Using Linear mixed-effects models – why, when and how

Rob Davies

r.davies1@lancaster.ac.uk

May 2016

Aims for the class

1. Understand the motivation for *linear mixed-effects models* – the requirements of handling multilevel structured data
2. Introduce a multilevel structured dataset
3. Recognize alternative methods for analyzing multilevel structured data
4. Practise running linear mixed-effects models in R
5. Evaluating models using information criteria

Phenomena and data sets in the social sciences often have a multilevel structure

Repeated measures or *clustered data*

- Test the same people multiple times
  - Pre and post treatment
  - Multiple stimuli – everyone sees the same stimuli
  - Repeated testing – follow learning, development within individuals – in longitudinal designs
- Do multi-stage sampling
  - Find (sample) classes or schools – test (sample) children within classes or schools
  - Find (sample) clinics – test (sample) patients within clinics

Phenomena and data sets in the social sciences often have a multilevel structure

Repeated measures or *clustered data*

- Test the same people multiple times
  - Pre and post treatment
  - Multiple stimuli – everyone sees the same stimuli
  - Repeated testing – follow learning, development within individuals – in longitudinal designs
- Do multi-stage sampling
  - Find (sample) classes or schools – test (sample) children within classes or schools
  - Find (sample) clinics – test (sample) patients within clinics

Phenomena and data sets in the social sciences often have a multilevel structure

Repeated measures or *clustered data*

- Test the same people multiple times
  - Pre and post treatment
  - Multiple stimuli – everyone sees the same stimuli
  - Repeated testing – follow learning, development within individuals – in longitudinal designs
- Do multi-stage sampling
  - Find (sample) classes or schools – test (sample) children within classes or schools
  - Find (sample) clinics – test (sample) patients within clinics

Phenomena and data sets in the social sciences often have a multilevel structure

Repeated measures or *clustered data*

- Test the same people multiple times
  - Pre and post treatment
  - Multiple stimuli – everyone sees the same stimuli
  - Repeated testing – follow learning, development within individuals – in longitudinal designs
- Do multi-stage sampling
  - Find (sample) classes or schools – test (sample) children within classes or schools
  - Find (sample) clinics – test (sample) patients within clinics

Phenomena and data sets in the social sciences often have a multilevel structure

Repeated measures or *clustered data*

- Test the same people multiple times
  - Pre and post treatment
  - Multiple stimuli – everyone sees the same stimuli
  - Repeated testing – follow learning, development within individuals – in longitudinal designs
- Do multi-stage sampling
  - Find (sample) classes or schools – test (sample) children within classes or schools
  - Find (sample) clinics – test (sample) patients within clinics

Phenomena and data sets in the social sciences often have a multilevel structure

Repeated measures or *clustered data*

- Test the same people multiple times
  - Pre and post treatment
  - Multiple stimuli – everyone sees the same stimuli
  - Repeated testing – follow learning, development within individuals – in longitudinal designs
- Do multi-stage sampling
  - Find (sample) classes or schools – test (sample) children within classes or schools
  - Find (sample) clinics – test (sample) patients within clinics

Phenomena and data sets in the social sciences often have a multilevel structure

Repeated measures or *clustered data*

- Test the same people multiple times
  - Pre and post treatment
  - Multiple stimuli – everyone sees the same stimuli
  - Repeated testing – follow learning, development within individuals – in longitudinal designs
- Do multi-stage sampling
  - Find (sample) classes or schools – test (sample) children within classes or schools
  - Find (sample) clinics – test (sample) patients within clinics

The key insight: observations are clustered – correlated – *not independent*

Dependence of observations could be treated as a nuisance because an assumption of linear models is that observations are independent so failing to take dependence into account may result in incorrect inferences – the non-independence of observations means you have less information than their total number of suggests you have

Where we are going: *linear mixed-effects modelling*

Capture sources of variance due to *fixed effects* e.g. frequency and *random effects* e.g. differences between sampling units like people or words in intercepts or slopes

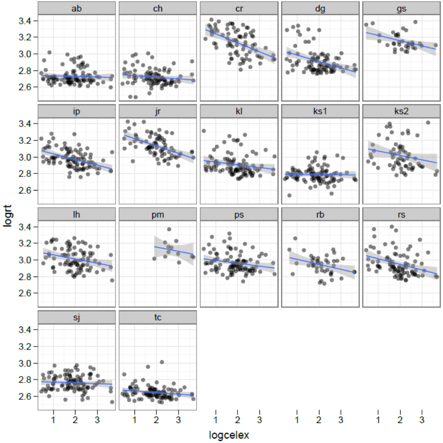


Figure : Effect of word frequency on word naming latencies of adult students

In psychological research, uniformity - the average participant - is a convenient simplification

We often average over individual differences to investigate experimental effects – or we study differences between participant groups averaging over responses to different stimuli

Figure : crowd-korean-CC-Eric-Lafforgue

Both approaches cause problems but neither are necessary with linear mixed-effects models

If we consider variability among individuals or sub-groups, focusing on the average appears risky

Figure : crowd-CC-CatWalker

ML study – A concrete usage example

We can investigate systematic variation in effects by looking for *interactions*

- Person-level effects: how reader attributes affect performance
- Word-level effects: how word attributes affect performance
- *Interactions*: how word-level effects are modulated by person-level effects

ML study – A concrete usage example

We can investigate systematic variation in effects by looking for *interactions*

- Person-level effects: how reader attributes affect performance
- Word-level effects: how word attributes affect performance
- *Interactions*: how word-level effects are modulated by person-level effects

ML study – A concrete usage example

We can investigate systematic variation in effects by looking for *interactions*

- Person-level effects: how reader attributes affect performance
- Word-level effects: how word attributes affect performance
- *Interactions*: how word-level effects are modulated by person-level effects

ML study – A concrete usage example

The data-set – experimental reading task – lexical decision

- All participants saw all 160 words and 160 matched non-words
- Effects of *TOWRE* measures of reading skill, age, ART measure of print exposure
- Effects of word attributes like length in letters, frequency of occurrence
- Interactions between effects of *who* you and effects of *what* you read e.g. TOWRE non-word score * word frequency

Get the data for practice – download and read in the ML datset of responses to words and nonwords

Having read in *subjects.behaviour.items-310114.csv*, use *subset()* to remove errors

```
ML.all <- read.csv("subjects.behaviour.items-310114.csv",
header=T, na.strings = "-999")

ML.all.correct <- subset(ML.all, RT > 200)
```

The logic of Analysis of Variance in linear model terms

$$X_{ij} = \mu + (\mu_j - \mu) + \varepsilon_{ij} = \mu + \tau_j + \varepsilon_{ij} \quad (1)$$

- X_{ij} – the score of person i in condition j
- μ – the mean of all subjects who could be tested in the experiment
- μ_j – the mean score in condition j
- τ_j – the extent to which the mean for condition j is different from the overall mean
- ε_{ij} – the amount to which person i in condition j differs from the mean for that group

The logic of Analysis of Variance in linear model terms

$$X_{ij} = \mu + (\mu_j - \mu) + \varepsilon_{ij} = \mu + \tau_j + \varepsilon_{ij} \quad (1)$$

- X_{ij} – the score of person *i* in condition *j*
- μ – the mean of all subjects who could be tested in the experiment
- μ_j – the mean score in condition *j*
- τ_j – the extent to which the mean for condition *j* is different from the overall mean
- ε_{ij} – the amount to which person *i* in condition *j* differs from the mean for that group

If you take repeated measures then observations will be *dependent* – correlated – within each person

For a slow responder, all their responses will be slow together

- Linear models assume independence of observations
- One way to take the dependence out is by centring all observations for each person on the means (for each person)

If you take repeated measures then observations will be *dependent* – correlated – within each person

For a slow responder, all their responses will be slow together

- Linear models assume independence of observations
- One way to take the dependence out is by centring all observations for each person on the means (for each person)

We can achieve the same thing as centering by accounting for between and within subject differences in our model

$$X_{ij} = \mu + \pi_i + \tau_j + \varepsilon_{ij} \tag{2}$$

- X_{ij} – the score of person i in condition j
- μ – the mean of all subjects who could be tested in the experiment
- π_i – add effect of being subject i – compared to average over all subjects
- τ_j – add effect of being in condition j – compared to average over all conditions
- ε_{ij} – the amount to which person i in condition j differs from the mean for that group

We can achieve the same thing as centering by accounting for between and within subject differences in our model

$$X_{ij} = \mu + \pi_i + \tau_j + \varepsilon_{ij} \tag{2}$$

- X_{ij} – the score of person i in condition j
- μ – the mean of all subjects who could be tested in the experiment
- π_i – add effect of being subject i – compared to average over all subjects
- τ_j – add effect of being in condition j – compared to average over all conditions
- ε_{ij} – the amount to which person i in condition j differs from the mean for that group

A more realistic repeated measures model

Suppose that effects vary between subjects

$$X_{ij} = \mu + \tau_j + \pi_i + \pi_i\tau_j + \varepsilon_{ij} \tag{3}$$

- X_{ij} – the score of person *i* in condition *j* – grand mean
- μ – the mean of all subjects who could be tested in the experiment
- τ_j – add effect of being in condition *j* – compare average over all conditions (grand mean)
- π_i – add effect of being subject *i* – compare average over all subjects (grand mean)
- $\pi_i\tau_j$ – a subject by treatment interaction – different subjects (or words) react to conditions in different ways
- ε_{ij} – the amount to which person *i* in condition *j* differs from the mean for that group

The language as fixed effect fallacy

We need to deal with effects of random variation due to random differences between stimuli as well as differences between people

The language as fixed effect fallacy – a very famous paper by Clark (1973)

- Historically, psychologists tested effects against error variance due to differences between people
- They ignored differences due to stimuli
- This meant they might find significant effects not because there were true differences between conditions
- But because there were also random differences between stimuli in the responses they elicited

The language as fixed effect fallacy – a very famous paper by Clark (1973)

- Historically, psychologists tested effects against error variance due to differences between people
- They ignored differences due to stimuli
- This meant they might find significant effects not because there were true differences between conditions
- But because there were also random differences between stimuli in the responses they elicited

The language as fixed effect fallacy – a very famous paper by Clark (1973)

- Historically, psychologists tested effects against error variance due to differences between people
- They ignored differences due to stimuli
- This meant they might find significant effects not because there were true differences between conditions
- But because there were also random differences between stimuli in the responses they elicited

The language as fixed effect fallacy – a very famous paper by Clark (1973)

- Historically, psychologists tested effects against error variance due to differences between people
- They ignored differences due to stimuli
- This meant they might find significant effects not because there were true differences between conditions
- But because there were also random differences between stimuli in the responses they elicited

A linear model taking into account *the random effects of items*

$$X_{ij} = \mu + \pi_i + \tau_j + \pi_i\tau_j + \beta_k + \pi_i\beta_k + \varepsilon_{ijk} \quad (4)$$

- β_k – effect of word *k* – unexplained differences in average response elicited by different stimuli
- $\pi_i\beta_k$ – the stimulus word by subject interaction – different people respond to different stimuli differently

A linear model taking into account *the random effects of items*

$$X_{ij} = \mu + \pi_i + \tau_j + \pi_i\tau_j + \beta_k + \pi_i\beta_k + \varepsilon_{ijk} \tag{4}$$

- β_k – effect of word k – unexplained differences in average response elicited by different stimuli
- $\pi_i\beta_k$ – the stimulus word by subject interaction – different people respond to different stimuli differently

Taking into account error variance due to subjects and items

Clark's (1973) $minF'$ solution

$$minF' = \frac{MS_{\tau}}{MS_{\pi\tau} + MS_{\beta_k}} = \frac{F_1 F_2}{F_1 + F_2} \tag{5}$$

- You start by *aggregating* your data
- By-subjects data – for each subject, take the average of their responses to all the items
- By-items data – for each item, take the average of all subjects' responses
- You do separate ANOVAs, one for by-subjects (F1) data and one for by-items (F2) data
- You put F1 and F2 together in the calculation of minF'

Taking into account error variance due to subjects and items

Clark's (1973) *minF′* solution

$$minF' = \frac{MS_{\tau}}{MS_{\pi\tau} + MS_{\beta_k}} = \frac{F_1 F_2}{F_1 + F_2} \tag{5}$$

- You start by *aggregating* your data
- By-subjects data – for each subject, take the average of their responses to all the items
- By-items data – for each item, take the average of all subjects' responses
- You do separate ANOVAs, one for by-subjects (F1) data and one for by-items (F2) data
- You put F1 and F2 together in the calculation of minF'

Taking into account error variance due to subjects and items

Clark's (1973) *minF′* solution

$$minF' = \frac{MS_\tau}{MS_{\pi\tau} + MS_{\beta_k}} = \frac{F_1 F_2}{F_1 + F_2} \tag{5}$$

- You start by *aggregating* your data
- By-subjects data – for each subject, take the average of their responses to all the items
- By-items data – for each item, take the average of all subjects' responses
- You do separate ANOVAs, one for by-subjects (F1) data and one for by-items (F2) data
- You put F1 and F2 together in the calculation of minF'

Taking into account error variance due to subjects and items

Clark's (1973) *minF′* solution

$$minF' = \frac{MS_{\tau}}{MS_{\pi\tau} + MS_{\beta_k}} = \frac{F_1 F_2}{F_1 + F_2} \quad (5)$$

- You start by *aggregating* your data
- By-subjects data – for each subject, take the average of their responses to all the items
- By-items data – for each item, take the average of all subjects' responses
- You do separate ANOVAs, one for by-subjects (F1) data and one for by-items (F2) data
- You put F1 and F2 together in the calculation of minF'

Taking into account error variance due to subjects and items

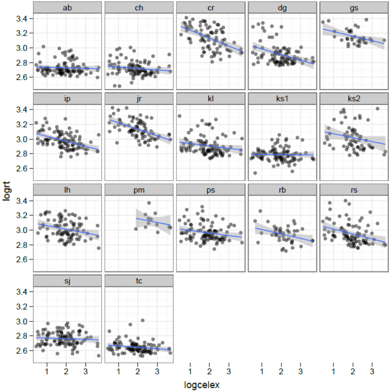
Clark's (1973) *minF′* solution

$$minF' = \frac{MS_{\tau}}{MS_{\pi\tau} + MS_{\beta_k}} = \frac{F_1 F_2}{F_1 + F_2} \tag{5}$$

- You start by *aggregating* your data
- By-subjects data – for each subject, take the average of their responses to all the items
- By-items data – for each item, take the average of all subjects' responses
- You do separate ANOVAs, one for by-subjects (F1) data and one for by-items (F2) data
- You put F1 and F2 together in the calculation of minF'

The problem with minF' is that it is only good for ANOVA and ANOVA is only good for testing the effects of categorical variables – *factors*

Many dealt with the Clark problem, and allowed themselves to include predictors that were continuous variables, by performing regression analyses of by-items data



Repeated measures regression analysis

The problem with regression on by-items means data

- Lorch & Myers (1990) argued there is a problem with multiple regression on by-items mean observations
- The approach reverses the language-as-fixed-effect problem
- Effects are assessed by comparison with an item-based error term
- Effects can be significant because of random variation between subjects in how they responded to items

Repeated measures regression analysis

The problem with regression on by-items means data

- Lorch & Myers (1990) argued there is a problem with multiple regression on by-items mean observations
- The approach reverses the language-as-fixed-effect problem
- Effects are assessed by comparison with an item-based error term
- Effects can be significant because of random variation between subjects in how they responded to items

Repeated measures regression analysis

The problem with regression on by-items means data

- Lorch & Myers (1990) argued there is a problem with multiple regression on by-items mean observations
- The approach reverses the language-as-fixed-effect problem
- Effects are assessed by comparison with an item-based error term
- Effects can be significant because of random variation between subjects in how they responded to items

Repeated measures regression analysis

The problem with regression on by-items means data

- Lorch & Myers (1990) argued there is a problem with multiple regression on by-items mean observations
- The approach reverses the language-as-fixed-effect problem
- Effects are assessed by comparison with an item-based error term
- Effects can be significant because of random variation between subjects in how they responded to items

Repeated measures regression analysis

The problem with regression on by-items means data

- Lorch & Myers (1990) suggested two solutions to the problem with multiple regression on by-items mean observations
  1. Code for subject with $n - 1$ dummy variables and complete regression using subject, and subject by effect, predictors
  2. Perform a regression (linear model) on each subject and complete a t-test or ANOVA on the resulting per-subject coefficients

Repeated measures regression analysis

The problem with regression on by-items means data

- Lorch & Myers (1990) suggested two solutions to the problem with multiple regression on by-items mean observations
  1. Code for subject with $n - 1$ dummy variables and complete regression using subject, and subject by effect, predictors
  2. Perform a regression (linear model) on each subject and complete a t-test or ANOVA on the resulting per-subject coefficients

Repeated measures regression analysis

The problem with regression on by-items means data

- Lorch & Myers (1990) suggested two solutions to the problem with multiple regression on by-items mean observations
  1. Code for subject with $n - 1$ dummy variables and complete regression using subject, and subject by effect, predictors
  2. Perform a regression (linear model) on each subject and complete a t-test or ANOVA on the resulting per-subject coefficients

Analyses of results with simulated data and alternate procedures suggested that the Lorch & Myers by-subjects regression approach does not really work

	$\beta_Z = 0$					
	$\alpha = 0.05$			$\alpha = 0.01$		
	X	Y	Z	X	Y	Z
lmerS: p(t)	0.609	0.990	**0.380**	0.503	0.982	**0.238**
lmerS: p(MCMC)	0.606	0.991	**0.376**	0.503	0.982	**0.239**
subj	0.677	0.995	**0.435**	0.519	0.979	**0.269**
item	0.210	0.873	0.063	0.066	0.670	0.012
lmer: p(t)	0.248	0.898	0.077	0.106	0.752	0.018
lmer: p(MCMC)	0.219	0.879	0.067	0.069	0.674	0.013
	$\beta_Z = 4$					
	$\alpha = 0.05$			$\alpha = 0.01$		
	X	Y	Z	X	Y	Z
lmerS: p(t)	0.597	0.989	0.925	0.488	0.978	0.867
lmerS: p(MCMC)	0.594	0.989	0.924	0.485	0.978	0.869
subj	0.650	0.992	0.931	0.487	0.979	0.868
item	0.183	0.875	0.574	0.055	0.642	0.295
lmer: p(t)	0.219	0.897	0.626	0.089	0.780	0.415
lmer: p(MCMC)	0.190	0.881	0.587	0.061	0.651	0.304

lmer: mixed-effect regression with crossed random effects for subject and item; lmerS: mixed-effect model with subject as random effect; Subj: by-subject regression; Item: by-item regression.

Figure : Baayen et al. (2008): simulated data with or without effects present: item = by-items means regression; lmerS = LM90 per-subject regression approach

If you do repeated measures studies of any kind, you need to take the 'language-as-fixed-effect fallacy' into account – participant and stimulus random effects

Dealing with clustered data – start by ignoring the multilevel structure

You could try to run an ordinary linear model including subject-level and item-level variables

$$RT = \beta_0 + \beta_{wordreadingability} + \beta_{itemtype} + \beta_{word*itemtype} + \epsilon \tag{6}$$

But it is usually incorrect to assume the multilevel structure can be represented by the explanatory variables alone

- What about effect of *random* variation between *participants*?
- What about effect of *random* variation between *stimuli*?

Dealing with clustered data – start by ignoring the multilevel structure

You could try to run an ordinary linear model including subject-level and item-level variables

$$RT = \beta_0 + \beta_{wordreadingability} + \beta_{itemtype} + \beta_{word*itemtype} + \epsilon \quad (6)$$

But it is usually incorrect to assume the multilevel structure can be represented by the explanatory variables alone

- What about effect of *random* variation between *participants*?
- What about effect of *random* variation between *stimuli*?

A more realistic repeated measures model of the item type and reading ability effects

Taking into account random effects of subjects and items

$$RT = \beta_0 + \beta_{wordreadingability} + \beta_{itemtype} + \beta_{ability*itemtype} + \beta_{subject} + \beta_{item} + \beta_{subject*itemtype} + \epsilon \quad (7)$$

- What about effect of random variation between participants?
- Allow *intercept varies* – random effect of subject – some have slower average some have faster average than average overall
- Allow *effect of item type varies* – random effect of subject – subjects can have effects of different direction or size

A more realistic repeated measures model of the item type and reading ability effects

Taking into account random effects of subjects and items

$$RT = \beta_0 + \beta_{wordreadingability} + \beta_{itemtype} + \beta_{ability*itemtype} + \beta_{subject} + \beta_{item} + \beta_{subject*itemtype} + \epsilon \quad (7)$$

- What about effect of random variation between participants?
- Allow *intercept varies* – random effect of subject – some have slower average some have faster average than average overall
- Allow *effect of item type varies* – random effect of subject – subjects can have effects of different direction or size

A more realistic repeated measures model of the item type and reading ability effects

Taking into account random effects of subjects and items

$$RT = \beta_0 + \beta_{wordreadingability} + \beta_{itemtype} + \beta_{ability*itemtype} + \beta_{subject} + \beta_{item} + \beta_{subject*itemtype} + \epsilon \quad (7)$$

- What about effect of random variation between participants?
- Allow *intercept varies* – random effect of subject – some have slower average some have faster average than average overall
- Allow *effect of item type varies* – random effect of subject – subjects can have effects of different direction or size

A more realistic repeated measures model of the item type and reading ability effects

Taking into account random effects of subjects and items

$$RT = \beta_0 + \beta_{wordreadingability} + \beta_{itemtype} + \beta_{ability*itemtype} + \beta_{subject} + \beta_{item} + \beta_{subject*itemtype} + \epsilon \quad (8)$$

- What about effect of random variation between stimuli?
- Allow *intercept varies* – random effect of items – some items harder and elicit slower average response and some easier and elicit faster responses on average than average overall
- Allow *effect of reading ability varies* – within-items effect of subject type can be different for different items

A more realistic repeated measures model of the item type and reading ability effects

Taking into account random effects of subjects and items

$$RT = \beta_0 + \beta_{wordreadingability} + \beta_{itemtype} + \beta_{ability*itemtype} + \beta_{subject} + \beta_{item} + \beta_{subject*itemtype} + \epsilon \quad (8)$$

- What about effect of random variation between stimuli?
- Allow *intercept varies* – random effect of items – some items harder and elicit slower average response and some easier and elicit faster responses on average than average overall
- Allow *effect of reading ability varies* – within-items effect of subject type can be different for different items

A more realistic repeated measures model of the item type and reading ability effects

Taking into account random effects of subjects and items

$$RT = \beta_0 + \beta_{wordreadingability} + \beta_{itemtype} + \beta_{ability*itemtype} + \beta_{subject} + \beta_{item} + \beta_{subject*itemtype} + \epsilon \quad (8)$$

- What about effect of random variation between stimuli?
- Allow *intercept varies* – random effect of items – some items harder and elicit slower average response and some easier and elicit faster responses on average than average overall
- Allow *effect of reading ability varies* – within-items effect of subject type can be different for different items

We do not model random effects directly – we just estimate the *spread* of variation in intercepts

random intercepts – predicted differences (adjustments) between the overall average and the group e.g. person average

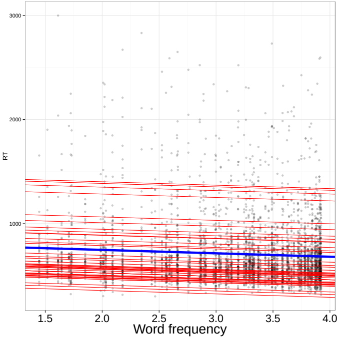


Figure : Learner data – random intercepts, fixed slope in frequency effect

In fact, we can allow for random differences in the *slopes* of the effects of theoretical interest

random slopes – predicted differences (adjustments) between the overall effect and the group e.g. per-person effect

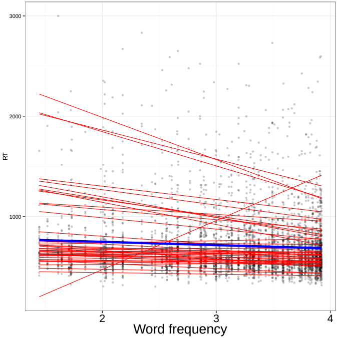


Figure : Individual differences in effect of word frequency on RTs

The advantages of mixed-effects models

- Approach is not restrictive about predictors or data structure
- ANOVA is OK for experimental designs, categorical factors, data sets without missing values
- Can test effects at different levels of hierarchy
- We can allow random effects of both subjects and items – solving the 'language-as-fixed-effect' problem
- Estimation robust to imbalances in data

The advantages of mixed-effects models

- Approach is not restrictive about predictors or data structure
- ANOVA is OK for experimental designs, categorical factors, data sets without missing values
- Can test effects at different levels of hierarchy
- We can allow random effects of both subjects and items – solving the 'language-as-fixed-effect' problem
- Estimation robust to imbalances in data

The advantages of mixed-effects models

- Approach is not restrictive about predictors or data structure
- ANOVA is OK for experimental designs, categorical factors, data sets without missing values
- Can test effects at different levels of hierarchy
- We can allow random effects of both subjects and items – solving the ‘language-as-fixed-effect’ problem
- Estimation robust to imbalances in data

The advantages of mixed-effects models

- Approach is not restrictive about predictors or data structure
  - ANOVA is OK for experimental designs, categorical factors, data sets without missing values
- Can test effects at different levels of hierarchy
- We can allow random effects of both subjects and items – solving the 'language-as-fixed-effect' problem
- Estimation robust to imbalances in data

We focus on building a series of models up to the most complex model supported by the data

- A minimal model of the data might assume that the data we observe can be predicted only by the average value of observations
- The overall average – intercept – and random effects of grouping variables like subjects or stimulus items
- The question is then whether our capacity to predict observations is improved by adding other terms

We focus on building a series of models up to the most complex model supported by the data

- A minimal model of the data might assume that the data we observe can be predicted only by the average value of observations
- The overall average – intercept – and random effects of grouping variables like subjects or stimulus items
- The question is then whether our capacity to predict observations is improved by adding other terms

We focus on building a series of models up to the most complex model supported by the data

- A minimal model of the data might assume that the data we observe can be predicted only by the average value of observations
- The overall average – intercept – and random effects of grouping variables like subjects or stimulus items
- The question is then whether our capacity to predict observations is improved by adding other terms

Examine the fixed effects then the random effects

- Start by examining models varying in the *fixed effects* but constant in the *random effects*
- Fitted using maximum likelihood (*REML* = *FALSE*) method
- Think about simpler models being *nested* inside – i.e. as simplifications of – more complex models
- Add effects of interest – because they were manipulated, are of theoretical or practical interest – fixed effects in series

Examine the fixed effects then the random effects

- Start by examining models varying in the *fixed effects* but constant in the *random effects*
- Fitted using maximum likelihood ($REML = FALSE$) method
- Think about simpler models being *nested* inside – i.e. as simplifications of – more complex models
- Add effects of interest – because they were manipulated, are of theoretical or practical interest – fixed effects in series

Examine the fixed effects then the random effects

- Start by examining models varying in the *fixed effects* but constant in the *random effects*
- Fitted using maximum likelihood ($REML = FALSE$) method
- Think about simpler models being *nested* inside – i.e. as simplifications of – more complex models
- Add effects of interest – because they were manipulated, are of theoretical or practical interest – fixed effects in series

Examine the fixed effects then the random effects

- Start by examining models varying in the *fixed effects* but constant in the *random effects*
- Fitted using maximum likelihood ($REML = FALSE$) method
- Think about simpler models being *nested* inside – i.e. as simplifications of – more complex models
- Add effects of interest – because they were manipulated, are of theoretical or practical interest – fixed effects in series

R code for running a linear mixed-effects model is similar to the code for running a linear model

```
full.lmer0 <- lmer(logrt ~
                        (1|subjectID) + (1|item_name),
data = ML.all.correct, REML = F)
```

- `lmer()` – run a *Linear Mixed-effects model* rather than *lm() linear model*
- `(1|subjectID) + (1|item_name)` specify random effects
- `(1|...)` – random effect on intercepts
- `subjectID` or `item_name` – effects of subjects or items – specified by subject identity code (ID) or item name in coding variables

R code for running a linear mixed-effects model is similar to the code for running a linear model

```
full.lmer0 <- lmer(logrt ~
                     (1|subjectID) + (1|item_name),
data = ML.all.correct, REML = F)
```

- `lmer()` – run a *Linear Mixed-effects model* rather than *lm() linear model*
- `(1|subjectID) + (1|item_name)` specify random effects
- `(1|...)` – random effect on intercepts
- `subjectID` or `item_name` – effects of subjects or items – specified by subject identity code (ID) or item name in coding variables

R code for running a linear mixed-effects model is similar to the code for running a linear model

```
full.lmer0 <- lmer(logrt ~

                  (1|subjectID) + (1|item_name),

data = ML.all.correct, REML = F)
```

- `lmer()` – run a *Linear Mixed-effects model* rather than *lm() linear model*
- `(1|subjectID) + (1|item_name)` specify random effects
- `(1|...)` – random effect on intercepts
- `subjectID` or `item_name` – effects of subjects or items – specified by subject identity code (ID) or item name in coding variables

R code for running a linear mixed-effects model is similar to the code for running a linear model

```
full.lmer0 <- lmer(logrt ~

                    (1|subjectID) + (1|item_name),

data = ML.all.correct, REML = F)
```

- `lmer()` – run a *Linear Mixed-effects model* rather than *lm() linear model*
- `(1|subjectID) + (1|item_name)` specify random effects
- `(1|...)` – random effect on intercepts
- `subjectID` or `item_name` – effects of subjects or items – specified by subject identity code (ID) or item name in coding variables

REML and ML estimation and model comparison

```
full.lmer0 <- lmer(logrt ~
                    (1|subjectID) + (1|item_name),
data = subjects.behaviour.items.nomissing, REML = F)
```

- `REML = F` – maximum likelihood estimation
- Maximum likelihood estimation seeks to find those parameter values that, given the data and our choice of model, make the model's predicted values most similar to the observed values

REML and ML estimation and model comparison

```
full.lmer0 <- lmer(logrt ~
                    (1|subjectID) + (1|item_name),
data = subjects.behaviour.items.nomissing, REML = F)
```

- `REML = F` – maximum likelihood estimation
- Maximum likelihood estimation seeks to find those parameter values that, given the data and our choice of model, make the model's predicted values most similar to the observed values

LMEs – build-up the *fixed* effects while holding the *random* effects constant

To empty model, for the ML study analysis, add subject then item attribute predictors

```
full.lmer1 <- lmer(logrt ~

zAge + zTOWRE_wordacc + zTOWRE_nonwordacc +

(1|subjectID) + (1|item_name),

data = ML.all.correct, REML = F)

summary(full.lmer1)
```

- `zAge + zTOWRE_wordacc` add fixed effects – just as in linear models
- *Fixed effects* reproducible effects – manipulated, selected – of theoretical or practical interest
- `summary(full.lmer1)` – print a model summary

LMEs – build-up the *fixed* effects while holding the *random* effects constant

To empty model, for the ML study analysis, add subject then item attribute predictors

```
full.lmer1 <- lmer(logrt ~

zAge + zTOWRE_wordacc + zTOWRE_nonwordacc +

(1|subjectID) + (1|item_name),

data = ML.all.correct, REML = F)

summary(full.lmer1)
```

- `zAge + zTOWRE_wordacc` add fixed effects – just as in linear models
- *Fixed effects* reproducible effects – manipulated, selected – of theoretical or practical interest
- `summary(full.lmer1)` – print a model summary

LMEs – build-up the *fixed* effects while holding the *random* effects constant

To empty model, for the ML study analysis, add subject then item attribute predictors

```
full.lmer1 <- lmer(logrt ~

zAge + zTOWRE_wordacc + zTOWRE_nonwordacc +

(1|subjectID) + (1|item_name),

data = ML.all.correct, REML = F)

summary(full.lmer1)
```

- `zAge + zTOWRE_wordacc` add fixed effects – just as in linear models
- *Fixed effects* reproducible effects – manipulated, selected – of theoretical or practical interest
- `summary(full.lmer1)` – print a model summary

How do we know if increasing *model complexity* by adding predictors actually helps us to account for variation in outcome values?

Simplicity and parsimony

- Trade-off between too much and too little simplicity in model selection – variable selection
- Models with too many parameters may tend to identify effects that are spurious
- Effects may be *unintuitive* and hard to explain *and* not reproduced in future samples

How do we know if increasing *model complexity* by adding predictors actually helps us to account for variation in outcome values?

Simplicity and parsimony

- Trade-off between too much and too little simplicity in model selection – variable selection
- Models with too many parameters may tend to identify effects that are spurious
- Effects may be *unintuitive* and hard to explain *and* not reproduced in future samples

How do we know if increasing *model complexity* by adding predictors actually helps us to account for variation in outcome values?

Simplicity and parsimony

- Trade-off between too much and too little simplicity in model selection – variable selection
- Models with too many parameters may tend to identify effects that are spurious
- Effects may be *unintuitive* and hard to explain *and* not reproduced in future samples

Akaike Information Criteria: *AIC*

Akaike showed you could estimate information loss in terms of the likelihood of the model given the data

$$AIC = -2ln(l) + 2k \tag{9}$$

- $-2ln(l)$ -2 times the log of the likelihood of the model given the data
- (l) – likelihood
  - Is proportional to the probability of observed data conditional on some hypothesis being true
- k – is the number of parameters in the model

Akaike Information Criteria: *AIC*

Akaike showed you could estimate information loss in terms of the likelihood of the model given the data

$$AIC = -2ln(l) + 2k \tag{9}$$

- $-2ln(l)$ -2 times the log of the likelihood of the model given the data
- (l) – likelihood
- Is proportional to the probability of observed data conditional on some hypothesis being true
- k – is the number of parameters in the model

Akaike Information Criteria: *AIC*

Akaike showed you could estimate information loss in terms of the likelihood of the model given the data

$$AIC = -2ln(l) + 2k \tag{9}$$

- $-2ln(l)$ -2 times the log of the likelihood of the model given the data
- (l) – likelihood
- Is proportional to the probability of observed data conditional on some hypothesis being true
- k – is the number of parameters in the model

Akaike Information Criteria: *AIC*

Akaike showed you could estimate information loss in terms of the likelihood of the model given the data

$$AIC = -2ln(l) + 2k \tag{9}$$

- $-2ln(l)$ -2 times the log of the likelihood of the model given the data
- (l) – likelihood
  - Is proportional to the probability of observed data conditional on some hypothesis being true
- k – is the number of parameters in the model

Bayesian Information Criteria: *BIC*

Schwartz proposed an alternative estimate

$$BIC = -2ln(l) + kln(N) \tag{10}$$

- $-2ln(l)$ – -2 times the log of the likelihood of the model given the data
- $+kln(N)$ – is the number of parameters in the model times the log of the sample size
- Crudely, the penalty for greater complexity is heavier in BIC

Bayesian Information Criteria: *BIC*

Schwartz proposed an alternative estimate

$$BIC = -2ln(l) + kln(N) \tag{10}$$

- $-2ln(l)$ – -2 times the log of the likelihood of the model given the data
- $+kln(N)$ – is the number of parameters in the model times the log of the sample size
- Crudely, the penalty for greater complexity is heavier in BIC

Bayesian Information Criteria: *BIC*

Schwartz proposed an alternative estimate

$$BIC = -2ln(l) + kln(N) \tag{10}$$

- $-2ln(l)$ – -2 times the log of the likelihood of the model given the data
- $+kln(N)$ – is the number of parameters in the model times the log of the sample size
- Crudely, the penalty for greater complexity is heavier in BIC

Likelihood ratio test comparison

- The test statistic is the comparison of the likelihood of the simpler model with the more complex model
- Comparison by division $2log\frac{likelihood-complex}{likelihood-simple}$
- The likelihood ratio is compared to the χ^2 distribution for a significance test
- Assuming the null hypothesis that the simpler model is adequate
- With degrees of freedom equal to the difference in the number of parameters of the models being compared

Likelihood ratio test comparison

- The test statistic is the comparison of the likelihood of the simpler model with the more complex model
- Comparison by division $2log\frac{likelihood-complex}{likelihood-simple}$
- The likelihood ratio is compared to the χ^2 distribution for a significance test
- Assuming the null hypothesis that the simpler model is adequate
- With degrees of freedom equal to the difference in the number of parameters of the models being compared

Likelihood ratio test comparison

- The test statistic is the comparison of the likelihood of the simpler model with the more complex model
- Comparison by division $2log\frac{likelihood-complex}{likelihood-simple}$
- The likelihood ratio is compared to the χ^2 distribution for a significance test
- Assuming the null hypothesis that the simpler model is adequate
- With degrees of freedom equal to the difference in the number of parameters of the models being compared

Likelihood ratio test comparison

- The test statistic is the comparison of the likelihood of the simpler model with the more complex model
- Comparison by division $2log\frac{likelihood-complex}{likelihood-simple}$
- The likelihood ratio is compared to the χ^2 distribution for a significance test
- Assuming the null hypothesis that the simpler model is adequate
- With degrees of freedom equal to the difference in the number of parameters of the models being compared

Likelihood ratio test comparison

- The test statistic is the comparison of the likelihood of the simpler model with the more complex model
- Comparison by division $2log\frac{likelihood-complex}{likelihood-simple}$
- The likelihood ratio is compared to the χ^2 distribution for a significance test
- Assuming the null hypothesis that the simpler model is adequate
- With degrees of freedom equal to the difference in the number of parameters of the models being compared

Model comparison

- AIC, BIC and LRT comparisons should be consistent in their indications – which model to prefer
- Can be tricky where dealing with complex sets of predictors – indicators may diverge
- Remember that BIC may penalise complexity more heavily – especially if conducting exploratory research
- Remember that may be obliged to include all effects built-in by design – if conducting confirmatory study

```
> anova(full.lmer0, full.lmer1)
Data: subjects.behaviour.items.nomissing
Models:
full.lmer0: logrt ~ (1 | subjectID) + (1 | item_name)
full.lmer1: logrt ~ cAge + cTOWRE_wordacc + cTOWRE_nonwordacc + cART_HRminusFR +
full.lmer1:     (1 | subjectID) + (1 | item_name)
           Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
full.lmer0  4 -17981 -17952 8994.4   -17989
full.lmer1  8 -17983 -17925 8999.4   -17999 10.116      4    0.03852 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model comparison

- AIC, BIC and LRT comparisons should be consistent in their indications – which model to prefer
- Can be tricky where dealing with complex sets of predictors – indicators may diverge
- Remember that BIC may penalise complexity more heavily – especially if conducting exploratory research
- Remember that may be obliged to include all effects built-in by design – if conducting confirmatory study

```
> anova(full.lmer0, full.lmer1)
Data: subjects.behaviour.items.nomissing
Models:
full.lmer0: logrt ~ (1 | subjectID) + (1 | item_name)
full.lmer1: logrt ~ cAge + cTOWRE_wordacc + cTOWRE_nonwordacc + cART_HRminusFR +
full.lmer1:     (1 | subjectID) + (1 | item_name)
           Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
full.lmer0  4 -17981 -17952 8994.4   -17989
full.lmer1  8 -17983 -17925 8999.4   -17999 10.116      4    0.03852 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Model comparison

- AIC, BIC and LRT comparisons should be consistent in their indications – which model to prefer
- Can be tricky where dealing with complex sets of predictors – indicators may diverge
- Remember that BIC may penalise complexity more heavily – especially if conducting exploratory research
- Remember that may be obliged to include all effects built-in by design – if conducting confirmatory study

```
> anova(full.lmer0, full.lmer1)
Data: subjects.behaviour.items.nomissing
Models:
full.lmer0: logrt ~ (1 | subjectID) + (1 | item_name)
full.lmer1: logrt ~ cAge + cTOWRE_wordacc + cTOWRE_nonwordacc + cART_HRminusFR +
full.lmer1:     (1 | subjectID) + (1 | item_name)
           Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
full.lmer0  4 -17981 -17952 8994.4   -17989
full.lmer1  8 -17983 -17925 8999.4   -17999 10.116      4    0.03852 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model comparison

- AIC, BIC and LRT comparisons should be consistent in their indications – which model to prefer
- Can be tricky where dealing with complex sets of predictors – indicators may diverge
- Remember that BIC may penalise complexity more heavily – especially if conducting exploratory research
- Remember that may be obliged to include all effects built-in by design – if conducting confirmatory study

```
> anova(full.lmer0, full.lmer1)
Data: subjects.behaviour.items.nomissing
Models:
full.lmer0: logrt ~ (1 | subjectID) + (1 | item_name)
full.lmer1: logrt ~ cAge + cTOWRE_wordacc + cTOWRE_nonwordacc + cART_HRminusFR +
full.lmer1:     (1 | subjectID) + (1 | item_name)
           Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
full.lmer0  4 -17981 -17952 8994.4   -17989
full.lmer1  8 -17983 -17925 8999.4   -17999 10.116      4    0.03852 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model comparisons among mixed-effects models – use *anova()* function

```
anova(full.lmer0, full.lmer1)
```

- `anova(...,....)` – compare pairs of models
- `full.lmer0` – a simpler model – more limited assumptions about sources of variance
- `full.lmer1` – a more complex model – more predictors – includes simpler model as a special case

Model comparisons among mixed-effects models – use *anova()* function

```
anova(full.lmer0, full.lmer1)
```

- `anova(...,....)` – compare pairs of models
- `full.lmer0` – a simpler model – more limited assumptions about sources of variance
- `full.lmer1` – a more complex model – more predictors – includes simpler model as a special case

Model comparisons among mixed-effects models – use *anova()* function

```
anova(full.lmer0, full.lmer1)
```

- `anova(...,....)` – compare pairs of models
- `full.lmer0` – a simpler model – more limited assumptions about sources of variance
- `full.lmer1` – a more complex model – more predictors – includes simpler model as a special case

Running the *anova*(,) comparison will deliver AIC, BIC, and likelihood comparisons for varying models

```
> anova(full.lmer1, full.lmer2)
Data: subjects.behaviour.items.nomissing
Models:
full.lmer1: logrt ~ cAge + cTOWRE_wordacc + cTOWRE_nonwordacc + cART_HRminusFR +
full.lmer1:     (1 | subjectID) + (1 | item_name)
full.lmer2: logrt ~ cAge + cTOWRE_wordacc + cTOWRE_nonwordacc + cART_HRminusFR +
full.lmer2:     item_type + cLength + cOrtho_N + cBG_Mean + (1 | subjectID) +
full.lmer2:     (1 | item_name)
           Df    AIC    BIC logLik deviance  Chisq Chi Df Pr(>Chisq)
full.lmer1  8 -17983 -17925 8999.4   -17999
full.lmer2 12 -18319 -18232 9171.3   -18343 343.81      4  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Figure : Comparison of model with subject attribute predictors and model also with item effects

Model selection and judgment

Using Information Criteria statistics

- Compare a simpler model, example: model 0, just random effects on intercepts; model 1, just subject main effects; model 2, subject and item main effects
- If the more complex model better approximates reality then it will be more likely given the data
- AIC will be closer to negative infinity
- e.g. 10 is better than 1000, -1000 better than -10
- Over and above any measure of the complexity of the model

Model selection and judgment

Using Information Criteria statistics

- Compare a simpler model, example: model 0, just random effects on intercepts; model 1, just subject main effects; model 2, subject and item main effects
- If the more complex model better approximates reality then it will be more likely given the data
- AIC will be closer to negative infinity
- e.g. 10 is better than 1000, -1000 better than -10
- Over and above any measure of the complexity of the model

Model selection and judgment

Using Information Criteria statistics

- Compare a simpler model, example: model 0, just random effects on intercepts; model 1, just subject main effects; model 2, subject and item main effects
- If the more complex model better approximates reality then it will be more likely given the data
- AIC will be closer to negative infinity
- e.g. 10 is better than 1000, -1000 better than -10
- Over and above any measure of the complexity of the model

Model selection and judgment

Using Information Criteria statistics

- Compare a simpler model, example: model 0, just random effects on intercepts; model 1, just subject main effects; model 2, subject and item main effects
- If the more complex model better approximates reality then it will be more likely given the data
- AIC will be closer to negative infinity
- e.g. 10 is better than 1000, -1000 better than -10
- Over and above any measure of the complexity of the model

Model selection and judgment

Using AIC and BIC

- Compare a simpler model: model 1, just main effects; model 2, main effects plus interactions
- If the more complex model better approximates reality then it will be more likely given the data
- AIC and BIC should move in the same direction – usually will
- AIC will tend to allow more complex models – may be necessary when want more accurate predictions
- BIC will tend to favour simpler models – may be necessary when seek models that replicate over the long run

Model selection and judgment

Using AIC and BIC

- Compare a simpler model: model 1, just main effects; model 2, main effects plus interactions
- If the more complex model better approximates reality then it will be more likely given the data
- AIC and BIC should move in the same direction – usually will
- AIC will tend to allow more complex models – may be necessary when want more accurate predictions
- BIC will tend to favour simpler models – may be necessary when seek models that replicate over the long run

Model selection and judgment

Using AIC and BIC

- Compare a simpler model: model 1, just main effects; model 2, main effects plus interactions
- If the more complex model better approximates reality then it will be more likely given the data
- AIC and BIC should move in the same direction – usually will
- AIC will tend to allow more complex models – may be necessary when want more accurate predictions
- BIC will tend to favour simpler models – may be necessary when seek models that replicate over the long run

Model selection and judgment

Using AIC and BIC

- Compare a simpler model: model 1, just main effects; model 2, main effects plus interactions
- If the more complex model better approximates reality then it will be more likely given the data
- AIC and BIC should move in the same direction – usually will
- AIC will tend to allow more complex models – may be necessary when want more accurate predictions
- BIC will tend to favour simpler models – may be necessary when seek models that replicate over the long run

Model selection and judgment

Using AIC and BIC

- Compare a simpler model: model 1, just main effects; model 2, main effects plus interactions
- If the more complex model better approximates reality then it will be more likely given the data
- AIC and BIC should move in the same direction – usually will
- AIC will tend to allow more complex models – may be necessary when want more accurate predictions
- BIC will tend to favour simpler models – may be necessary when seek models that replicate over the long run

Reporting standards

Likelihood Ratio Test comparisons

- Recommendations (Bates et al., 2015; glmm.wikidot) – to compare models of varying complexity
- Use Likelihood Ratio Test

Reporting standards

Likelihood Ratio Test comparisons

- Recommendations (Bates et al., 2015; glmm.wikidot) – to compare models of varying complexity
- Use Likelihood Ratio Test