id), data=dat)
5. Add correlation parameter to this reduced model. E.g., lmm <- lmer(DV ~ 1 + c1 + c3 + (1 + c1 + c3 | id), data=dat)
6. Perform a PCA to check whether all PC explain non-zero variance. If not, repeat steps (3-6). Otherwise this is the parsimonious model.

Alternative / additional criterion: Use model comparison based on AIC, BIC, or (for nested models) the log likelihood ratio test for model selection

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