Introduction to mixed-effects regression for (psycho)linguists

Martijn Wieling

Department of Humanities Computing, University of Groningen

Groningen, April 21, 2015

Overview

- Introduction
- Recap: multiple regression
- Mixed-effects regression analysis: explanation
- Methodological issues
- Case-study: Lexical decision latencies (Baayen, 2008: 7.5.1)
- Conclusion

Introduction

- Consider the following situation (taken from Clark, 1973):
  - Mr. A and Mrs. B study reading latencies of verbs and nouns
  - Each randomly selects 20 words and tests 50 participants
  - Mr. A finds (using a sign test) verbs to have faster responses
  - Mrs. B finds nouns to have faster responses
- How is this possible?

Introduction

- Consider the following situation (taken from Clark, 1973):
  - Mr. A and Mrs. B study reading latencies of verbs and nouns
  - Each randomly selects 20 words and tests 50 participants
  - Mr. A finds (using a sign test) verbs to have faster responses
  - Mrs. B finds nouns to have faster responses
- How is this possible?

The language-as-fixed-effect fallacy

- The problem is that Mr. A and Mrs. B disregard the variability in the words (which is huge)
  - Mr. A included a difficult noun, but Mrs. B included a difficult verb
  - Their set of words does not constitute the complete population of nouns and verbs, therefore their results are limited to their words
- This is known as the language-as-fixed-effect fallacy (LAFEF)
  - Fixed-effect factors have repeatable and a small number of levels
  - Word is a random-effect factor (a non-repeatable random sample from a larger population)

Why linguists are not always good statisticians

- LAFEF occurs frequently in linguistic research until the 1970's
  - Many reported significant results are wrong (the method is anti-conservative)!
- Clark (1973) combined a by-subject (F_1) analysis and by-item (F_2) analysis in a measure called *min F'*
  - Results are significant and generalizable across subjects and items when *min F'* is significant
  - Unfortunately many researchers (>50%!) incorrectly interpreted this study and may report wrong results (Raaijmakers et al., 1999)
  - E.g., they only use F_1 and F_2 and not *min F'* or they use F_2 while unneccesary (e.g., counterbalanced design)

Our problems solved...

- Apparently, analyzing this type of data is difficult...
- Fortunately, using mixed-effects regression models solves these problems!
  - The method is easier than using the approach of Clark (1973)
  - Results can be generalized across subjects and items
  - Mixed-effects models are robust to missing data (Baayen, 2008, p. 266)
  - We can easily test if it is necessary to treat item as a random effect
- But first some words about regression...

Our problems solved...

- Apparently, analyzing this type of data is difficult...
- Fortunately, using mixed-effects regression models solves these problems!
  - The method is easier than using the approach of Clark (1973)
  - Results can be generalized across subjects and items
  - Mixed-effects models are robust to missing data (Baayen, 2008, p. 266)
  - We can easily test if it is necessary to treat item as a random effect
- But first some words about regression...

Regression vs. ANOVA

- Most people either use ANOVA or regression
  - ANOVA: categorical predictor variables
  - Regression: continuous predictor variables
- Both can be used for the same thing!
  - ANCOVA: continuous and categorical predictors
  - Regression: categorical (dummy coding) and continuous predictors
- Why I use regression as opposed to ANOVA
  - No temptation to dichotomize continuous predictors
  - Intuitive interpretation (your mileage may vary)
  - Mixed-effects analysis is relatively easy to do and does not require a balanced design (which is generally necessary for repeated-measures ANOVA)
- This course will focus on regression

Recap: multiple regression

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

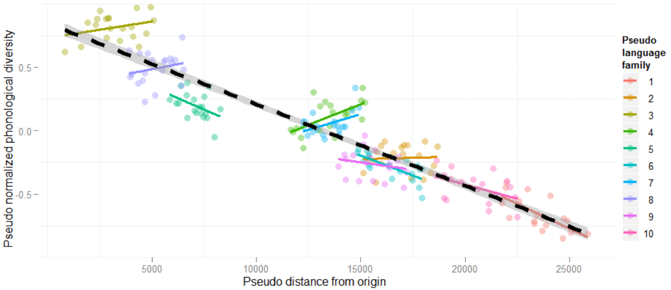
Mixed-effects regression modeling: introduction

- Mixed-effects regression modeling distinguishes fixed-effect and random-effect factors
- Fixed-effect factors:
  - Repeatable levels
  - Small number of levels (e.g., Gender, Word Category)
  - Same treatment as in multiple regression (treatment coding)
- Random-effect 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: e.g., non-native participants might benefit more from frequent words than native participants
    = Random Slope for Item Frequency per Subject
  - The effect of subject age on RT might be different for one word than another: e.g., modern words might be recognized faster by younger participants
    = Random Slope for Subject Age per Item
- Note that it is essential to test for random slopes!

Random slopes are necessary!



		Estimate	Std. Error	t value	Pr(>|t|)
Linear regression	DistOrigin	-6.418e-05	1.808e-06	-35.49	<2e-16
+ Random intercepts	DistOrigin	-2.224e-05	6.863e-06	-3.240	<0.001
+ Random slopes	DistOrigin	-1.478e-05	1.519e-05	-0.973	n.s.

This example is explained at http://hlplab.wordpress.com

Specific models for every observation

- Mixed-effects regression analysis 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
  - Parsimony: a single parameter (standard deviation) models this variation for every random slope or intercept (a normal distribution with mean 0 is assumed)
  - The adjustments to population slopes and intercepts are Best Linear Unbiased Predictors (BLUPs)
  - AIC comparisons 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)

Specific models for every observation

- $RT = 200 - 5WF + 3WL + 10SA$ (general model)
  - The intercepts and slopes may vary (according to the estimated standard variation for each parameter) and this influences the word- and subject-specific values

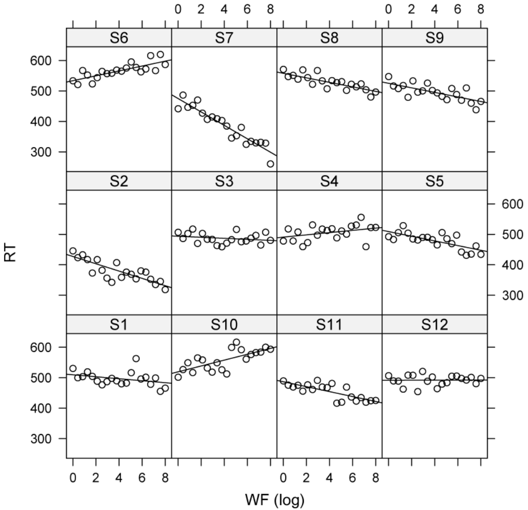
- $RT = 400 - 5WF + 3WL - 2SA$ (word: scythe)
- $RT = 300 - 5WF + 3WL + 15SA$ (word: twitter)
- $RT = 300 - 7WF + 3WL + 10SA$ (subject: non-native)
- $RT = 150 - 5WF + 3WL + 10SA$ (subject: fast)

- And it is easy to use!

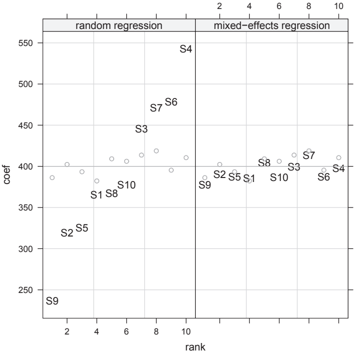
```
> lmer( RT ~ WF + WL + SA + (1+SA|Wrd) + (1+WF|Subj) )
```

- `lmer` figures out by itself if the random-effects are nested (schools-pupils), or crossed (participants-items)

Specific models for every subject



BLUPs of `lmer` do not suffer from shrinkage



- The BLUPS (i.e. adjustment to the model estimates per item/participant) are close to the real adjustments, as `lmer` takes into account regression towards the mean (fast subjects will be slower next time, and slow subjects will be faster) thereby avoiding overfitting and improving prediction – see Efron & Morris (1977)

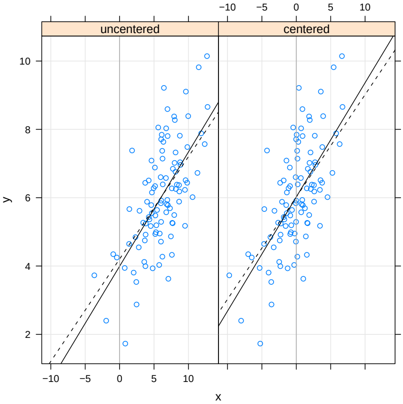
Methodological issues

- Parsimony
- Centering
- Assumptions about the residuals
  - Normally distributed and homoskedastic
  - No trial-by-trial dependencies
- Assumptions about the predictors
  - We assume linearity, if this is not suitable you can use *Generalized additive mixed-effects regression modeling* (Wood, 2006)
- Model criticism
- How to select the “best” model?

Parsimony

- 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

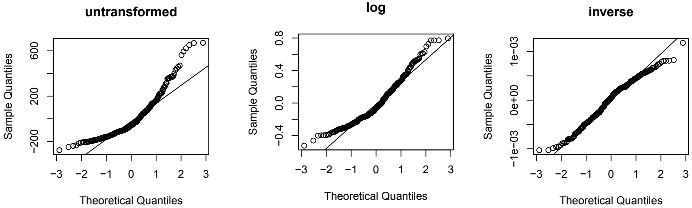
Center your variables (i.e. subtract the mean)

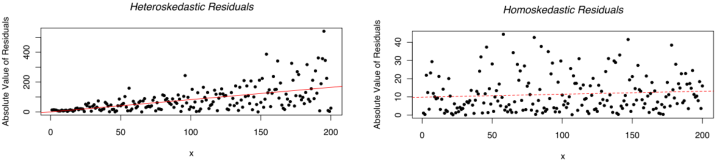


- Otherwise random slopes and intercepts may show a spurious correlation
- Also helps the interpretation of factorial predictors in model (marking differences at means of other variables, rather than at values equal to 0)

Residuals: normally distributed and homoskedastic

- The errors should follow a normal distribution with mean zero and the same standard deviation for any cell in your design, and for any covariate
  - If not then transform the dependent variable: log(Y), or -1000/Y
  - And use mixed-effects regression
  - Note: not required for *logistic* regression





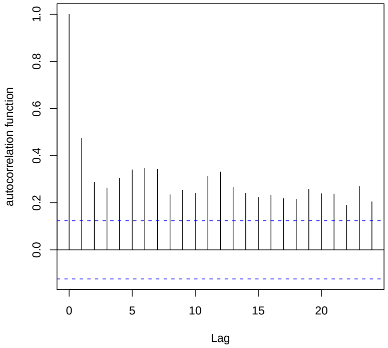
Residuals: no trial-by-trial dependencies

- Residuals should be independent
  - With trial-by-trial dependencies, this assumption is violated, which may result in models that underperform
- Possible remedies:
  - Include trial as a predictor in your model
  - Include the value of the dependent variable at the previous trial as a predictor in your model

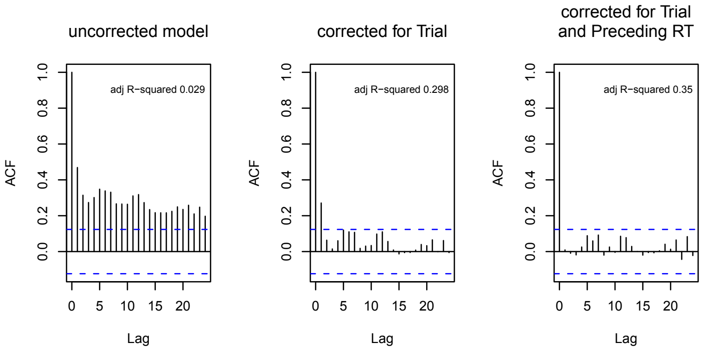
Trial-by-trial dependencies in a word naming task

- Word naming (reading aloud) of Dutch verbs
- Trial-by-trial dependencies in the residuals of the model

```
> acf(resid(model), main=" ", ylab="autocorrelation function")
```



Taking into account trial-by-trial dependencies



Model criticism

- Check the distribution of residuals: if not normally distributed then transform dependent variable (as illustrated before)
- Check outlier characteristics and refit the model when large outliers are excluded to verify that your effects are not ‘carried’ by these outliers
- Important: no *a priori* exclusion of outliers without a clear reason
  - A good reason is not that the value is over 2.5 SD above the mean
  - A good reason (e.g.,) is that the response is faster than possible

Model selection I

- "The data analyst knows more than the computer." (Henderson & Velleman, 1981) - Get to know your data!
- There is no adequate *automatic* procedure to find the best model
- My stepwise variable-selection procedure (for exploratory analysis):
  - Include random intercepts
  - Add other potential explanatory variables one-by-one
  - Insignificant predictors are dropped
  - Test predictors for inclusion which were excluded at an earlier stage
  - Test possible interactions (don't make it too complex)
  - Try to break the model by including significant predictors as random slopes
  - Only choose a more complex model if it decreases AIC by at least 2
    - Then the model with the lowest AIC is at least 2.7 ($e^{\frac{\Delta \text{AIC}}{2}}$) times more likely
- Starting from the most complex model in the context of a mixed-effects regression model is frequently impossible (the model with the full random-effects structure may not converge)

Model selection II

- For a hypothesis-driven analysis, stepwise selection is problematic
  - Confidence intervals too narrow: *p*-values too low (multiple comparisons)
  - See, e.g., Burnham & Anderson (2002)
- Solutions:
  - Careful specification of potential *a priori* models lining up with the hypotheses (including random intercepts and slopes) and evaluating only these models (e.g., via AIC)
  - Validating a stepwise procedure via cross validation (e.g., bootstrap analysis)

Case study: long-distance priming

- De Vaan, Schreuder & Baayen (The Mental Lexicon, 2007)
- Design
  - long-distance priming (39 intervening items)
  - base condition (`baseheid`): base preceded neologism (fluffy - fluffiness)
  - derived condition (`heid`): identity priming (fluffiness - fluffiness)
  - only items judged to be real words are included
- Prediction
  - Subjects in the derived condition (`heid`) would be faster than those in the base condition (`baseheid`)

A first model: counterintuitive results!

(note: $t > 2 \Rightarrow p < 0.05$, for $N \gg 100$)

```
> library(lme4) # version 1.1.7 (NOT compatible with version 0.9...)
> dat = read.table('datprevrt.txt', header=T) # adapted primingHeid data set

> dat.lmer1 = lmer(RT ~ Condition + (1|Word) + (1|Subject), data=dat)
> summary(dat.lmer1)

...

Random effects:
 Groups   Name        Variance Std.Dev.
 Word     (Intercept) 0.003411 0.05841
 Subject  (Intercept) 0.040843 0.20210
 Residual             0.044084 0.20996
Number of obs: 832, groups:  Word, 40; Subject, 26

Fixed effects:
              Estimate Std. Error t value
(Intercept)    6.60296    0.04215  156.66
Conditionheid  0.03127    0.01467    2.13  # slower...

...
```

Evaluation

- Counterintuitive inhibition
- But various potential factors are not accounted for in the model
  - Longitudinal effects: trial rank, RT to preceding trial
  - RT to prime as predictor
  - Response to the prime (correct/incorrect): a yes response to a target associated with a previously rejected prime may take longer
  - The presence of atypical outliers

An effect of trial?

```
> dat.lmer2 = lmer(RT ~ Trial + Condition + (1|Subject) + (1|Word),
                   data=dat)
> round( summary(dat.lmer2)$coef, 3 )

                Estimate Std. Error t value
(Intercept)        6.633      0.047 142.153
Trial              0.001      0.001  -1.519
Conditionheid      0.031      0.015   2.114
```

An effect of previous trial RT?

```
> dat$PrevRT = log(dat$PrevRT) # RT is already log-transformed
> dat.lmer3 = lmer(RT ~ PrevRT + Condition + (1|Subject) + (1|Word),
                   data=dat)
> round( summary(dat.lmer3)$coef, 3 )

              Estimate Std. Error t value
(Intercept)      5.805      0.223  26.032
PrevRT           0.121      0.033   3.633
Conditionheid    0.028      0.015   1.903
```

An effect of RT to prime?

```
> dat.lmer4 = lmer(RT ~ RTtoPrime + PrevRT + Condition + (1|Subject)
                   + (1|Word), data=dat)
> round( summary(dat.lmer4)$coef, 3 )

               Estimate Std. Error t value
(Intercept)       4.749      0.295  16.080
RTtoPrime         0.164      0.032   5.141
PrevRT            0.119      0.033   3.605
Conditionheid    -0.006      0.016  -0.383
```

An effect of the decision for the prime?

```
> dat.lmer5 = lmer(RT ~ RTtoPrime + ResponseToPrime + PrevRT + Condition
                   + (1|Subject) + (1|Word), data=dat)
> round( summary(dat.lmer5)$coef, 3 )

                          Estimate Std. Error t value
(Intercept)                  4.763      0.292  16.299
RTtoPrime                    0.165      0.031   5.242
ResponseToPrimeincorrect     0.100      0.023   4.445
PrevRT                       0.114      0.033   3.495
Conditionheid               -0.018      0.016  -1.107
```

Interaction for prime-related predictors?

```
> dat.lmer6 = lmer(RT ~ RTtoPrime * ResponseToPrime + PrevRT + Condition
                   + (1|Subject) + (1|Word), data=dat)
> round( summary(dat.lmer6)$coef, 3 )

                                    Estimate Std. Error t value
(Intercept)                            4.324      0.315  13.720
RTtoPrime                              0.228      0.036   6.334
ResponseToPrimeincorrect               1.455      0.405   3.590
PrevRT                                 0.118      0.033   3.640
Conditionheid                         -0.027      0.016  -1.642
RTtoPrime:ResponseToPrimeincorrect    -0.202      0.061  -3.344
```

- Interpretation: the RT to the prime is only predictive for the RT of the target word when the prime was judged to be a correct word

An effect of base frequency?

(Note the lower variance of the random intercept for word: previous value was 0.0034)

```
> dat.lmer7 = lmer(RT ~ RTtoPrime * ResponseToPrime + PrevRT + BaseFrequency
                 + Condition + (1|Subject) + (1|Word), data=dat)
> summary(dat.lmer7)

Random effects:
 Groups   Name        Variance Std.Dev.
 Word     (Intercept) 0.001151 0.03393
 Subject  (Intercept) 0.023991 0.15489
 Residual             0.042240 0.20552
Number of obs: 832, groups: Word, 40; Subject, 26

Fixed effects:
                                   Estimate Std. Error t value
(Intercept)                        4.440979   0.319605  13.895
RTtoPrime                          0.218242   0.036153   6.037
ResponseToPrimeincorrect           1.397052   0.405164   3.448
PrevRT                             0.115424   0.032456   3.556
BaseFrequency                     -0.009243   0.004371  -2.115
Conditionheid                     -0.024656   0.016179  -1.524
RTtoPrime:ResponseToPrimeincorrect -0.193987   0.060550  -3.204
```

Testing random slopes: no main frequency effect!

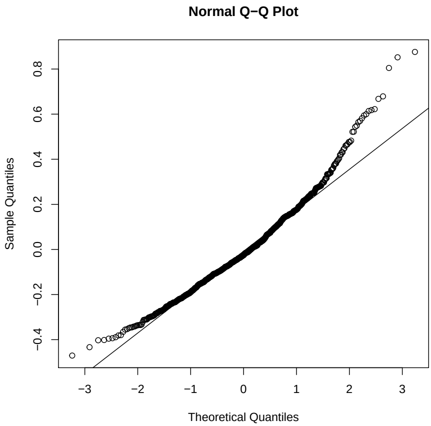
```
> dat.lmer7a = lmer(RT ~ RTtoPrime * ResponseToPrime + PrevRT
                     + BaseFrequency + Condition + (1|Subject)
                     + (0+BaseFrequency|Subject) + (1|Word), data=dat)
> AIC(dat.lmer7) - AIC(dat.lmer7a) # compare AIC
[1] 3.647772 # at least 2 lower: dat.lmer7a is better

# Alternative to AIC comparison: Likelihood Ratio Test
> anova(dat.lmer7,dat.lmer7a,refit=F) # compares models

           Df     AIC     BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
dat.lmer7  10 -125.55 -78.315 72.777  -145.55
dat.lmer7a 11 -129.20 -77.239 75.601  -151.20 5.6478      1    0.01748 *

> round( summary(dat.lmer7a)$coef, 3 )

                                  Estimate Std. Error t value
(Intercept)                          4.482      0.317  14.124
RTtoPrime                            0.218      0.036   6.067
ResponseToPrimeincorrect             1.417      0.402   3.524
PrevRT                               0.108      0.032   3.354
BaseFrequency                       -0.008      0.005  -1.485
Conditionheid                       -0.025      0.016  -1.530
RTtoPrime:ResponseToPrimeincorrect  -0.197      0.060  -3.274
```

Testing for correlation parameters in random effects

```
> dat.lmer7b = lmer(RT ~ RTtoPrime * ResponseToPrime + PrevRT
                         + BaseFrequency + Condition
                         + (1+BaseFrequency|Subject) + (1|Word), data=dat)
> summary(dat.lmer7b)

...
Random effects:
 Groups   Name          Variance  Std.Dev. Corr
 Word     (Intercept)   0.0011857 0.03443
 Subject  (Intercept)   0.0166775 0.12914
          BaseFrequency 0.0001861 0.01364  0.41
 Residual               0.0414192 0.20352
Number of obs: 832, groups: Word, 40; Subject, 26
...

> AIC(dat.lmer7a) - AIC(dat.lmer7b)
[1] -1.138734

> anova(dat.lmer7a,dat.lmer7b,refit=F)

           Df     AIC     BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
dat.lmer7a 11 -129.20 -77.239 75.601  -151.20
dat.lmer7b 12 -128.06 -71.377 76.031  -152.06 0.8613      1     0.3534
```

Model criticism

```
> qqnorm(resid(dat.lmer7a))
> qqline(resid(dat.lmer7a))
```



The trimmed model

```
> dat2 = dat[ abs(scale(resid(dat.lmer7a))) < 2.5 , ]
> dat2.lmer7a = lmer(RT ~ RTtoPrime * ResponseToPrime + PrevRT
                         + BaseFrequency + Condition + (1|Subject)
                         + (0+BaseFrequency|Subject) + (1|Word), data=dat2)
> round( summary(dat2.lmer7a)$coef, 3 )

                                  Estimate Std. Error t value
(Intercept)                          4.447      0.286  15.551
RTtoPrime                            0.235      0.032   7.347
ResponseToPrimeincorrect             1.560      0.356   4.388
PrevRT                               0.096      0.029   3.266
BaseFrequency                       -0.008      0.005  -1.775
Conditionheid                       -0.038      0.014  -2.657
RTtoPrime:ResponseToPrimeincorrect  -0.216      0.053  -4.066
```

The trimmed model

- Just 2% of the data removed

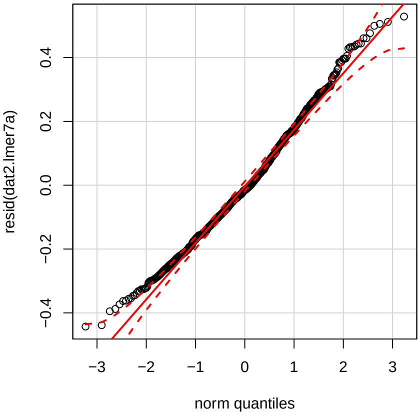
```
> noutliers = sum(abs(scale(resid(dat.lmer7a)))>=2.5)
> noutliers
[1] 17

> noutliers/nrow(dat)
[1] 0.02043269
```

- Improved fit (explained variance):

```
> cor(dat$RT,fitted(dat.lmer7a))^2
[1] 0.5210606

> cor(dat2$RT,fitted(dat2.lmer7a))^2
[1] 0.5717716
```

Checking the residuals of trimmed model

```
> library(car)
> qqp(resid(dat2.lmer7a))
```



Bootstrap sampling to validate results

```
> library(boot)
> bs.lmer7a = confint(dat2.lmer7a, method="boot", nsim = 1000, level = 0.95)
> bs.lmer7a

                                           2.5 %        97.5 %
sd_(Intercept)|Word                  0.000000000  0.0388402702
sd_BaseFrequency|Subject             0.003271268  0.0218704020
sd_(Intercept)|Subject               0.096063055  0.1898946632
sigma                                0.170499205  0.1900119424
(Intercept)                          3.877547385  5.0447555391
RTtoPrime                            0.172898282  0.2990675635
ResponseToPrimeincorrect             0.862874611  2.3142086875
PrevRT                               0.033960630  0.1591520465
BaseFrequency                       -0.017415646  0.0007102762 # n.s.
Conditionheid                       -0.064896445 -0.0084641311
RTtoPrime:ResponseToPrimeincorrect -0.329769967 -0.1099138944
```

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 easy with `lmer` in `R`
- Don't forget model criticism!

- Lab session to illustrate the commands used here
  `http://www.let.rug.nl/wieling/statscourse/lecture1`
  - Lab session contains additional information: how to do multiple comparisons, using other optimizers, and conducting logistic regression

Thank you for your attention!