Statistics for Linguistics

Session 7

Linear Mixed Effects Regression Models

Part 2 – Modelling

A Thought Experiment

- Imagine you are a parent of 6 children
- For some years, you have recorded their height on their birthdays

	Kate	Eve	Tess	Max	Neil	Jack
12	149.8	156.3	145.8	149.1	143.3	159.3
13	156.7	163.2	153.7	156.2	150.4	166.4
14	158.7	165.2	160.7	163.8	158.0	174
15	159.7	166.2	162.7	170.1	164.3	180.3
16	162.5	169.0	167.5	173.4	167.6	183.6
17	162.5	169.0	172.5	175.2	169.4	185.4
18	163.0	169.5	178.0	175.7	169.9	185.9

A Thought Experiment

- Using your knowledge on **simple linear regression**, you fit a model:

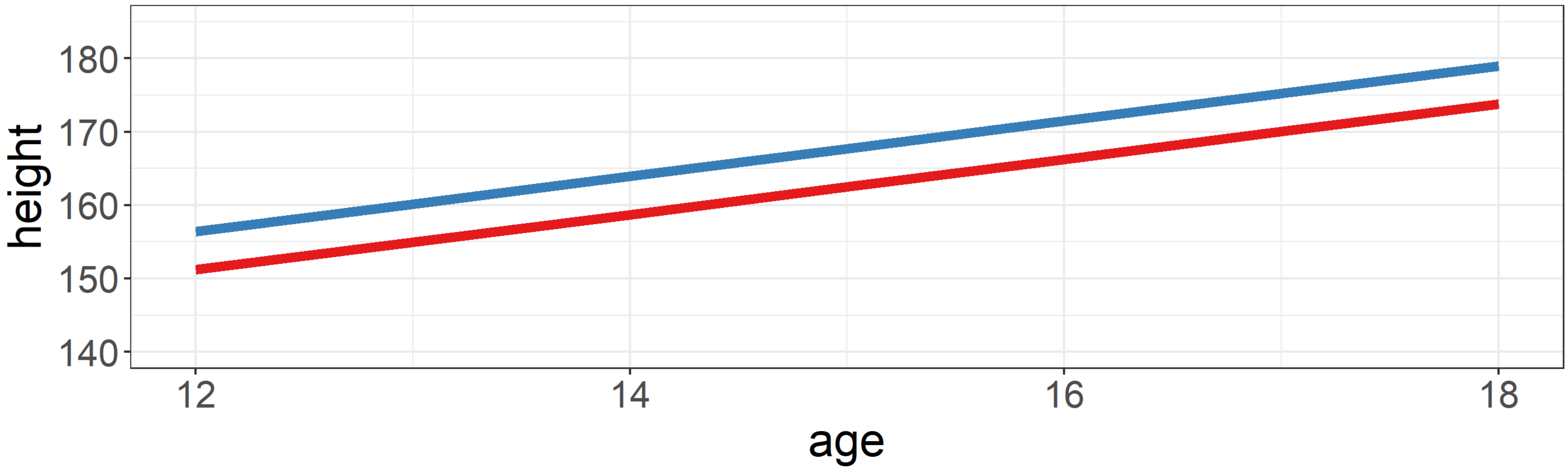
```
lm(height ~ age, data_h)
```

- According to your model, all children grow with the same *speed* (i.e. slope)

A Thought Experiment

- Using your knowledge on **multiple linear regression**, you fit a model:

```
lm(height ~ age + bsex, data_h)
```



- According to your model, all children grow with the same *speed* (i.e. slope), but girls are constantly shorter than boys (i.e. intercept)

A Thought Experiment

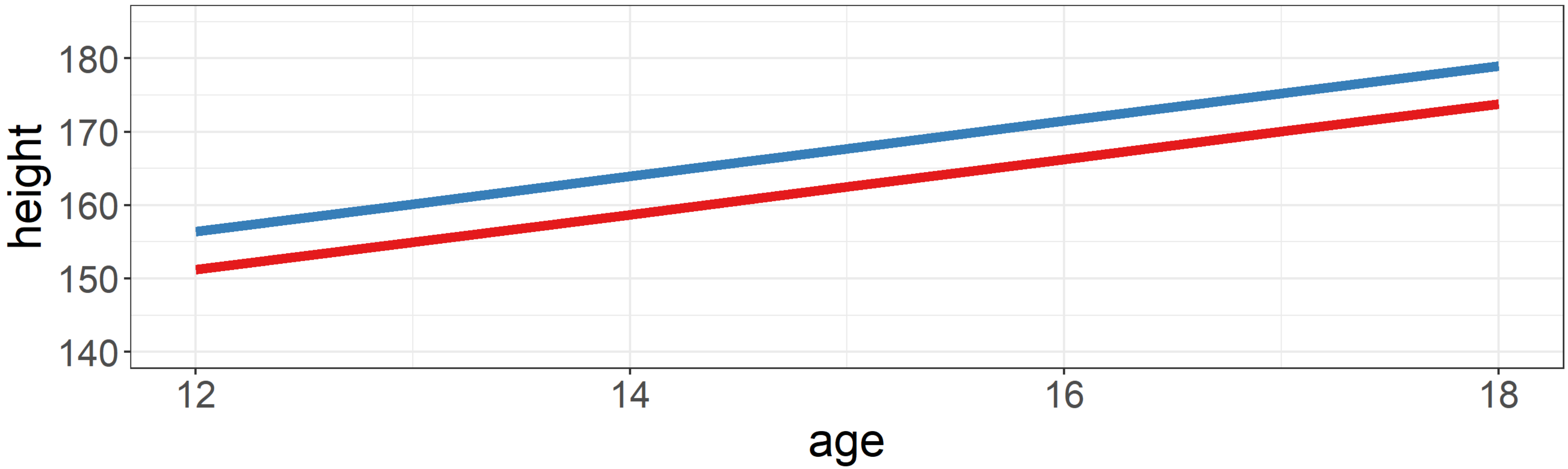
Are they, tough?

	Kate	Eve	Tess	Max	Neil	Jack
12	149.8	156.3	145.8	149.1	143.3	159.3
13	156.7	163.2	153.7	156.2	150.4	166.4
14	158.7	165.2	160.7	163.8	158.0	174
15	159.7	166.2	162.7	170.1	164.3	180.3
16	162.5	169.0	167.5	173.4	167.6	183.6
17	162.5	169.0	172.5	175.2	169.4	185.4
18	163.0	169.5	178.0	175.7	169.9	185.9

A Thought Experiment

- Using your knowledge on **multiple linear regression**, you fit a model:

```
lm(height ~ age + bsex, data_h)
```

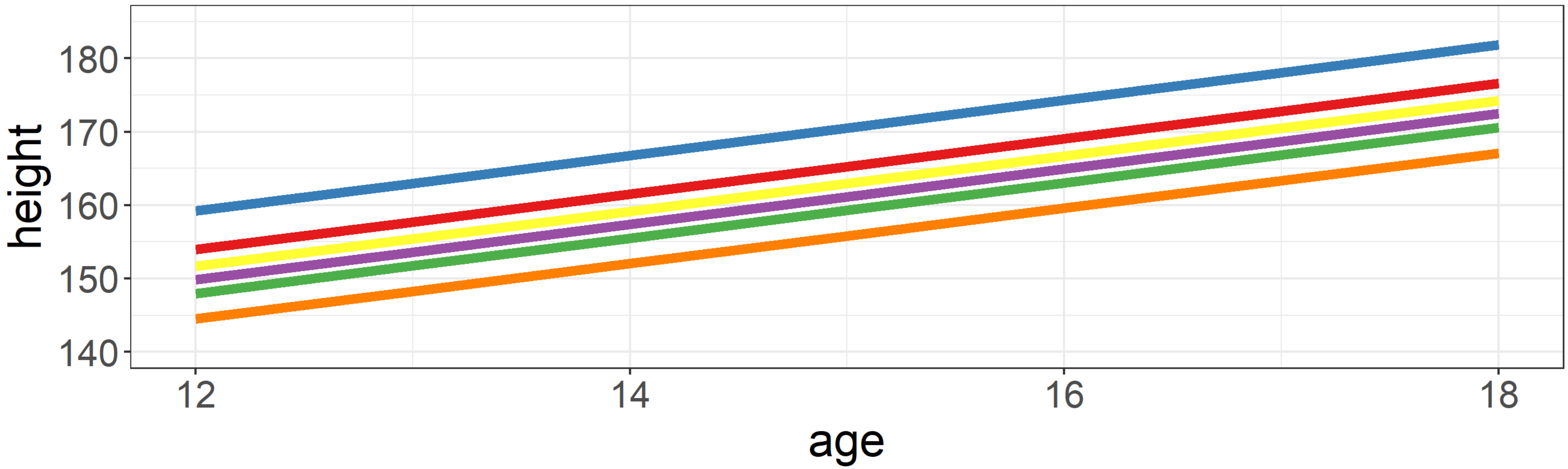


- **Question:** What do we need to do to make this model more realistic?

A Thought Experiment

- Using **linear mixed effects regression**, we fit a **random intercept** model:

```
lm(height ~ age + bsex + (1 | name), data_h)
```



- According to our model, each child starts with an individual height (i.e. intercept) while they grow with the same *speed* (i.e. slope)

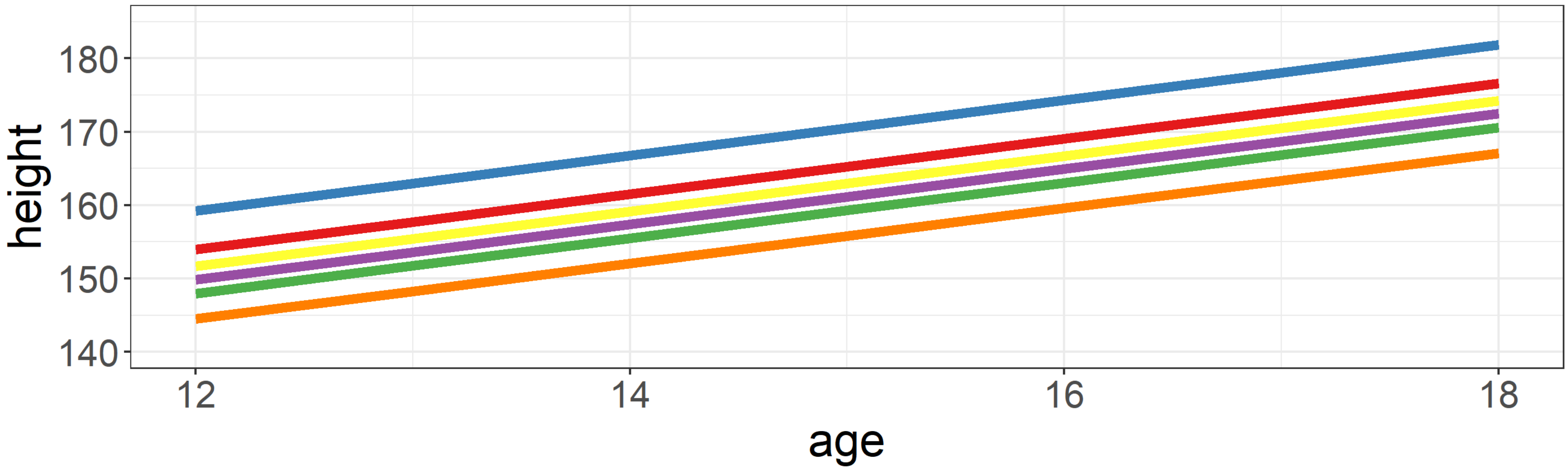
A Thought Experiment

Do they, tough?

	Kate		Tess	
12	149.8		145.8	
13	156.7	**6.9**	153.7	**4.9**
14	158.7	**2.0**	160.7	**7.0**
15	159.7	**1.0**	162.7	**2.0**
16	162.5	**2.5**	167.5	**4.8**
17	162.5	**0.0**	172.5	**5.0**
18	163.0	**0.5**	178.0	**5.5**

A Thought Experiment

- While this already is more realistic, it still assumes that all children grow with an identical speed, i.e. change of height per year

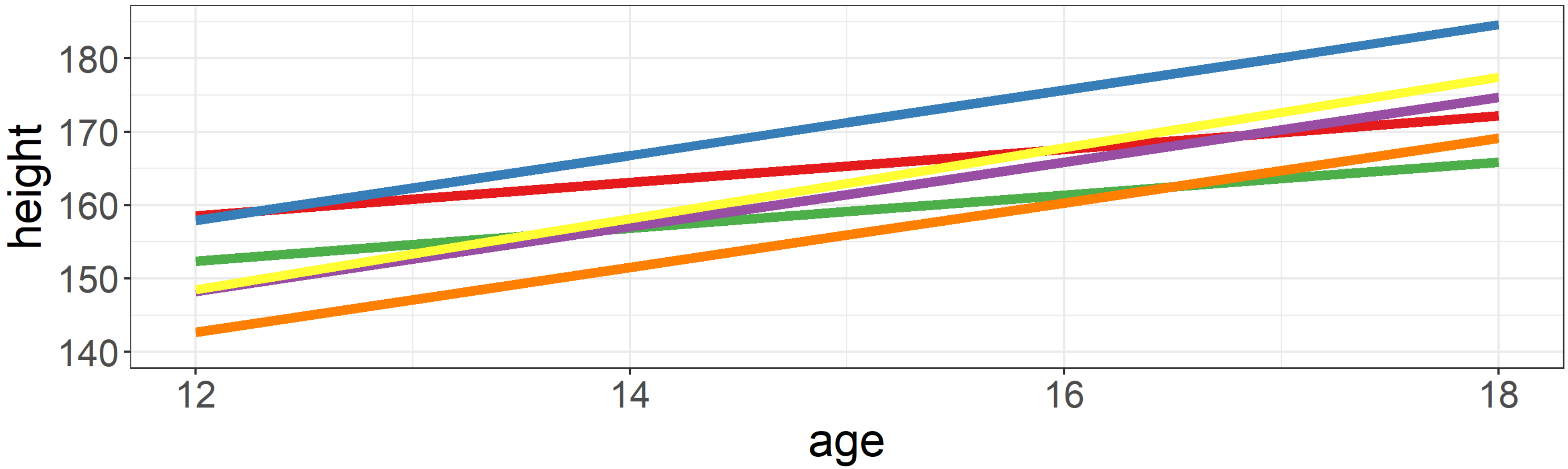


- **Question:** What do we need to do to make this model more realistic?

A Thought Experiment

- Using **linear mixed effects regression**, we fit a **random slope** model:

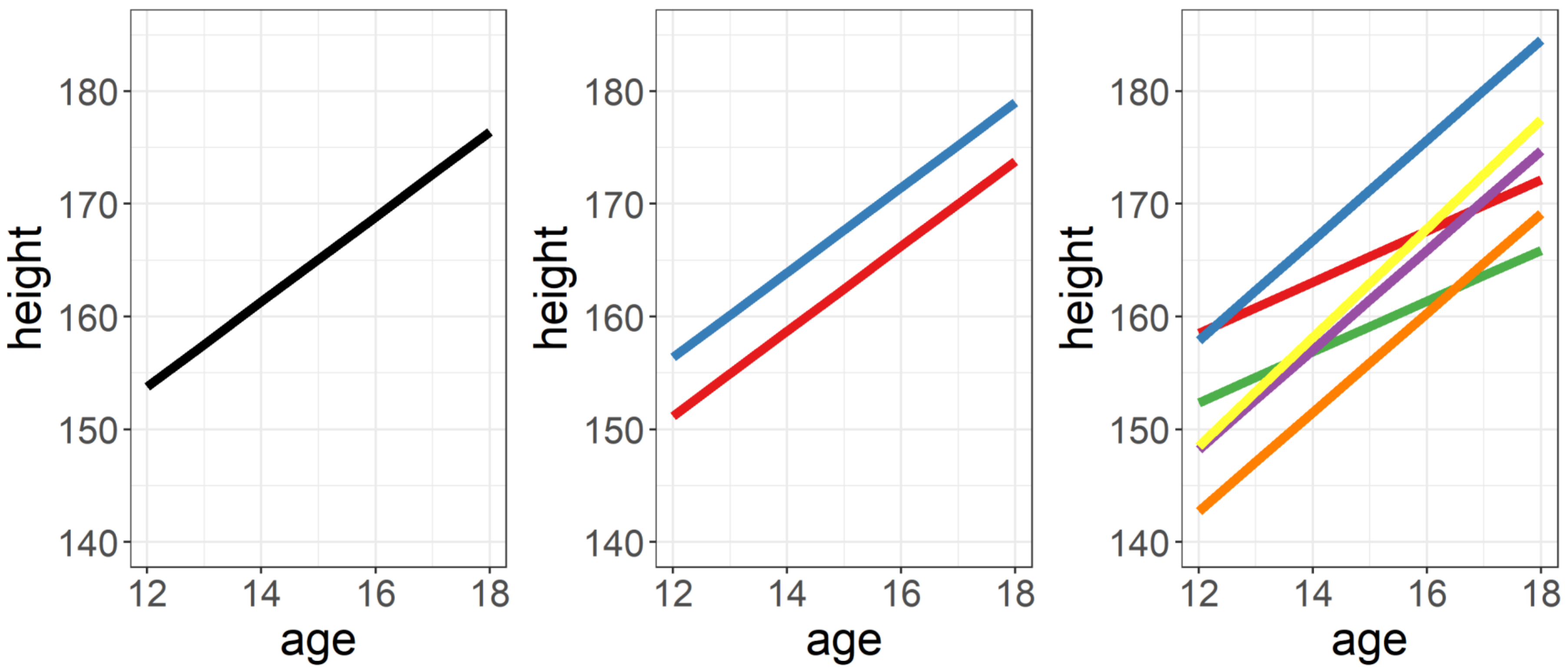
```
lm(height ~ age + bsex + (age | name), data_h)
```



- According to our model, each child starts with an individual height (i.e. intercept) and they grow with different *speed* (i.e. slope)

A Thought Experiment

- Apparently, **linear mixed effects regression models** capture reality better than simple or multiple linear regression models



Simple Linear Regression Formula

continuous
dependent variable

predictor variable

$$Y = \beta_1 + \beta_2 X + \epsilon$$

intercept

slope

error term /
residuals

Multiple Linear Regression Formula

continuous
dependent variable

predictor variable 2

predictor variable 1

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \epsilon$$

intercept

slope
of variable 1

slope
of variable 2

error term /
residuals

Random Intercept Formula

continuous
dependent variable

predictor variable 1

$$Y = \beta_0 + u_0 + \beta_1 X_1 + \cdots + \beta_i X_i + \epsilon$$

intercept

intercept adjustment
e.g. per name

slope
of variable 1

error term /
residuals

Random Slope Formula

continuous dependent variable

predictor variable 1

$$Y = \beta_0 + u_0 + (u_1\beta_1)X_1 + \cdots + \epsilon$$

intercept

intercept adjustment
e.g. per name

slope adjustment
e.g. for age

slope of variable 1

error term / residuals

Example Data

- For the following illustrations we will use data collected in a study on

Compensatory Vowel Shortening in German[1]

- Stressed vowels are shortened depending on how many segments follow within the same word
- e.g. /a:/ in /**ma:**/ is longer than in /**ma:m**/

  /a:/ in /**ma:m**/ is longer than in /**ma:ms**/

  /a:/ in /**ma:ms**/ is longer than in /**ma:ms.la**/

[1]Schmitz et al. (2018)

Example Data

- For the following illustrations we will use data collected in a study on

Compensatory Vowel Shortening in German[1]

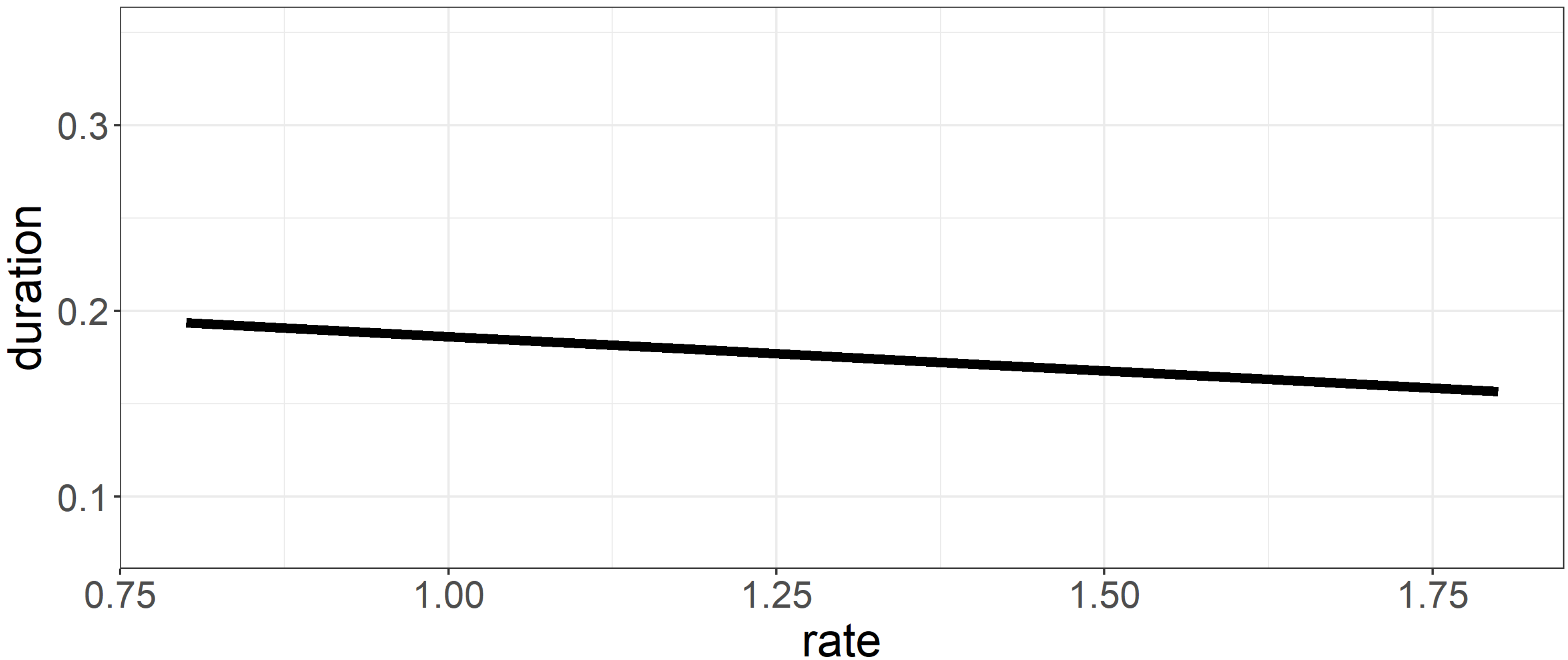
- Independent of shortening, open vowels should be shorter than mid vowels, which in turn should be shorter than closed vowels
- i.e. /**i:**, **u:**/ < /**e:**, **o:**/ < /**a:**/

[1]Schmitz et al. (2018)

Example Data

- So far, we tried to model vowel durations with

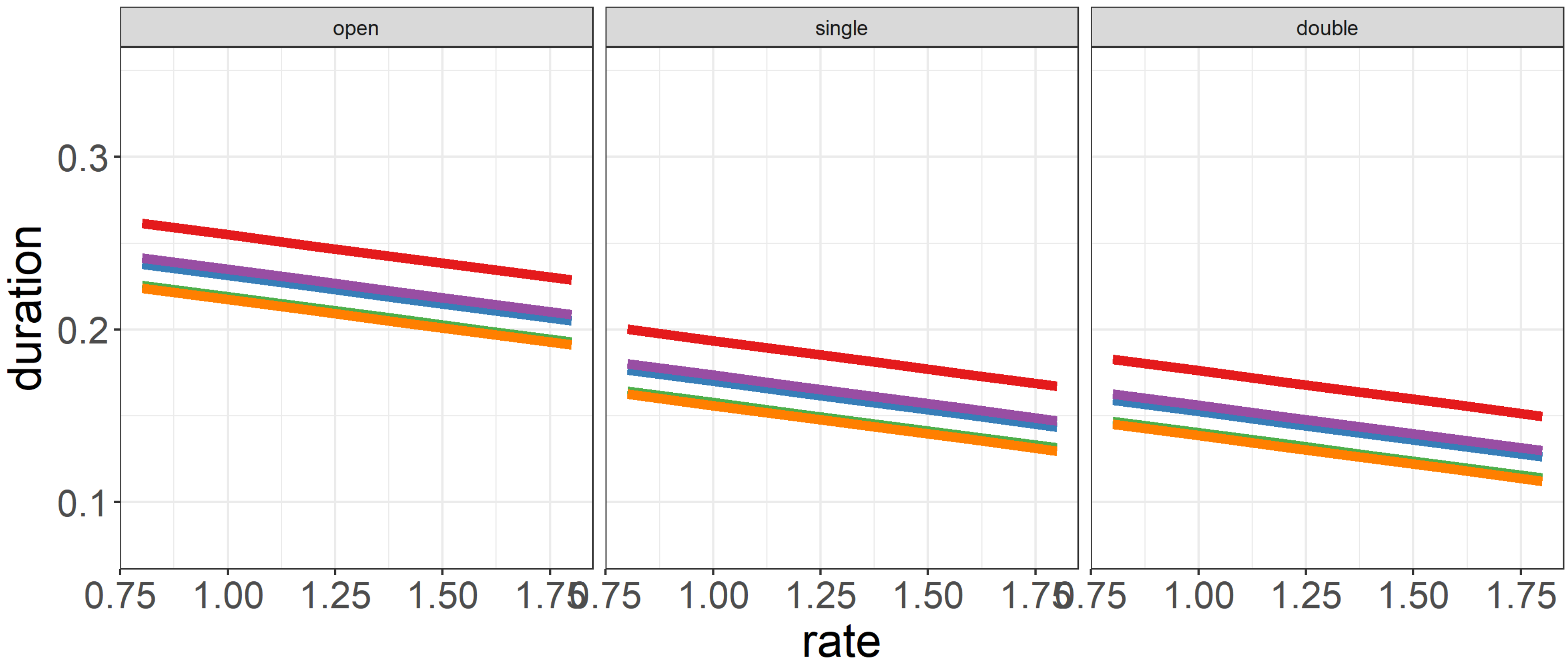
```
lm(duration ~ rate, data_v)
```



Example Data

- … and with

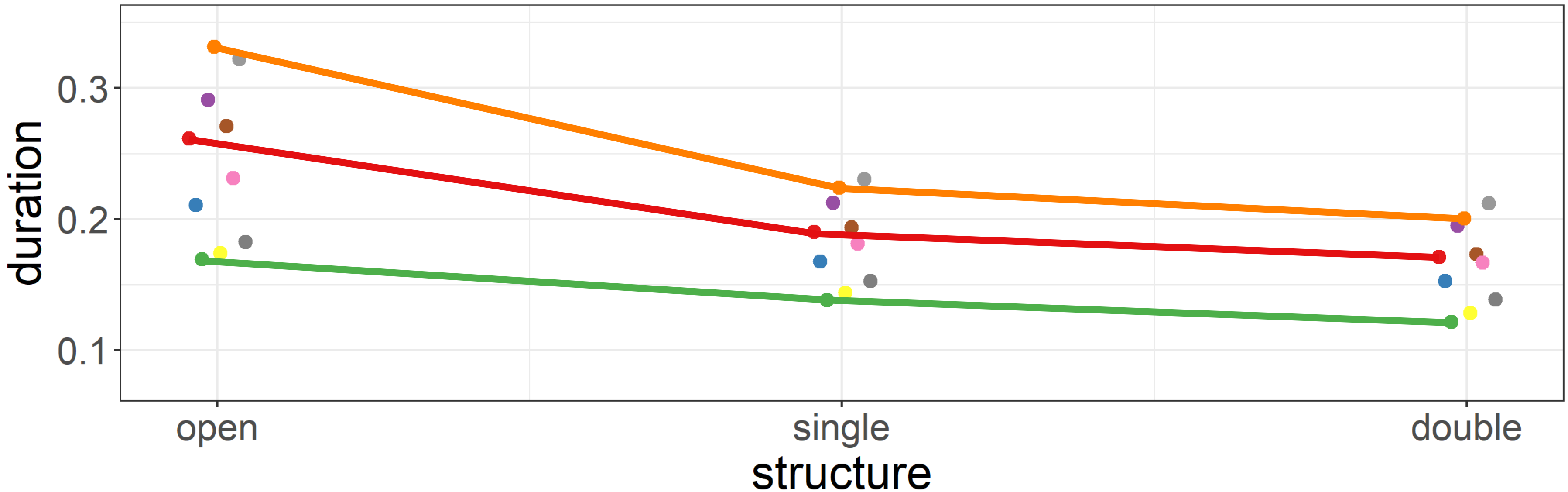
```
lm(duration ~ rate + structure + vowel, data_v)
```



Example Data

- However, we now know that we can – and should – also include random effects, for example

```
lmer(duration ~ rate + structure + vowel +
(structure | speaker), data_v)
```



Fixed & Random Effects

- In linear mixed effects regression, we work with two types of predictors

1. **fixed effects**
   - explanatory variables under investigation
   - repeatable
2. **random effects**
   - further sources of variation
   - random / not repeatable

Fixed & Random Effects

- Our model on vowel duration

predictors

`duration`	~	`rate + structure + vowel`	+	`(structure | speaker)`
dependent variable		independent variables		confounding variables
		fixed effects		random effects

Fixed & Random Effects

- Thus, when analysing new data, we must not only decide which variables we wish to use…
- But also which variables are better fit as **fixed effects** and which should be included as **random effects**

Fixed & Random Effects

- Typical **fixed effect** variables are measures for which predictions exist in the literature, e.g.
  - frequency — more frequent = shorter
  - neighbourhoods — denser = shorter
  - measured durations — longer base = longer affixes
  - etc.

Fixed & Random Effects

- Typical **random effect** variables are measures for which we have no fixed predictions, e.g.
  - speaker — no two speakers are the same
  - items — no two words are the same
  - order of items — can cause all sorts of random stuff
  - etc.

Fixed & Random Effects

- So, coming back to our example on vowel durations, we have the following predictor variables:
  - rate — faster = shorter
  - structure — more complex = shorter
  - vowel — more open = longer
  - speaker — ???
  - word — ???

Fixed & Random Effects

- So, coming back to our example on vowel durations, we have the following predictor variables:

rate	faster = shorter	**fixed effects**
structure	more complex = shorter	
vowel	more open = longer	
speaker	???	**random effects**
word	???	

Multiple Linear Regression in R

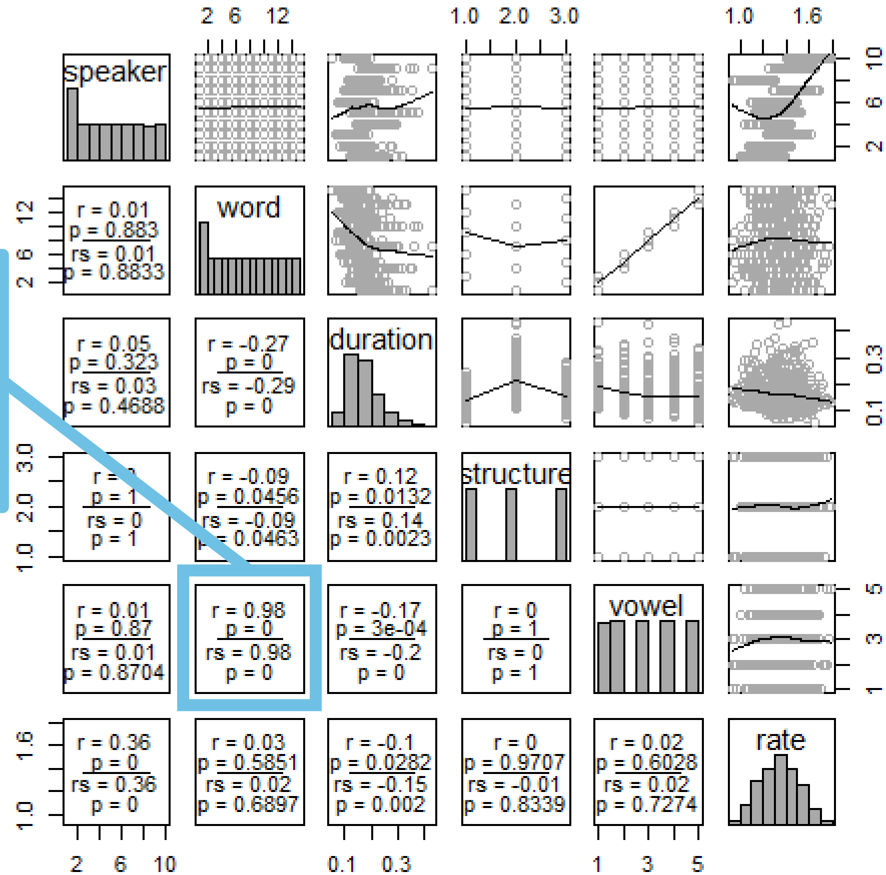
- More variables make the modelling procedure a little more time consuming
- Typical steps are
  1. Check variable distributions
  2. Check correlations; take action to avoid collinearity issues
  3. Create a ‘full’ model
  4. Find the ‘best’ model
  5. Check assumptions
  6. Interpret the model

Multiple Linear Regression in R

- More variables make the modelling procedure a little more time consuming
- Typical steps are
  1. Check variable distributions ✓
  2. Check correlations; take action to avoid collinearity issues
  3. Create a ‘full’ model
  4. Find the ‘best’ model
  5. Check assumptions
  6. Interpret the model

Step 2: Avoid Collinearity Issues

not a problem as one is a fixed effect and one is a random effect variable



Step 3: Full Model Creation

- Let's create the full model:

```
library(lme4)

mdl = lmer(durationLog ~ structure + vowel + rate +
                         (structure | speaker),
                         data_v)
```

Step 4: Find Best Model

- As before, we can use the `step()` function to do this

```
step(mdl)

...

...

...

Model found:

durationLog ~ structure + vowel + (structure | speaker)
```

Step 5: Check Assumptions

- Multiple Linear Regression Models follow the same **assumptions** as Simple Linear Regression Models
  - Linearity
  - Homoscedasticity
  - Normality
  - Independence
- **Disclaimer:** The SfL data sets are too small to create meaningful mixed-effects models, thus assumptions are mostly violated

Step 6: Interpretation

- In general, we are interested in two things
  1. the ***p*-values** of individual predictors
  2. the **effects** of the individual predictors

Step 6: Interpretation – *p*-Values

1. Using the `anova()` function, we can obtain ***p*-values**

```
Type III Analysis of Variance Table with Satterthwaite's method
          Sum Sq Mean Sq NumDF  DenDF F value    Pr(>F)    
structure 3.6769 1.83845     2  11.76 100.222 4.111e-08 ***
vowel     3.6894 0.92234     4 423.03  50.281 < 2.2e-16 ***
```

Step 6: Interpretation – Effects

2. Using the `summary()` function, we can take a closer look at the **effects** of the individual predictors

```
Fixed effects:
                 Estimate Std. Error        df t value Pr(>|t|)    
(Intercept)      -1.83695    0.07645   9.54001 -24.029 7.41e-10 ***
structureopen     0.43271    0.03064   9.03857  14.125 1.82e-07 ***
structuresingle   0.12182    0.01797  18.54869   6.777 2.04e-06 ***
vowele           -0.15059    0.02031 423.07910  -7.414 6.73e-13 ***
voweli           -0.24876    0.02031 423.07910 -12.248  < 2e-16 ***
vowelo           -0.13248    0.02031 423.07910  -6.523 1.98e-10 ***
vowelu           -0.24566    0.02031 423.07910 -12.095  < 2e-16 ***
```

Step 6: Interpretation – Effects

3. Using the `tukey()` function, we can take a closer look at the **effects** of the individual predictors

```
  Estimate Std. Error z value Pr(>|z|)
open - double == 0    0.43271    0.03064  14.125   <1e-10 ***
single - double == 0  0.12182    0.01797   6.777   <1e-10 ***
single - open == 0   -0.31089    0.02832 -10.979   <1e-10 ***
```

Step 6: Interpretation – Effects

3. Using the `tukey()` function, we can take a closer look at the **effects** of the individual predictors

```
 Estimate Std. Error z value Pr(>|z|)
e - a == 0 -0.150590   0.020311  -7.414  < 1e-05 ***
i - a == 0 -0.248762   0.020311 -12.248  < 1e-05 ***
o - a == 0 -0.132478   0.020311  -6.523  < 1e-05 ***
u - a == 0 -0.245664   0.020311 -12.095  < 1e-05 ***
i - e == 0 -0.098171   0.020190  -4.862 1.22e-05 ***
o - e == 0  0.018113   0.020190   0.897    0.898
u - e == 0 -0.095074   0.020190  -4.709 2.10e-05 ***
o - i == 0  0.116284   0.020190   5.759  < 1e-05 ***
u - i == 0  0.003098   0.020190   0.153    1.000
u - o == 0 -0.113186   0.020190  -5.606  < 1e-05 ***
```