

# Session 10: Gemischte Modelle 1

Dominic Schmitz & Janina Esser

Verein für Diversität in der Linguistik

# Gedankenexperiment



- Stell dir vor, dass du Elternteil von 6 Kindern bist
- Jedes Jahr misst du die Körpergröße deiner Kinder an ihren Geburtstagen

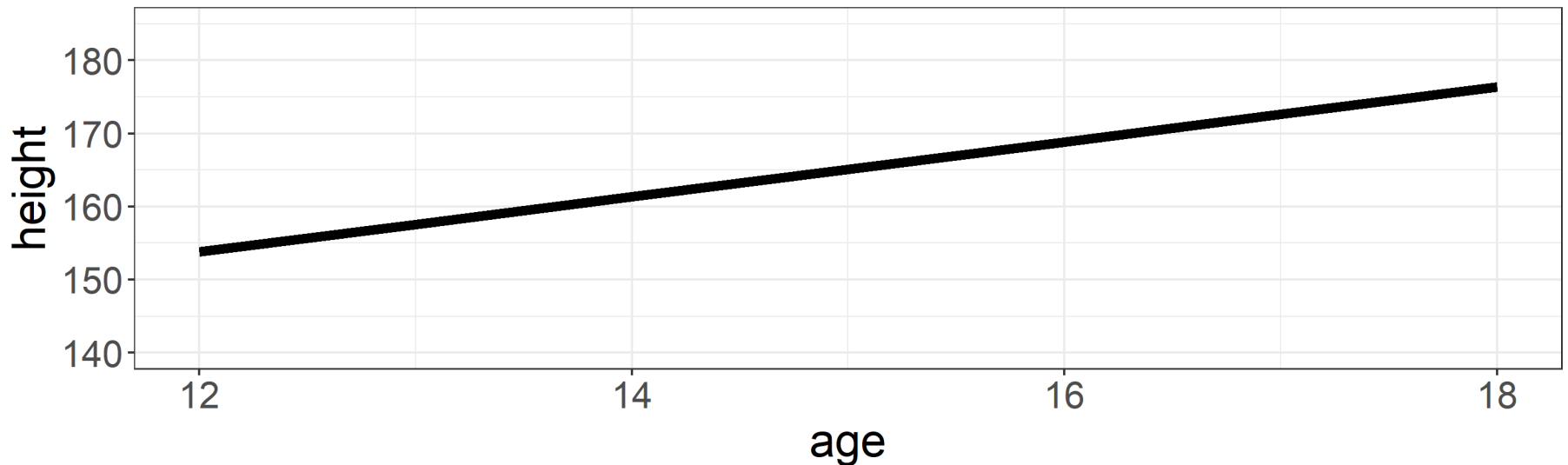
	Kate	Eve	Tess	Max	Neil	Jack
12	149.8	156.3	145.8	149.1	143.3	159.3
13	156.7	163.2	153.7	156.2	150.4	166.4
14	158.7	165.2	160.7	163.8	158.0	174
15	159.7	166.2	162.7	170.1	164.3	180.3
16	162.5	169.0	167.5	173.4	167.6	183.6
17	162.5	169.0	172.5	175.2	169.4	185.4
18	163.0	169.5	178.0	175.7	169.9	185.9

# Gedankenexperiment



- Mit deinem Wissen über Simple Lineare Regression erstellst du ein Model:

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



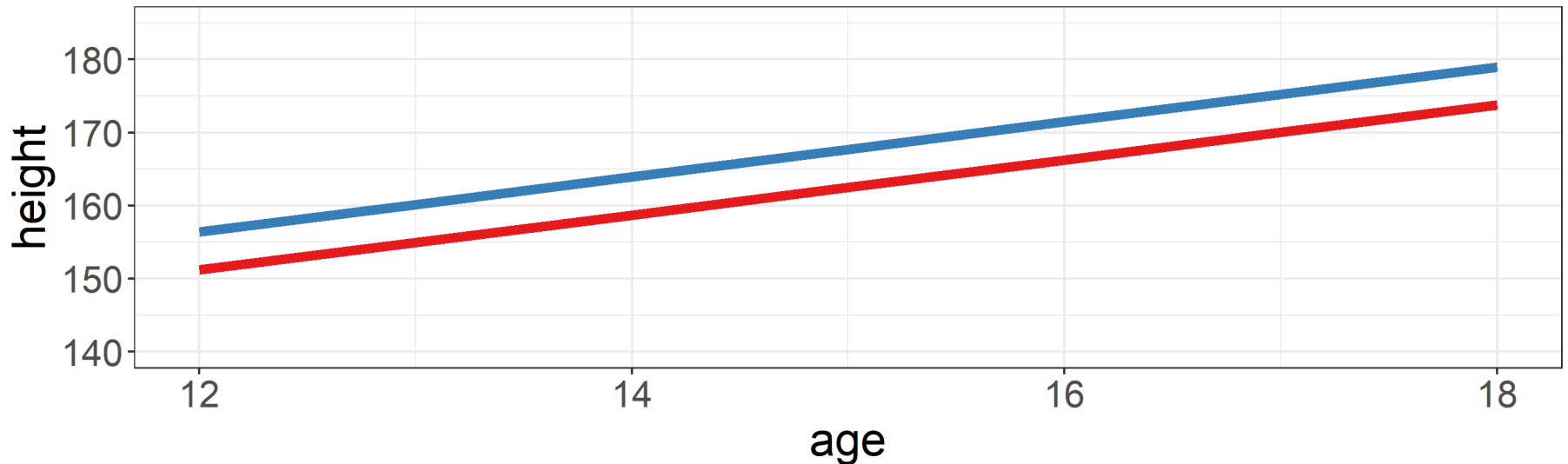
- Laut des Modells wachsen alle Kinder mit gleicher Geschwindigkeit (Slope)

# Gedankenexperiment



- Mit deinem Wissen über Multiple Lineare Regression erstellst du ein Model:

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



- Laut des Modells wachsen alle Kinder mit gleicher Geschwindigkeit (Slope), aber Mädchen sind konsistent kleiner als Jungs (Intercepts)

# Gedankenexperiment



- Aber stimmt das?

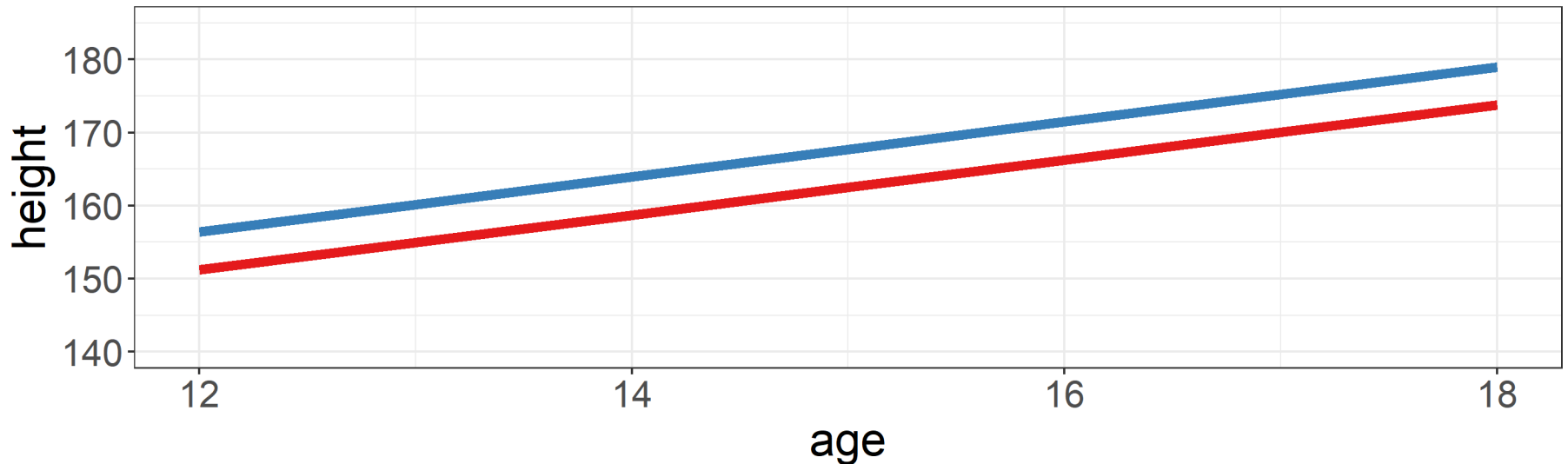
	Kate	Eve	Tess	Max	Neil	Jack
12	149.8	156.3	145.8	149.1	143.3	159.3
13	156.7	163.2	153.7	156.2	150.4	166.4
14	158.7	165.2	160.7	163.8	158.0	174
15	159.7	166.2	162.7	170.1	164.3	180.3
16	162.5	169.0	167.5	173.4	167.6	183.6
17	162.5	169.0	172.5	175.2	169.4	185.4
18	163.0	169.5	178.0	175.7	169.9	185.9

# Gedankenexperiment



- Mit deinem Wissen über Multiple Lineare Regression erstellst du ein Model:

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



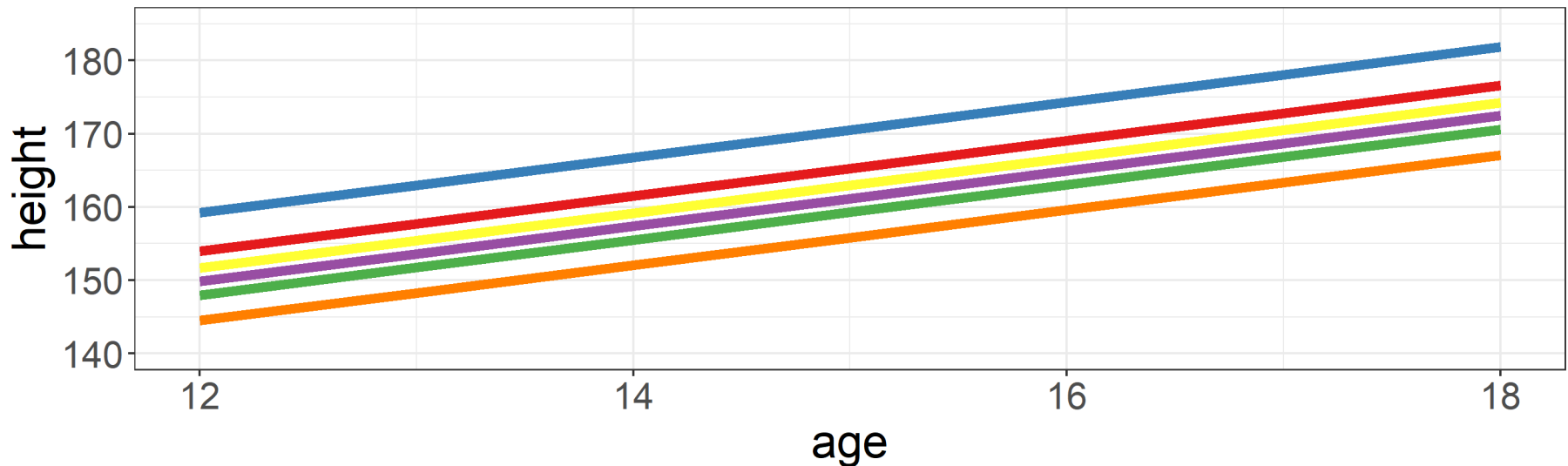
- **Frage:** Was müssen wir tun, damit das Modell realistischer wird?

# Gedankenexperiment



- Mit deinem baldigen Wissen über Gemischte Modelle erstellst du ein Model:

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



- Laut des Modells startet jedes Kind mit einer eigenen Größe (Intercepts), während alle Kinder mit gleicher Geschwindigkeit (Slope) wachsen

# Gedankenexperiment



- Aber stimmt das?

	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

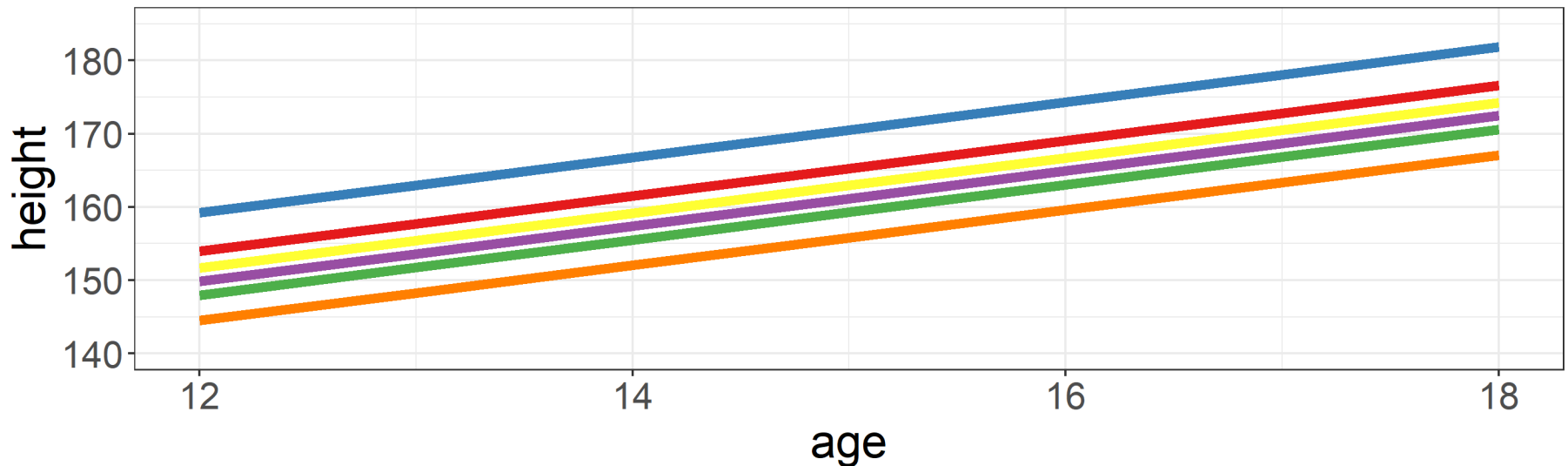


# Gedankenexperiment



- Mit deinem baldigen Wissen über Gemischte Modelle erstellst du ein Model:

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



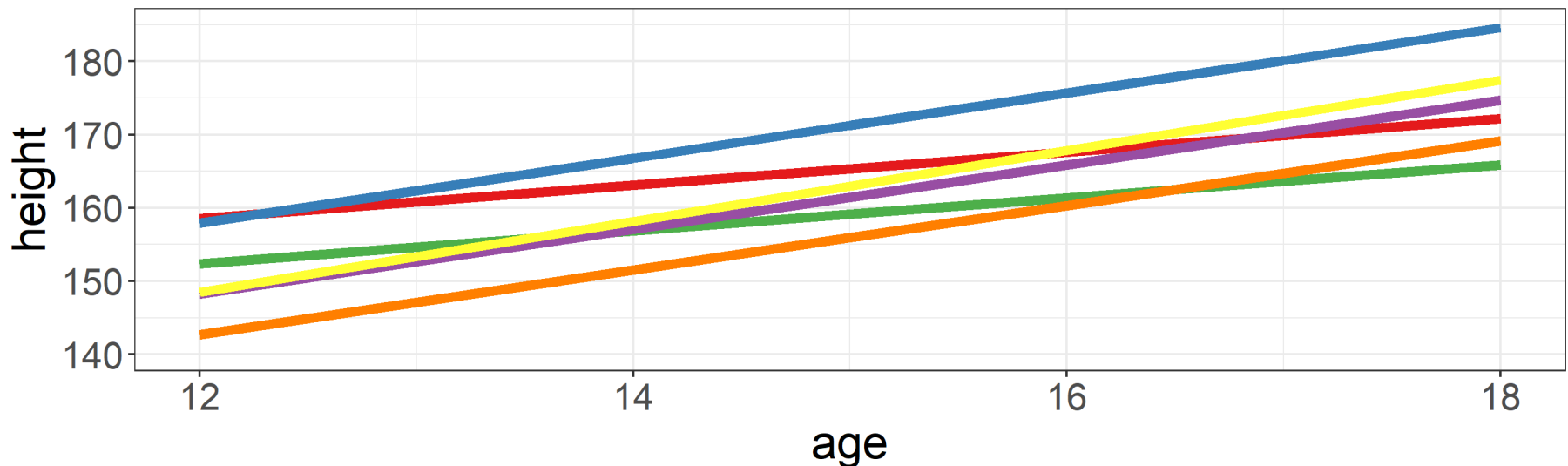
- **Frage:** Was müssen wir tun, damit das Modell realistischer wird?

# Gedankenexperiment



- Mit deinem baldigen Wissen über Gemischte Modelle erstellst du ein Model:

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

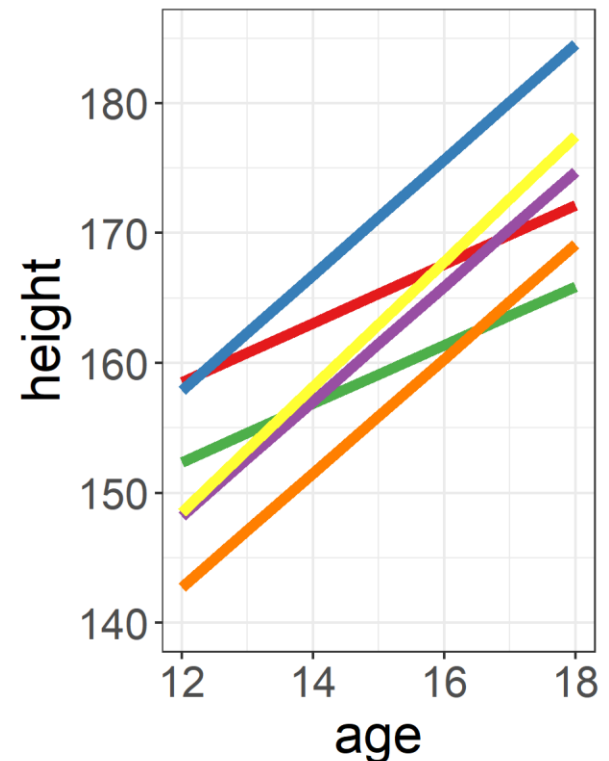
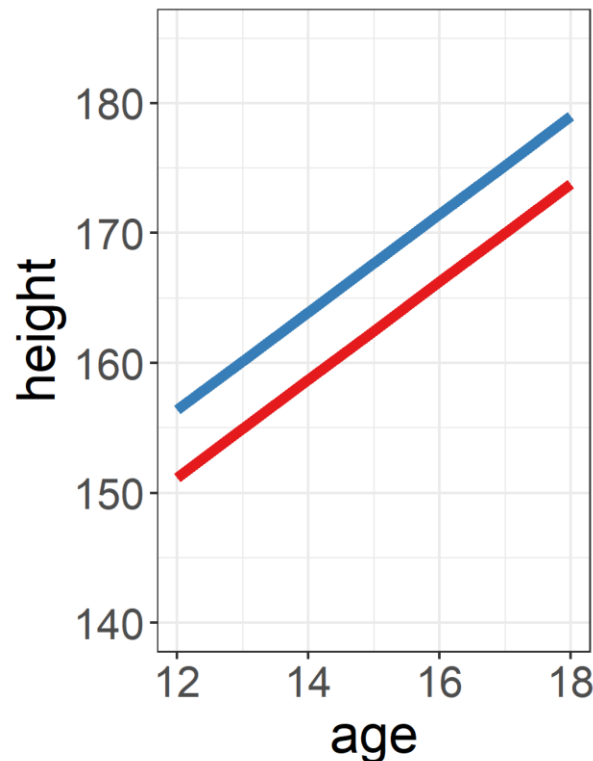
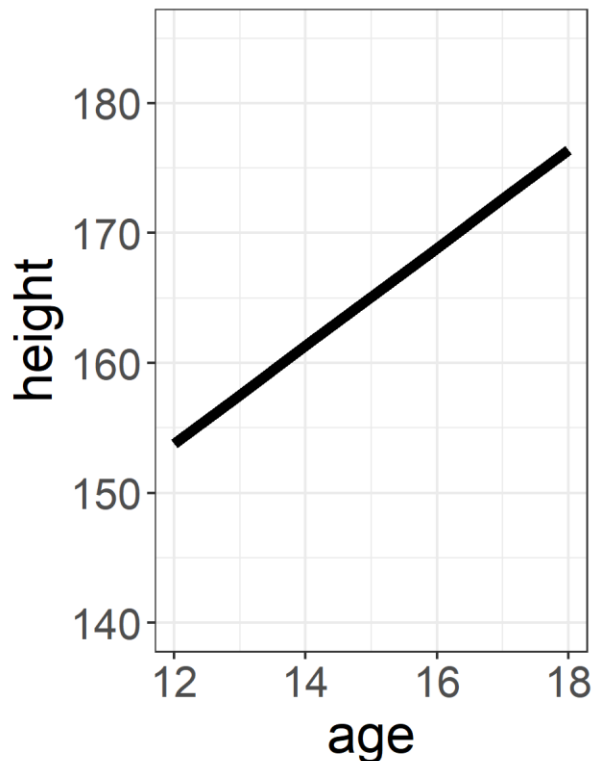


- Laut des Modells startet jedes Kind mit einer eigenen Größe (Intercepts) und wächst mit einer individuellen Geschwindigkeit (Slopes)

# Gedankenexperiment



- Offenbar sind Gemischte Modelle besser darin als Simple oder Multiple Lineare Modelle, die Realität (der Daten) zu erfassen



# Simple Lineare Regression



kontinuierliche  
abhängige Variable

unabhängige  
Predictor Variable

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

Intercept

Slope

Error Term /  
Residuen

# Multiple Lineare Regression: Formel



kontinuierliche  
abhängige Variable

unabhängige  
Predictor Variable 2

unabhängige  
Predictor Variable 1

unabhängige  
Predictor Variable i

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

Intercept

Slope von  
Variable 1

Slope von  
Variable 2

Slope von  
Variable i

Error Term /  
Residuen

# Random Intercept Formula



kontinuierliche  
abhängige Variable

unabhängige  
Predictor Variable 1

unabhängige  
Predictor Variable i

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

Intercept

**intercept  
adjustment**

Slope von  
Variable 1

Slope von  
Variable i

Error Term /  
Residuen

# Random Slope Formula



kontinuierliche  
abhängige Variable

unabhängige  
Predictor Variable 1

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

Intercept

intercept  
adjustment

slope  
adjustment

Slope von  
Variable 1

Error Term /  
Residuen

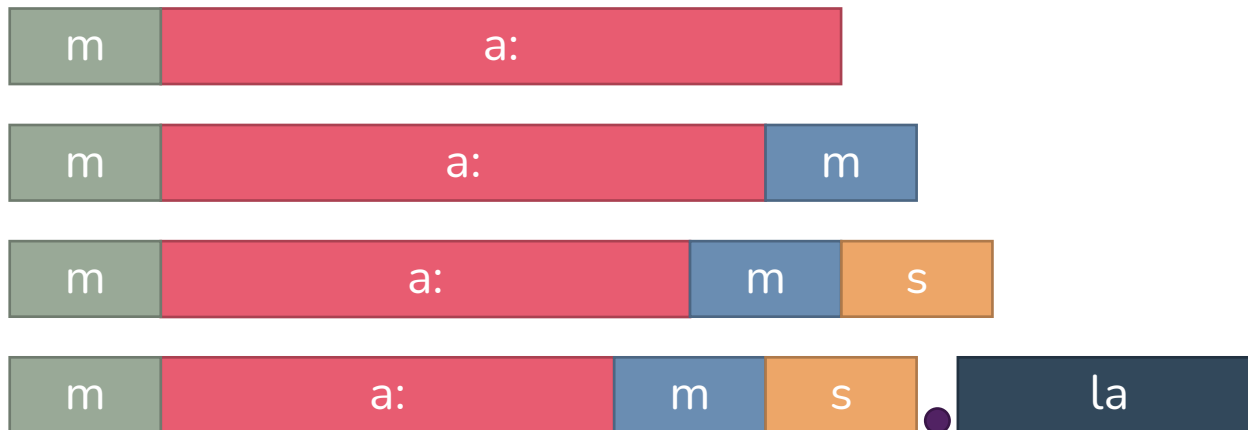
# Beispieldaten



- Für die folgenden Beispiele werden wir Daten folgender Studie nutzen:

## Compensatory Vowel Shortening in German<sup>1</sup>

- Stressed Vowels sind kürzer je nachdem wie viele Konsonanten ihnen folgen:



<sup>1</sup> Schmitz, D., Cho, H.-E., & Niemann, H. (2018). Vowel shortening in German as a function of syllable structure. Proceedings 13. Phonetik Und Phonologie Tagung (P&P13), 181–184.



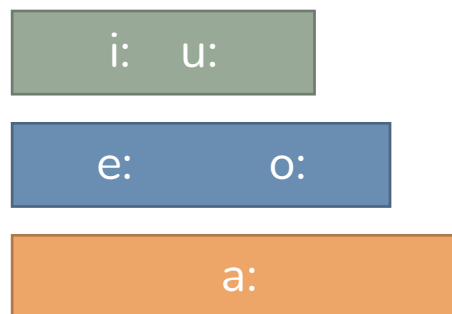
# Beispieldaten



- Für die folgenden Beispiele werden wir Daten folgender Studie nutzen:

## Compensatory Vowel Shortening in German<sup>1</sup>

- Unabhängig von diesem Vowel Shortening gilt, dass offene Vokale länger sind als halb-offene Vokale, und halb-offene Vokale sind länger als geschlossene Vokale:



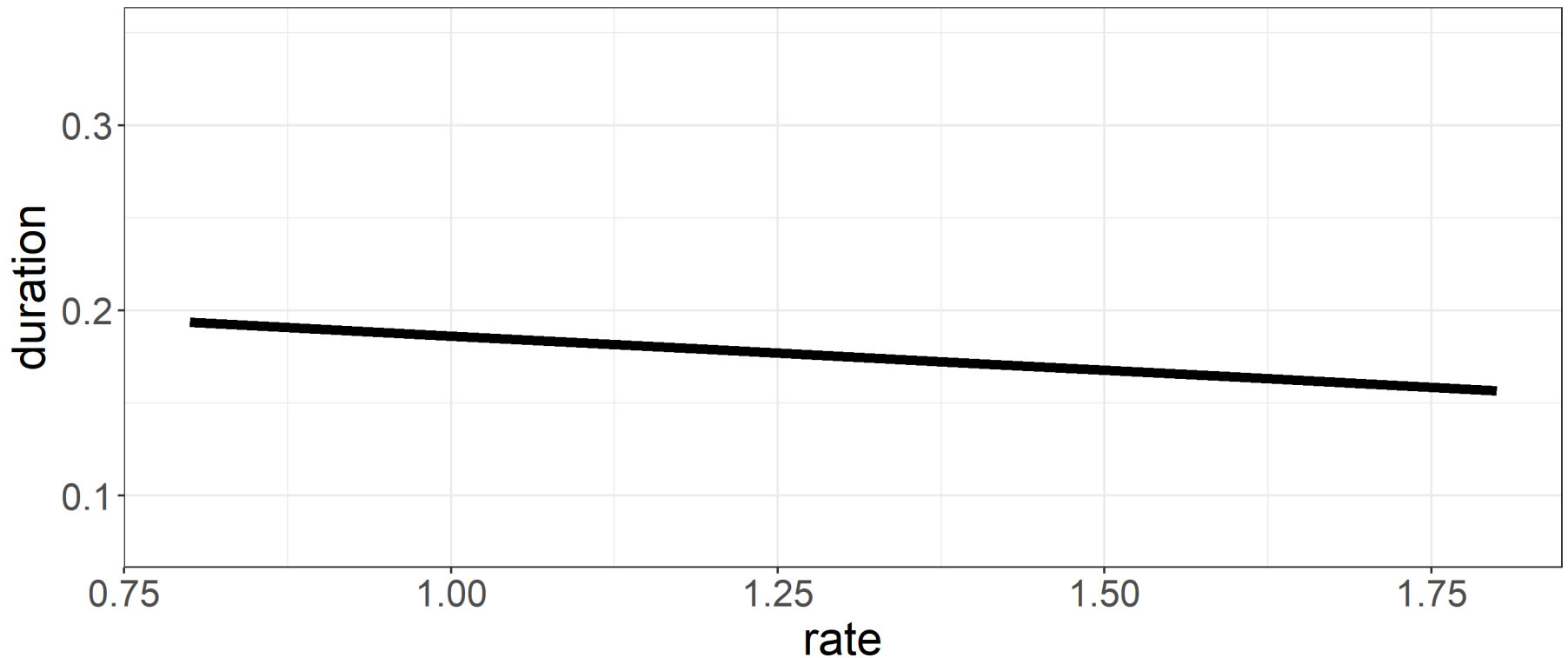
<sup>1</sup> Schmitz, D., Cho, H.-E., & Niemann, H. (2018). Vowel shortening in German as a function of syllable structure. Proceedings 13. Phonetik Und Phonologie Tagung (P&P13), 181–184.

# Beispieldaten



- Bisher haben wir Vokaldauer mit Simpler Linearer Regression...

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

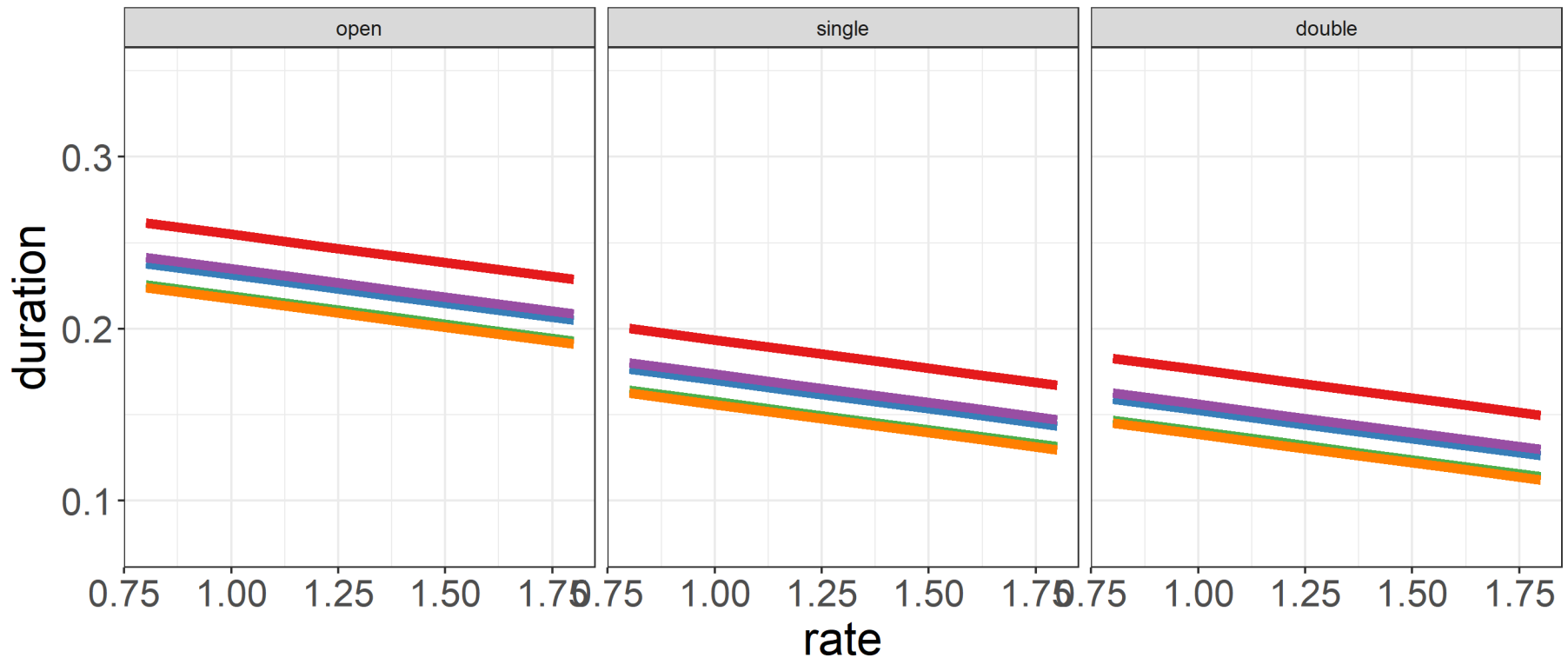


# Beispieldaten



- ... und Multipler Linearer Regression gemodelt

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

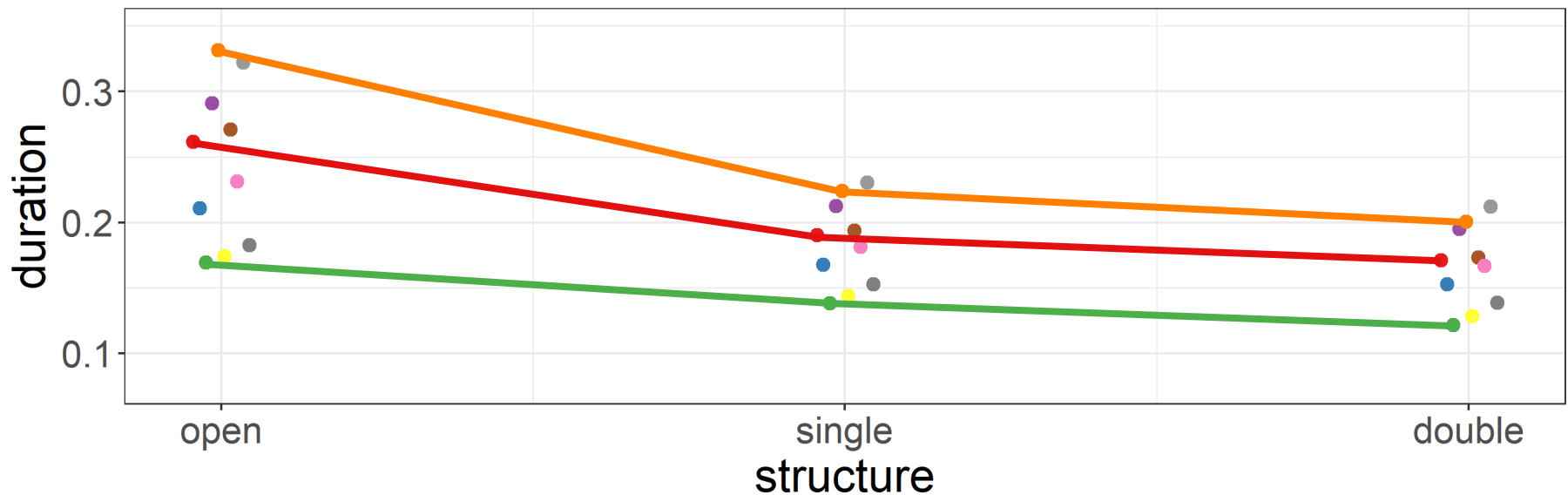


# Beispieldaten



- Mittlerweile wissen wir aber, dass Gemischte Modelle besser geeignet sind, zum Beispiel:

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



# Fixed & Random Effects



- In Gemischten Modellen arbeiten wir mit zwei Arten von Prediktoren:

## 1. Fixed Effects

- erklärende Variablen
- Variablen, die im Mittelpunkt stehen
- wiederholbar

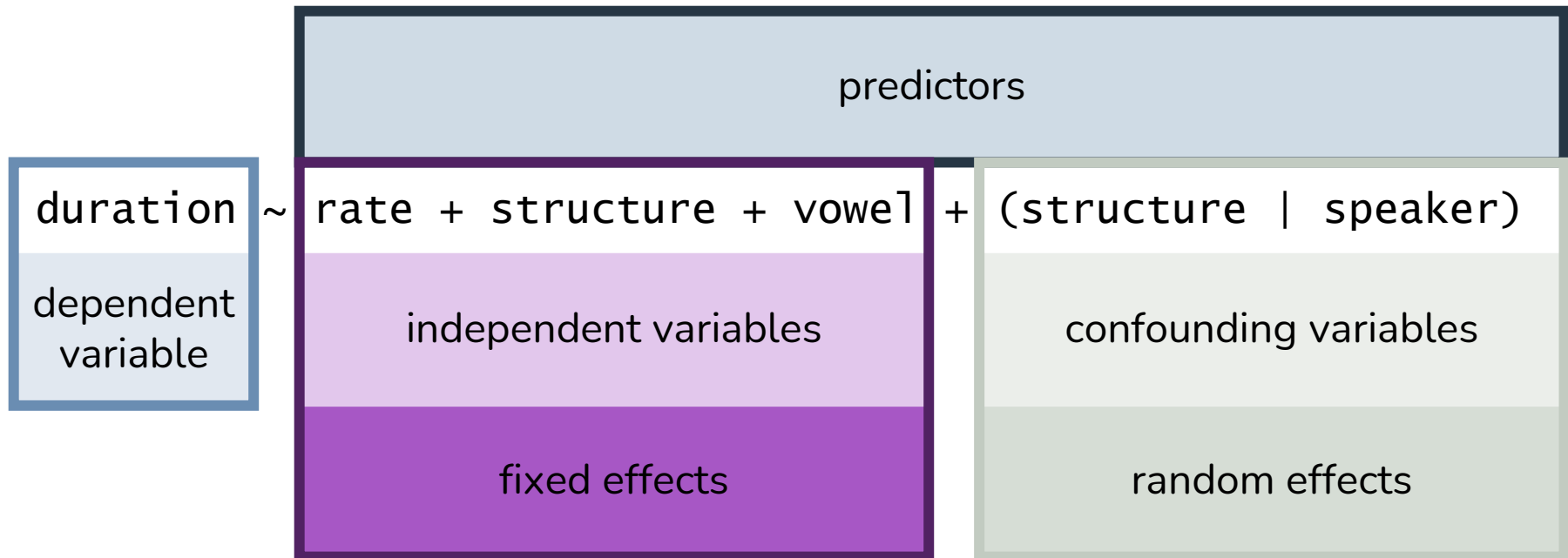
## 2. Random Effects

- Ursprung chaotischer Variation in den Daten
- zufällig, nicht wiederholbar

# Fixed & Random Effects



- Das vorherige Modell zu Vokaldauern:



# Fixed & Random Effects



- Wenn wir also Daten anhand Gemischter Modelle analysieren möchten, müssen wir nicht nur entscheiden, welche Variablen wir sinnvoller Weise nutzen sollten...
- sondern auch, welche Variablen sich als Fixed Effects eignen und welche Variablen eher Random Effects entsprechen

# Fixed & Random Effects



- Typische Beispiele für Fixed Effects sind Variablen, für welche es Vorhersagen durch vorherige Studien und Literatur gibt, z.B.
  - Frequenz                      frequenter = kürzere Dauer, kürzere RT
  - Neighbourhoods              denser = kürzere Dauer
  - gemessene Dauern              lange Base = lange Affix
  - Wortlänge                      mehr Buchstaben = höhere RT
  - Videospieelfrequenz              frequenter = kürzere RT
  - ...



# Fixed & Random Effects



- Typische Beispiele für Random Effects sind Variablen, für welche es keine klaren Vorhersagen gibt, z.B.
  - subject                      alle TN sind unterschiedlich
  - items                         alle Wörter sind unterschiedlich
  - item order                  Priming? wer weiß
  - ...

# Fixed & Random Effects



- Zurück zum Beispiel der Vokaldauern; hier haben wir folgende Variablen:
  - speech rate                      höher = kürzere Dauer
  - structure                        komplexer = kürzere Dauer
  - vowel                              offener = längere Dauer
  - speaker                          ???
  - word                                ???

# Fixed & Random Effects



- Zurück zum Beispiel der Vokaldauern; hier haben wir folgende Variablen:

• speech rate	höher = kürzere Dauer	fixed effects
• structure	komplexer = kürzere Dauer	
• vowel	offener = längere Dauer	
• speaker	???	random effects
• word	???	

# Gemischte Modelle in R



- Wie wir bereits festgestellt haben, bedeuten mehr Variablen auch mehr Arbeitsschritte
- Typische Schritte sind
  1. Distribution der abhängigen Variable überprüfen
  2. Check der Korrelationen wegen Kollinearität
  3. „volles“ Modell erstellen
  4. „bestes“ Modell finden
  5. Assumptions überprüfen
  6. Modell interpretieren

# Gemischte Modelle in R

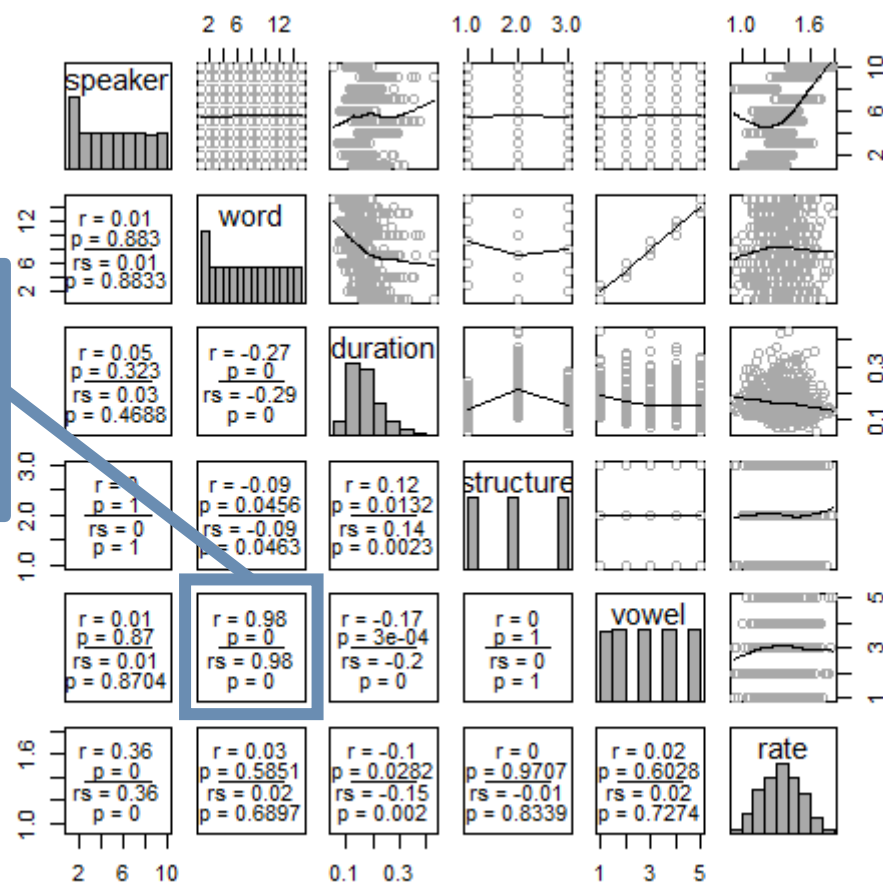


- Wie wir bereits festgestellt haben, bedeuten mehr Variablen auch mehr Arbeitsschritte
- Typische Schritte sind
  1. Distribution der abhängigen Variable überprüfen ✓
  2. Check der Korrelationen wegen Kollinearität
  3. „volles“ Modell erstellen
  4. „bestes“ Modell finden
  5. Assumptions überprüfen
  6. Modell interpretieren

# Check der Korrelationen



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



# Gemischte Modelle in R



- Wie wir bereits festgestellt haben, bedeuten mehr Variablen auch mehr Arbeitsschritte
- Typische Schritte sind
  1. Distribution der abhängigen Variable überprüfen ✓
  2. Check der Korrelationen wegen Kollinearität ✓
  3. „volles“ Modell erstellen
  4. „bestes“ Modell finden
  5. Assumptions überprüfen
  6. Modell interpretieren

# „Volles“ Modell erstellen



- Erstellen eines vollen Modells:

```
library(lme4)
```

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



# Gemischte Modelle in R



- Wie wir bereits festgestellt haben, bedeuten mehr Variablen auch mehr Arbeitsschritte
- Typische Schritte sind
  1. Distribution der abhängigen Variable überprüfen ✓
  2. Check der Korrelationen wegen Kollinearität ✓
  3. „volles“ Modell erstellen ✓
  4. „bestes“ Modell finden
  5. Assumptions überprüfen
  6. Modell interpretieren

# „Bestes“ Modell finden



- Finden des „besten“ Modells

```
step(md1)
```

```
...
```

```
...
```

```
...
```

Model found:

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

# Gemischte Modelle in R



- Wie wir bereits festgestellt haben, bedeuten mehr Variablen auch mehr Arbeitsschritte
- Typische Schritte sind
  1. Distribution der abhängigen Variable überprüfen ✓
  2. Check der Korrelationen wegen Kollinearität ✓
  3. „volles“ Modell erstellen ✓
  4. „bestes“ Modell finden ✓
  5. Assumptions überprüfen
  6. Modell interpretieren

# Assumptions überprüfen



- Multiple Linear Regression folgt den gleichen Assumptions, denen auch Simple und Multiple Linear Regression folgen
  - ▶ Linearity
  - ▶ Homoscedasticity
  - ▶ Normality
  - ▶ Independence
- **Hinweis:** Die SfL Datensätze sind i.d.R. zu klein um Gemischte Modelle zu erstellen, die allen Assumptions entsprechen.

# Gemischte Modelle in R



- Wie wir bereits festgestellt haben, bedeuten mehr Variablen auch mehr Arbeitsschritte
- Typische Schritte sind
  1. Distribution der abhängigen Variable überprüfen ✓
  2. Check der Korrelationen wegen Kollinearität ✓
  3. „volles“ Modell erstellen ✓
  4. „bestes“ Modell finden ✓
  5. Assumptions überprüfen ✓
  6. Modell interpretieren

# Interpretation des Modells



- Generell sind wir an zwei Dingen interessiert:
  1. die  $p$ -Werte der einzelnen Predictors
  2. die Effekte der einzelnen Predictors

# Interpretation des Modells



1. Mit der `anova()` Funktion erhalten wir  $p$ -Werte

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 ***

# Interpretation des Modells



2. Mit der `summary()` Funktion können wir einen Blick auf die einzelnen Effekte der Predictors werfen

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	***
vowel <sub>e</sub>	-0.15059	0.02031	423.07910	-7.414	6.73e-13	***
vowel <sub>i</sub>	-0.24876	0.02031	423.07910	-12.248	< 2e-16	***
vowel <sub>o</sub>	-0.13248	0.02031	423.07910	-6.523	1.98e-10	***
vowel <sub>u</sub>	-0.24566	0.02031	423.07910	-12.095	< 2e-16	***



# Interpretation des Modells



Der s.g. „Tukey-Contrast“ zeigt uns die Unterschiede innerhalb eines kategorischen 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 ***

# Interpretation des Modells



Der s.g. „Tukey-Contrast“ zeigt uns die Unterschiede innerhalb eines kategorischen 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	***