An Introduction to Linear Mixed Effects Models for Psychological Science

Dr. Lotte Meteyard (Dr Rob Davies)

Meteyard & Davies (in prep) Best practice for linear mixed effect models in psychological science

Mixed models

- Analysis technique that includes both fixed and random effects.
- *NB: they are not new*

- Related terms..

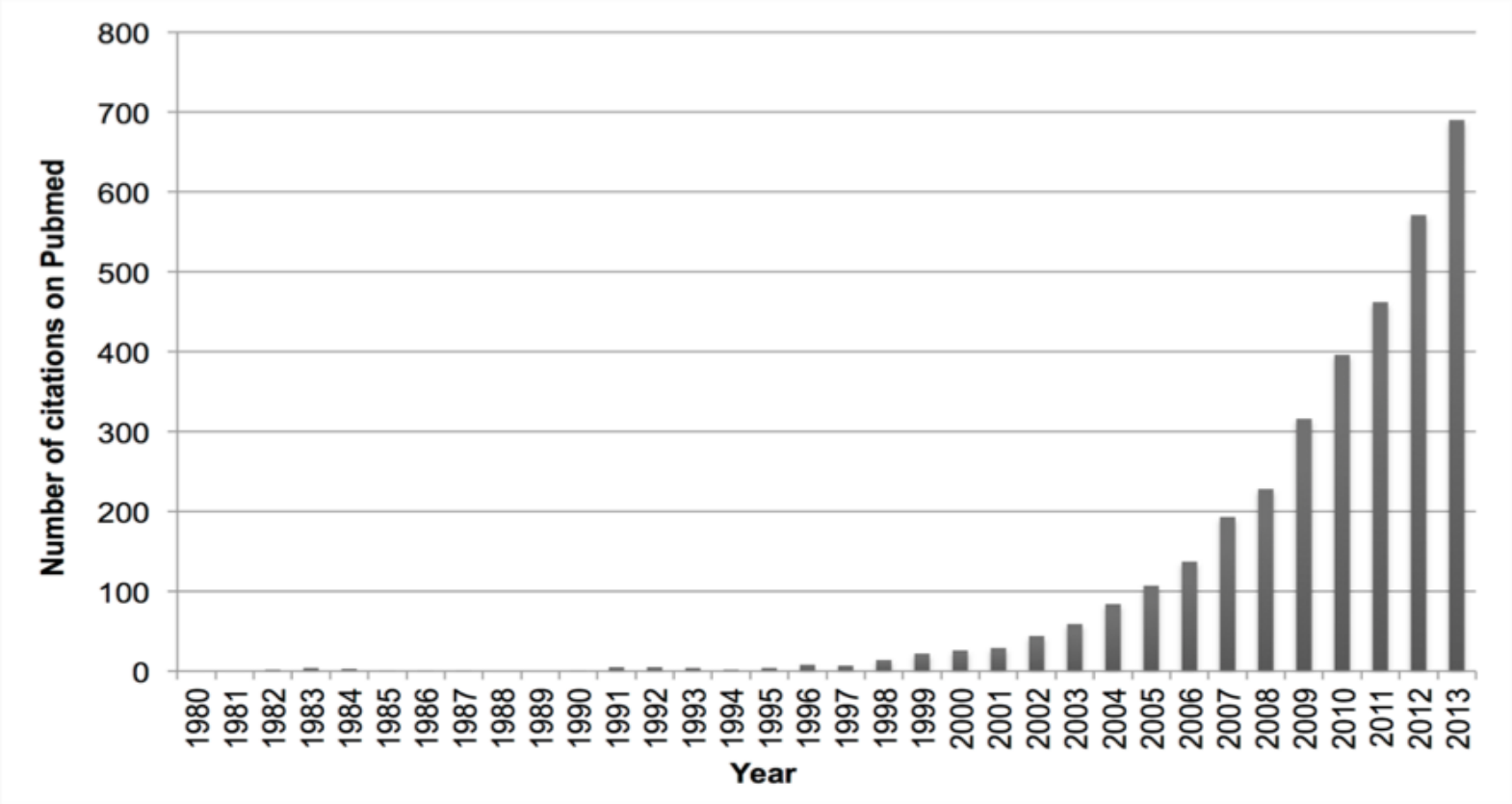
  Linear/generalised mixed models

  Multilevel models/analysis

  Heirarchical models

Mixed models

Number of Pubmed citations for ‘Linear Mixed Models’ by year



PCLS November 2018; Meteyard & Davies (in prep)

LMM Angst

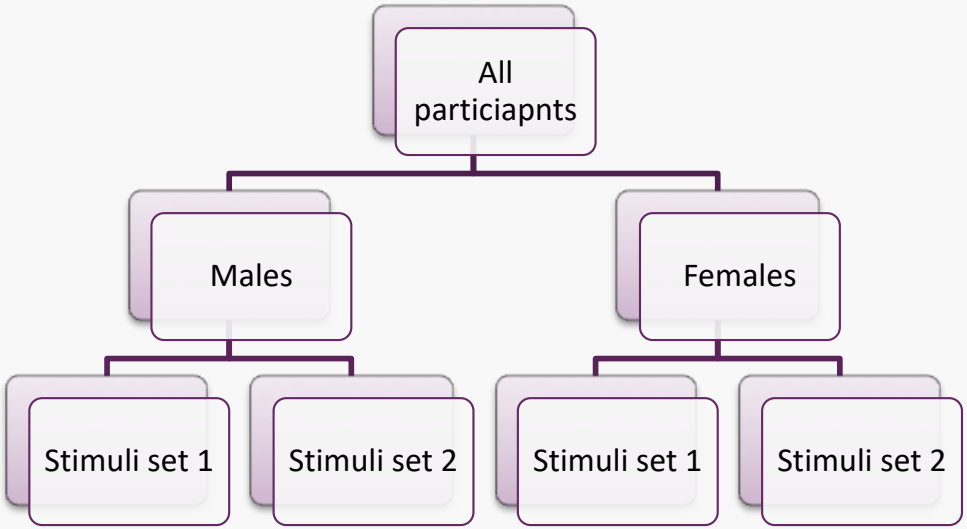
Reported challenge	%
Lack of standardized procedures	26
Selecting and specifying models	25
Researcher reports lack of knowledge	18
Understanding and interpreting random effects	14
Lack of training/guidelines for analysis, interpretation and reporting	13
Use of new and unfamiliar software	12

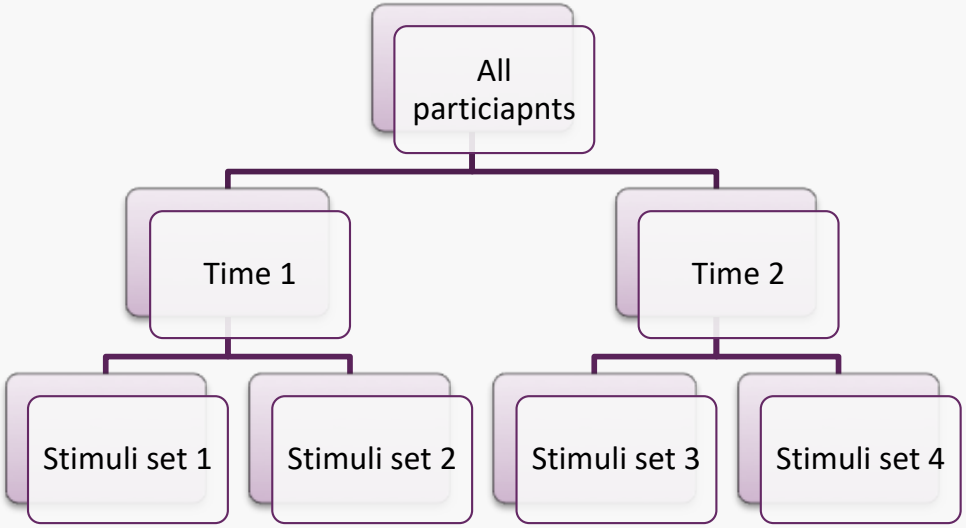
Reported challenge	%
General concern over use of LMM for own analysis	75
Reporting results	15
Model selection	14
Learning and understanding analysis	14
Lack of established standards	11
General concern over use of LMMs for discipline	73
Lack of standards	23
LMMs used when not fully understood	23
Misuse of models	17
Reporting is inconsistent and lacks detail	17
Peer review of LMMs is not robust	10

PCLS November 2018; Meteyard & Davies (in prep)

What is a LMM?

- Used for any data where there is a hierarchical or multi-level structure. That is, groups & sub-groups.





PCLS November 2018; Meteyard & Davies (in prep)

Multilevel/Hierarchical structures

- Pre and post tests
- Multiple stimuli, multiple grouped samples
- Repeated testing
- For example

  - data collected from the same people at different times *(correlated)*

  - data collected from the same place/location (correlated)

  - data collected with different groups of items/stimuli *(confound of different items)*

Multilevel/Hierarchical structures

- Particularly relevant for experiments / studies in which items are confounded with conditions

  e.g. faces vs houses, nouns vs verbs…

- Issue in these studies of the 'Items as a fixed effect fallacy" (Clark, 1973).

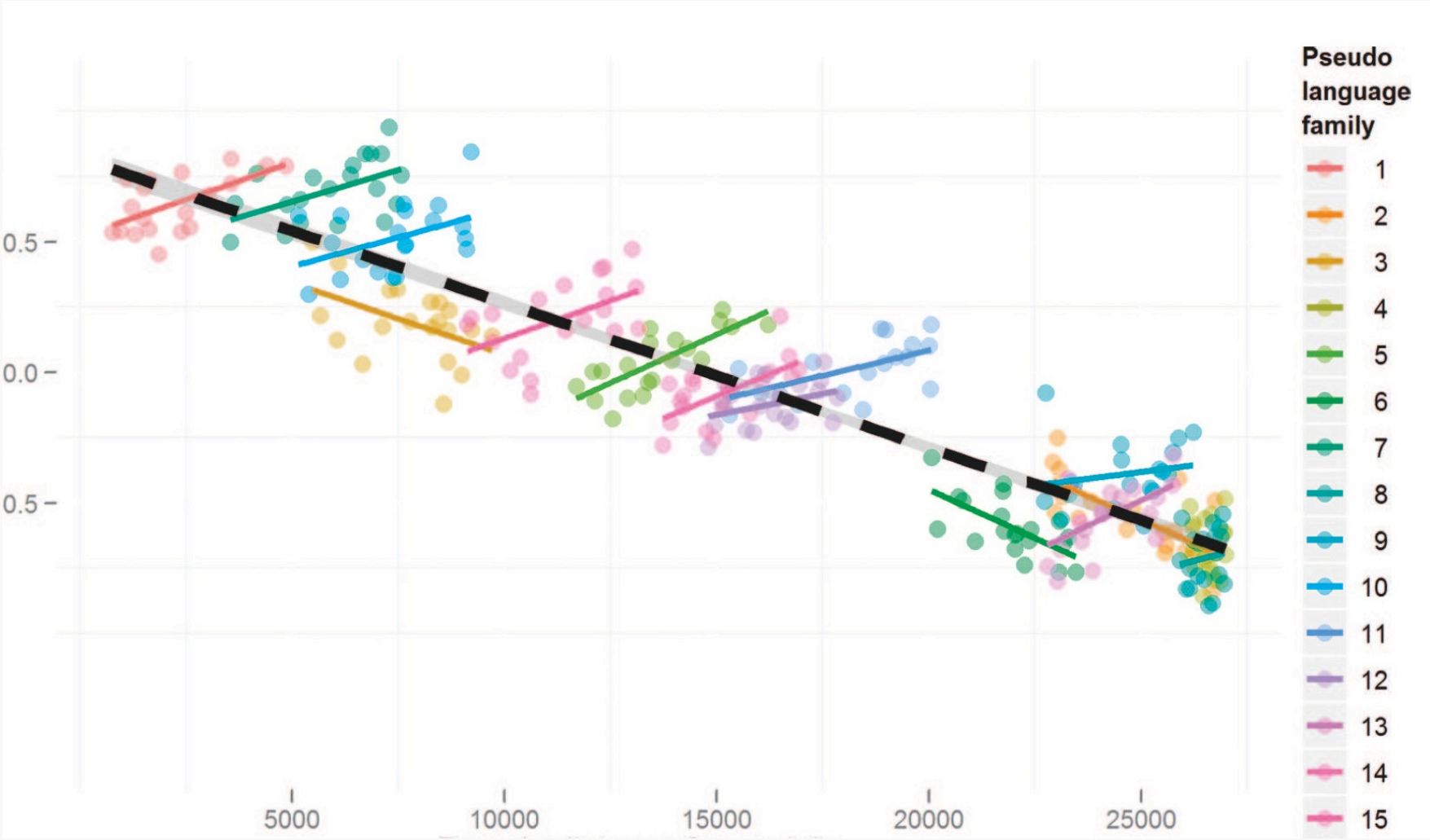
Multilevel/Hierarchical structures

- Hierarchical structures are a problem because

  (1) Most analyses assume data is independent – *standard errors smaller if data assumed to be independent when it isn't*

  (2) Averaging or aggregating across groups/sub-groups can 'hide' the true pattern in the data -
  we assume taking the average is OK, but this is usually a convenient oversimplification

Jaeger et al (2011)



Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. Linguistic Typology, 15(2), 281-320.

When to use LMMs

- Whenever you want to manage these problems…

  (1) Model subject & item variance (e.g. F1 vs F2) - may replace your usual ANOVA analysis

  (2) Model groupings in your data
  (e.g. over time, hierarchical designs)

  (3) Individual differences and effects within specific groups

What should my data look like?

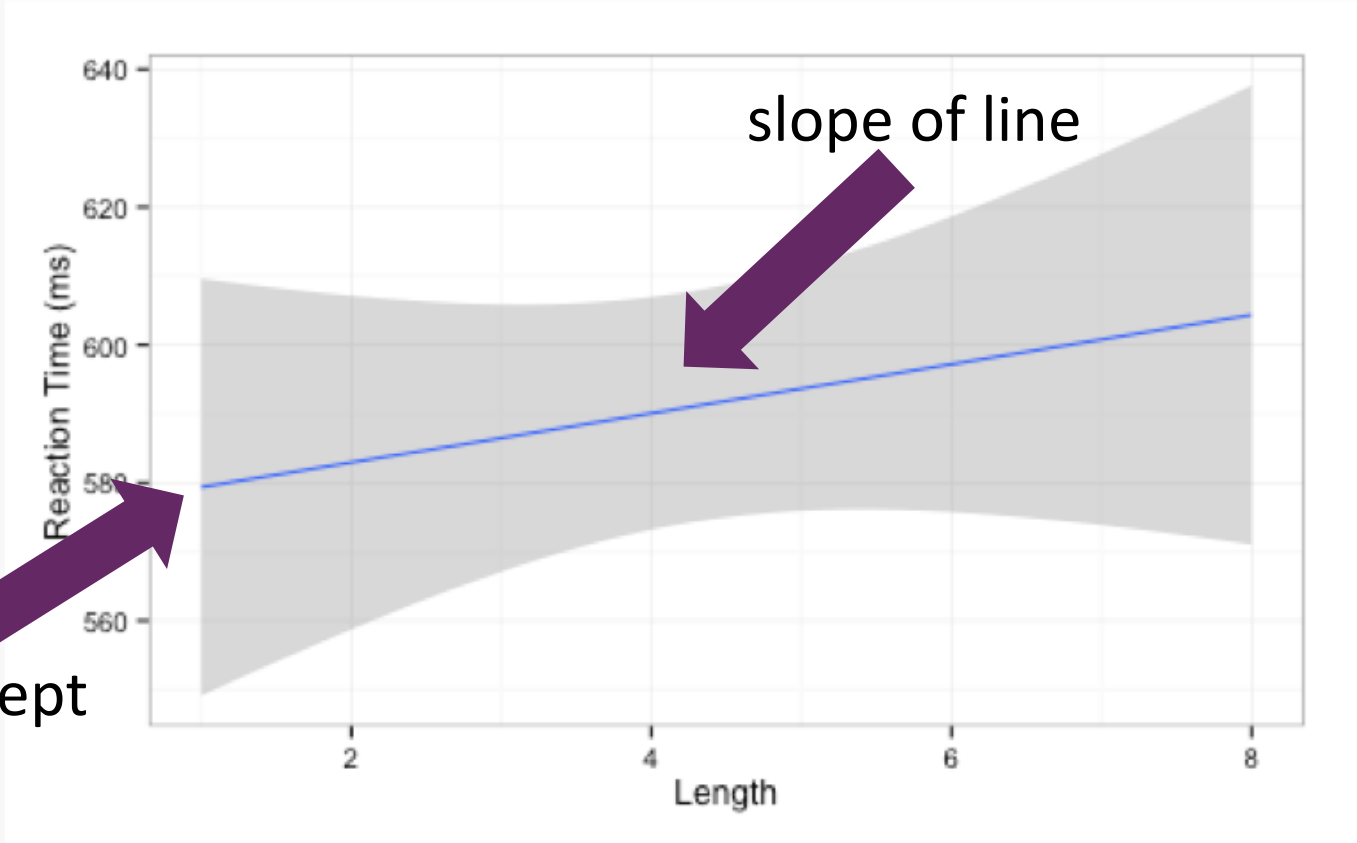
- LMMs should be seen as a form of REGRESSION

  So..
  (1) Lots of data points to power the analysis
  (2) Good/sufficient variance in your data
  (2) Data meets assumptions of regression
  NB: *linear* mixed effects models

- If you are planning on using LMMs,
  then *design your experiment to fit LMM analysis*.

What do LMMs do?

- Standard regression



RT ~ intercept + slope(Length)

Regression

Experiment with RT by Length of Word (IV)

RT ~ intercept + slope(Length)

y ~ a + b(x)

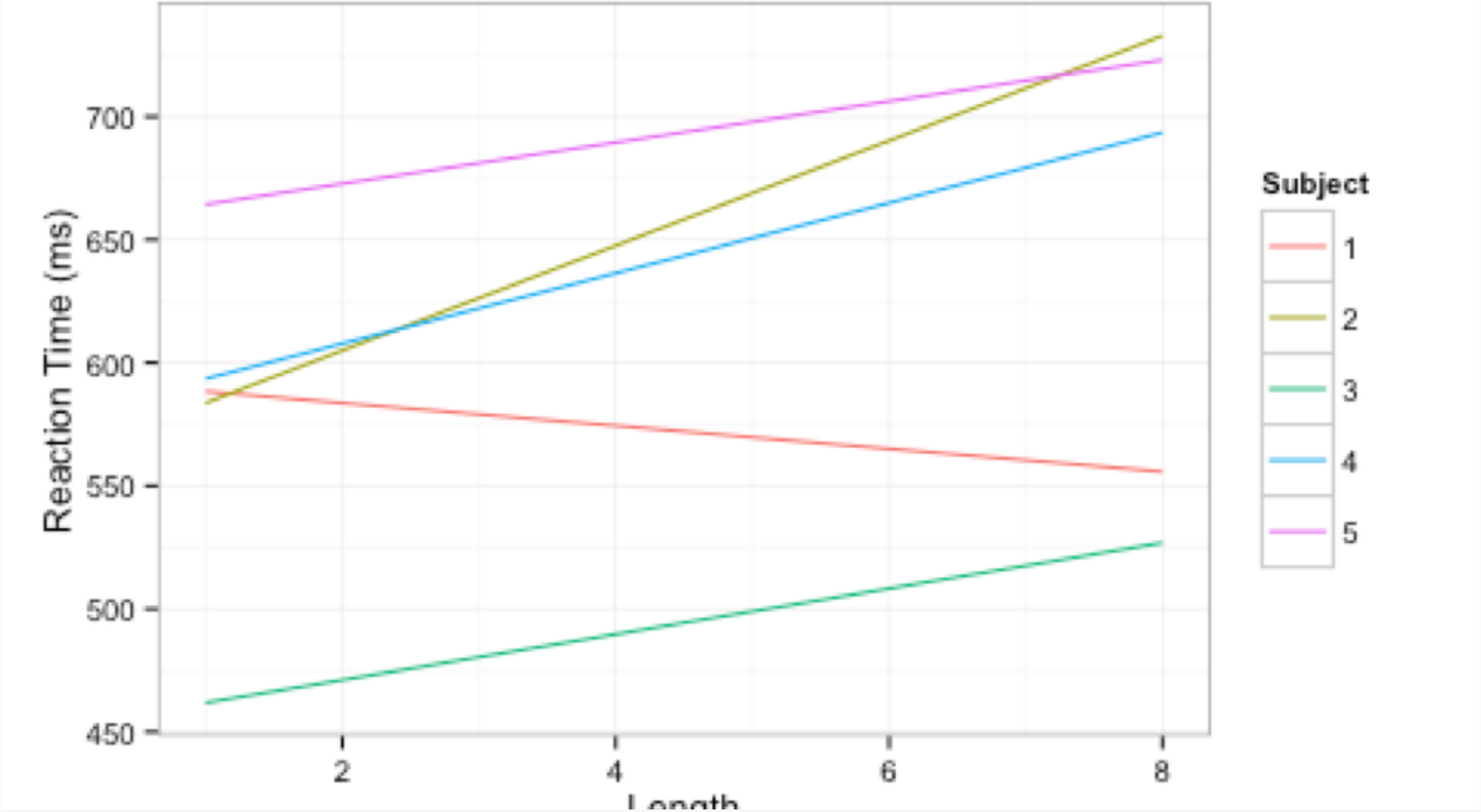
LMM & regression

For LMMs we can model *how the intercept and slope will vary across different groups*

Because we are modeling *variation* these are called *random effects* (i.e. the group has some random variation associated with it). **Measured in standard deviations**

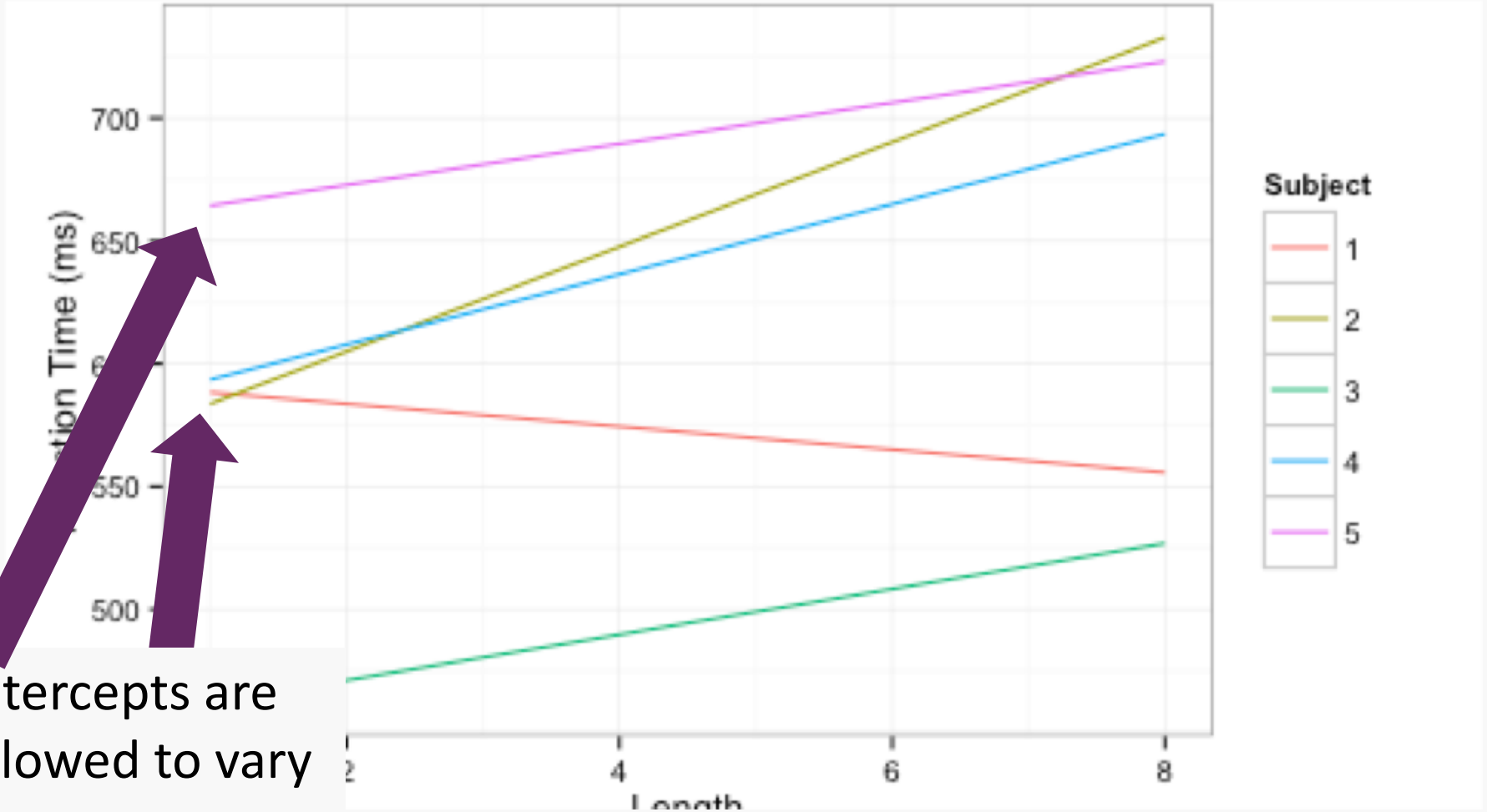
Fixed effects are your normal predictors/IVs.
Fixed because we want to know *on average* what they do. **Measured in means**.

What do LMMs do?



PCLS November 2018; Meteyard & Davies (in prep)

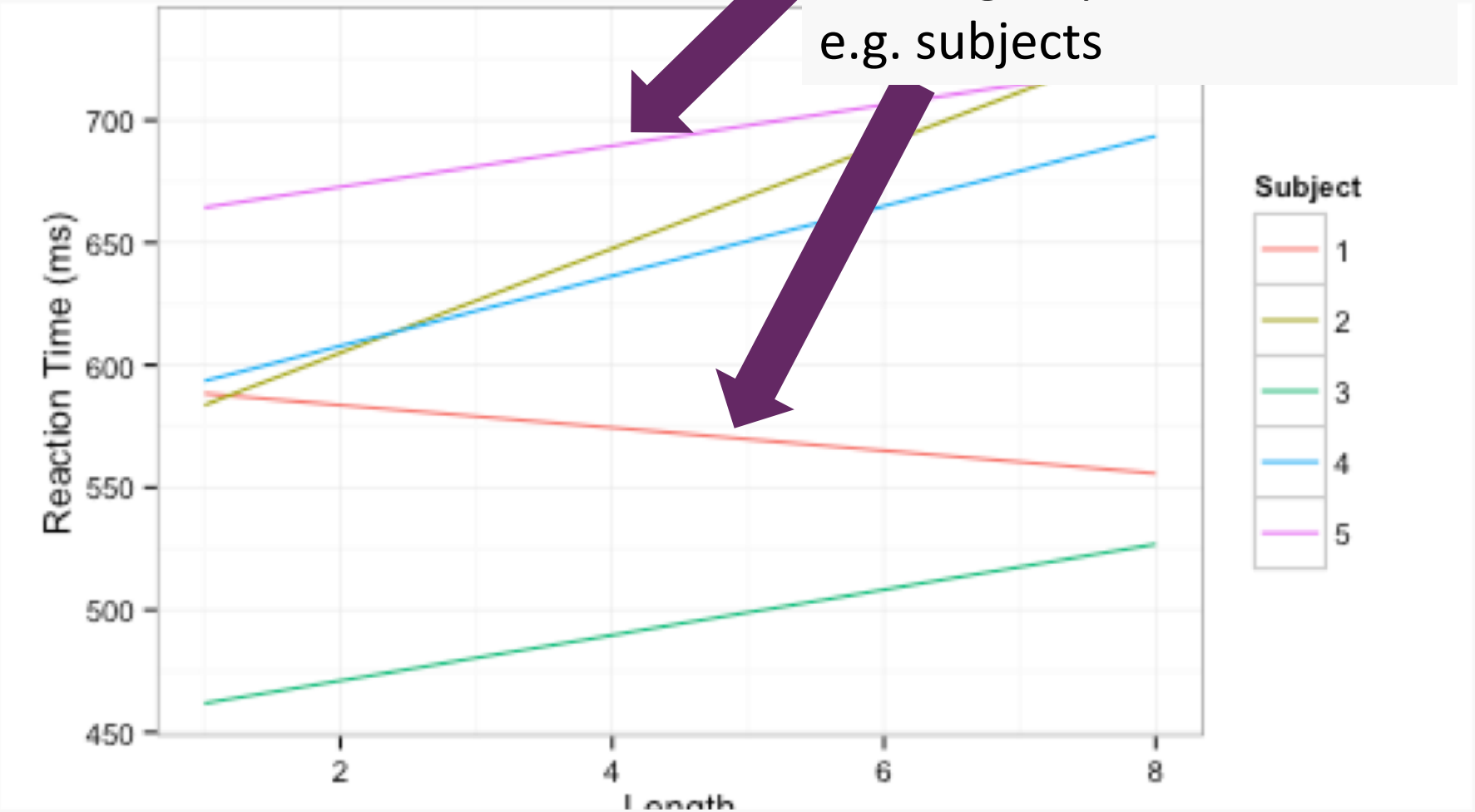
What do LMMs do?



intercepts are allowed to vary

What do LMMs do?

slope is allowed to vary across groups
e.g. subjects



What do LMMs do

For LMMs we can model *how the intercept and slope will vary across different groups*

For example, model the difference (i.e. the variance/deviation) for each subject from the average slope and the average intercept for all subjects.

Similarly, model the difference for each item from the average slope and the average intercept for all items.

What do LMMs do?

- That is basically it!
- Gets complicated because you can end up with very large random effect structures

  e.g. intercepts +/- slopes for subjects, items, and more

  slopes varying for different predictors
  for different groupings

  slopes of interactions between fixed effects across different groups

Intercepts and slopes

Experiment with Subjects, Items and Length of Word (IV)

- Random intercepts only

RT ~ Length + (1|Subjects) + (1|Items)

- Random intercepts and slopes (nb: 1 = correlated)

RT ~ Length + (1 + Length|Subjects)
+ (1 + Length|Items)

Intercepts and slopes

Experiment with Subjects, Ite /ord (IV)

Intercept varies by Subjects

- Random intercepts or

RT ~ Length + (1|Subjects) + (1|Items)

Intercept varies by Items

- Random intercepts and slopes (nb: 1 = correlated)

RT ~ Length + (1 + Length|Subjects)
+ (1 + Length|Items)

Intercepts and slopes

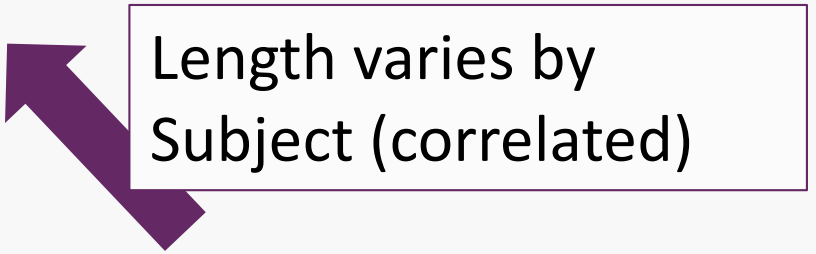
Experiment with Subjects, Items and Length of Word (IV)

- Random intercepts only

RT ~ Length + (1|Subjects) + (1|Items)

- Random intercepts and slopes (nb: 1 = correlated)

RT ~ Length + (1 + Length|Subjects)



```
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to degrees of freedom [lmerM
Formula: logrt ~ zAge + zTOWRE_wordacc + zTOWRE_nonwordacc + item_type +
    zLength + zOrtho_N + (1 | subjectID) + (1 | item_name)
   Data: ML.all.correct

     AIC      BIC   logLik deviance df.resid
-18319.5 -18247.1   9169.7 -18339.5    10244

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5521 -0.6479 -0.1575  0.4713  5.2553

Random effects:
 Groups    Name        Variance  Std.Dev.
 item_name (Intercept) 0.0006949 0.02636
 subjectID (Intercept) 0.0024557 0.04955
 Residual              0.0092889 0.09638
Number of obs: 10254, groups:  item_name, 320; subjectID, 34

Fixed effects:
                   Estimate Std. Error         df t value Pr(>|t|)
(Intercept)       2.866e+00  8.859e-03  3.907e+01 323.552  < 2e-16 ***
zAge              1.318e-02  8.871e-03  3.397e+01   1.486 0.146521
zTOWRE_wordacc    1.078e-04  1.146e-02  3.398e+01   0.009 0.992550
zTOWRE_nonwordacc -1.747e-02  1.138e-02  3.398e+01  -1.534 0.134173
item_typeword     -8.655e-02  3.517e-03  3.035e+02 -24.610  < 2e-16 ***
zLength            8.645e-03  2.198e-03  3.044e+02   3.934 0.000104 ***
zOrtho_N           2.254e-03  2.195e-03  3.046e+02   1.027 0.305203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to degrees of freedom [lmerM
Formula: logrt ~ zAge + zTOWRE_wordacc + zTOWRE_nonwordacc + item_type +
    zLength + zOrtho_N + (1 | subjectID) + (1 | item_name)
   Data: ML.all.correct

     AIC      BIC   logLik deviance df.resid
-18319.5 -18247.1   9169.7 -18339.5    10244

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5521 -0.6479 -0.1575  0.4713  5.2553

Random effects:
 Groups    Name        Variance  Std.Dev.
 item_name (Intercept) 0.0006949 0.02636
 subjectID (Intercept) 0.0024557 0.04955
 Residual              0.0092889 0.09638
Number of obs: 10254, groups:  item_name, 320; subjectID, 34

Fixed effects:
                   Estimate Std. Error         df t value Pr(>|t|)
(Intercept)       2.866e+00  8.859e-03  3.907e+01 323.552  < 2e-16 ***
zAge              1.318e-02  8.871e-03  3.397e+01   1.486 0.146521
zTOWRE_wordacc    1.078e-04  1.146e-02  3.398e+01   0.009 0.992550
zTOWRE_nonwordacc -1.747e-02  1.138e-02  3.398e+01  -1.534 0.134173
item_typeword     -8.655e-02  3.517e-03  3.035e+02 -24.610  < 2e-16 ***
zLength            8.645e-03  2.198e-03  3.044e+02   3.934 0.000104 ***
zOrtho_N           2.254e-03  2.195e-03  3.046e+02   1.027 0.305203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Random effects
- varaince/SD associated with subject/item etc.
(e.g. difference between each subject and mean)

```
Linear mixed model fit by maximum likelihood t-tests use Satterthwaite approximations to degrees of freedom [lmerM
Formula: logrt ~ zAge + zTOWRE_wordacc + zTOWRE_nonwordacc + item_type +
    zLength + zOrtho_N + (1 | subjectID) + (1 | item_name)
   Data: ML.all.correct

     AIC      BIC   logLik deviance df.resid
-18319.5 -18247.1   9169.7 -18339.5    10244

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5521 -0.6479 -0.1575  0.4713  5.2553

Random effects:
 Groups    Name        Variance  Std.Dev.
 item_name (Intercept) 0.0006949 0.02636
 subjectID (Intercept) 0.0024557 0.04955
 Residual              0.0092889 0.09638
Number of obs: 10254, groups:  item_name, 320; subjectID, 34

Fixed effects:
                   Estimate Std. Error         df t value Pr(>|t|)
(Intercept)        2.866e+00  8.859e-03  3.907e+01 323.552  < 2e-16 ***
zAge               1.318e-02  8.871e-03  3.397e+01   1.486 0.146521
zTOWRE_wordacc     1.078e-04  1.146e-02  3.398e+01   0.009 0.992550
zTOWRE_nonwordacc -1.747e-02  1.138e-02  3.398e+01  -1.534 0.134173
item_typeword     -8.655e-02  3.517e-03  3.035e+02 -24.610  < 2e-16 ***
zLength            8.645e-03  2.198e-03  3.044e+02   3.934 0.000104 ***
zOrtho_N           2.254e-03  2.195e-03  3.046e+02   1.027 0.305203
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fixed effects
- Typically same coefficients estimates as OLS regression
- Standard errors will be different

LMMs in R

- Models estimate values for fixed and random effects
  NB: you are modeling your data
  i.e. fitted values estimated from the data

- Order of entry doesn't matter (simultaneous)

- Tries to find the solution that makes the observed data most likely (*maximum likelihood*)
  ML – assumes fixed effects are precise
  REML – averages fixed effects and then calculates random effects

- Sometimes this modeling fails (*convergence*)

Convergence

- This is not just a problem with the 'mathematical engine'

"determine whether the problem lies in a failure of the…optimization stage, as opposed to a case of model misspecification or unidentifiability or a problem with the underlying PLS algorithm. To date we have only observerd PLS failures.." Bates et al
(page 24; http://cran.r-project.org/web/packages/lme4/vignettes/lmer.pdf)

- Optimiziation
- Model misspecification
- Variance or covariance in random effects (0/1)
- Unidentifiability / PLS failures
- Cf: Bates et al (2015) – parsimonious models and Brauer & Curtin (2018)

Model building

There is no recipe book

- Each statistical analysis has "boundary conditions" (Baayen)
- Identify boundary conditions and *check* them
- Mixed models:

(1) Variance / number data points

(2) Normality / distribution of data (e.g. residuals, outliers)

(3) Rationale & selection of random & fixed effects

(4) 'Final' model – Does it look sensible?

Can you interpret it?

There is no recipe book

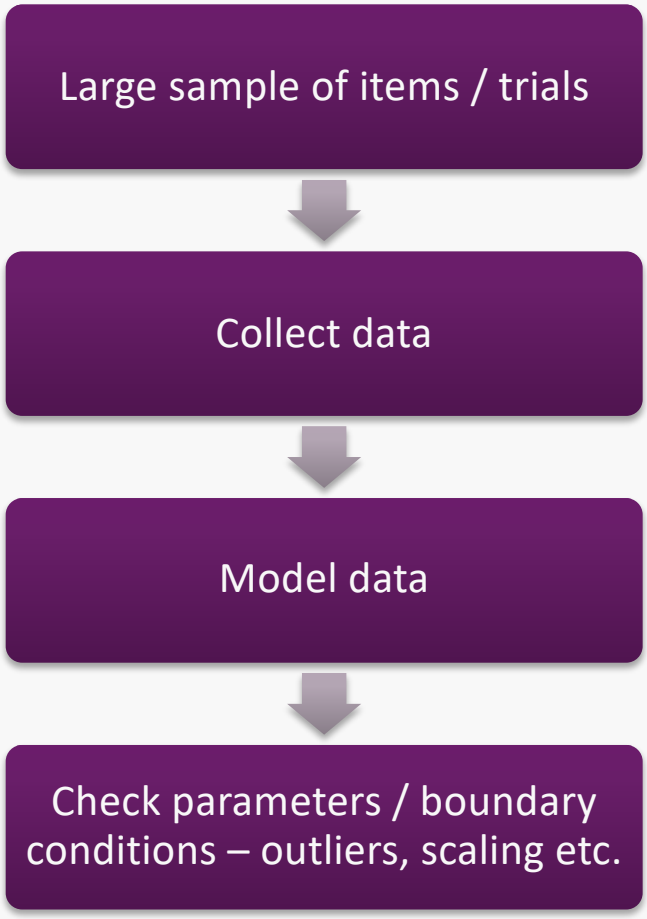
- Model checking and model selection

- Say what model checking was done
  (e.g. residuals, optimization, re)

- Say what approach and why
  (e.g. maximal to minimal; control/null model up;
  inclusion or not of higher order interactions)

- Report model selection / testing / comparison process

A mixed-model workflow



Maximise variance

Iterate

REPORT iterations

Most issues can be addressed by transparency..

PCLS November 2018; Meteyard & Davies (in prep)

Power for LMMs

- Non-trivial to compute
- E.g. random effect variances are essential for computing a-prior power (Westfall, Kenny & Judd, 2014)
- Just adding more items OR more participants doesn't work
- Need as many sampling units as possible, since this is the main limitation on power (Snijders, 2005).
- Some 'rules of thumb' appearing: 40 x 40 (Brysbaert & Stevens, 2018), or no less than 30-50 per sampling group (Bell et al, 2010; Maas & Hox, 2005; 2006)

Power for LMMs

- Remember it is regression with some bells and whistles..

Building models: many methods

- Simple model upwards:
  intercept only → + random effects → + fixed effects

  intercept = overall mean
  intercept assumed in lmer/R

- Models with control variables → + fixed effects
- Models with main effects → + interactions

- Simple regression/ordinary least squares models for sub-groups/per subject/per item, to inform LMM (Gelman & Hill, 2006)

Building models

- Add in fixed effect predictors:
  should see random effect variance *reduce* if the predictors are doing any work
- Compare successive models – does adding in a predictor make the model better?

- *Change either RE or FE for comparisons*
- Model comparison with **AIC** or **LRT**

Likelihood – *of data given the model*
Information criterion – *trade off between likelihood of data and number of parameters*

Building models

- Once have a reliable model, see if you can 'break' the effects you have

  e.g. change logRT to reciprocal (1/RT)?
  higher order interactions that modulate effects?
  other variables you expect to matter?

- Consider reporting multiple models

Random effects

- Sampling groups: subjects, items, school, etc.

- You expect or predict some variation
  e.g. interaction of subject ability with item

- LMM will tell you how much variance associated with a grouping - if very small, consider removing it.
  - include random effect because it is part of experiment design
  - include random effect because it is something you are interested in
  - can test models with and without (Baayen et al 2008)
  LRTs comparing models with same FE, differing in RE.

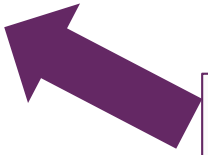
Model checks

- Variance in intercepts and slopes can covary –
  e.g. as people get slower, effects get bigger…

```
Random effects:
 Groups    Name           Variance  Std.Dev. Corr
 item_name (Intercept)    7.340e-04 0.027092
 subjectID (Intercept)    3.018e-03 0.054936
           item_typeword  1.483e-03 0.038507 -0.43
           zLength        1.032e-05 0.003212  0.78 -0.90
           zOrtho_N       1.526e-05 0.003906  0.44 -1.00  0.90
 Residual                 8.784e-03 0.093724
Number of obs: 10254, groups:  item_name, 320; subjectID, 34
```

- Correlations between random effects are high (~0/1)
  - model cannot solve the variance problem
  - is looking for nothing

```
Correlation of Fixed Effects:
            (Intr) zAge   zTOWRE_w zTOWRE_n itm_ty zLngth
zAge         0.012
zTOWRE_wrdc -0.006  0.126
zTOWRE_nnwr  0.014  0.093 -0.644
item_typwrd -0.200  0.000  0.000    0.001
zLength     -0.008  0.000  0.000    0.000    0.031
zOrtho_N    -0.011  0.000  0.000    0.000    0.047  0.602
```



Correlation between different predictors
(check for collinearity)

Mean centering & scaling (i.e. z scores of data) helps to reduce this problem

Reporting models

- Report model comparisons and Likelihood Ratio Tests (LRT) / AIC / BIC for different models
- Explain the model selection choice, based on aims of study and information criteria comparisons
- Summary of fixed effects: like linear models, with CIs
- Summary of random effects: variance and covariance (if applicable)

Reporting models

- Get over reporting ‘one’ analysis of your data

> A number of highly sensible papers now give a similar message – when multiple analyses (and models) are possible, best practice is to report them all and use them as a test of the robustness of effects (Carp, 2012; Steegen et al, 2016; Patel, Burford & Ioannidis, 2015).

p values

Significance and fitting

- Calculation of p values is non-trivial, identifying degrees of freedom
- Risk of false positive with some methods of significance estimation

- Least worst option is Satterthwaite and Kenward-Rogers approximations for dfs for models estimated with REML (Luke, 2016)
- Consider also model comparison vs coefficient significance

Hypothesis driven model selection

Building and choosing models

- "There is no such thing as an all purpose statistical method"
  Nagin & Odgers, 2010, pg 132
- Give rationale for the analysis & model choice
- Be explicit about technical choices
  (e.g. random and fixed effects structures, model selection process, model checking)
- Make programming script available
- Report your basic data that are going into the model (e.g. central tendency, variance, distribution for the DV/response variable)

Building models

- Too many parameters = false positives
- Complex to explain and understand

- Too few….

  Have you included all the important parameters?

  (hypotheses, motivation, previous evidence..)

- Model that will replicate across samples?
- What complexity do you want for your research questions?

Hypothesis driven approach

- “formulate your hypotheses and model without reference to the data (and ideally even prior to their collection)” (Roger Mundry, 2014, MPI EVA)

- The problem of model complexity: even with a few predictors, there is a potentially huge number of different models that can be fit.
  E.g. main effects, interactions, non-linear patterns…

Model building

(1) Identify ‘test’ predictors (fixed effects) that you have hypotheses about

(2) Identify anything you want to control for (control fixed effects)

(3) Identify random effects – this is independent from control vs test predictors. *Random effects are groupings of data you expect to cause systematic variation.*

(4) Compare more complex model against ‘null’ model (e.g. RE and control variables only).

Model building

- Interactions – **think** about what interactions might mean, and why you would test for them.
- Mundry (2014) advice:

Write down:

- why each of the terms in the model is in there
- what the different terms in the model represent (i.e. what they mean, what process they represent)
- what the model and the terms in it could reveal (and not reveal)
- ask yourself (and all researchers involved) if the model represents what you're aiming at

Recommended reading (for clarity)

- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*, 127-135.
- Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. Linguistic Typology, 15(2), 281-320.
- Gelman, A., & Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge University Press.

Recommended reading (for clarity)

- Papers by Reinhold Kliegl (Potsdam)

- Dan Mirman's R resources page
  (Alabama, Birmingham)
  http://www.danmirman.org/r-resources

Thank you for listening

"Essentially, all models are wrong, but some are useful"

George E.P. Box

Box, G. E. P. & Draper, N.R. (1987). *Empirical Model-Building and Response Surfaces*, p. 424, Wiley