12
Gemischte Lineare Regression

Dominic Schmitz & Janina Esser

Danger

Danger

Gedankenexperiment

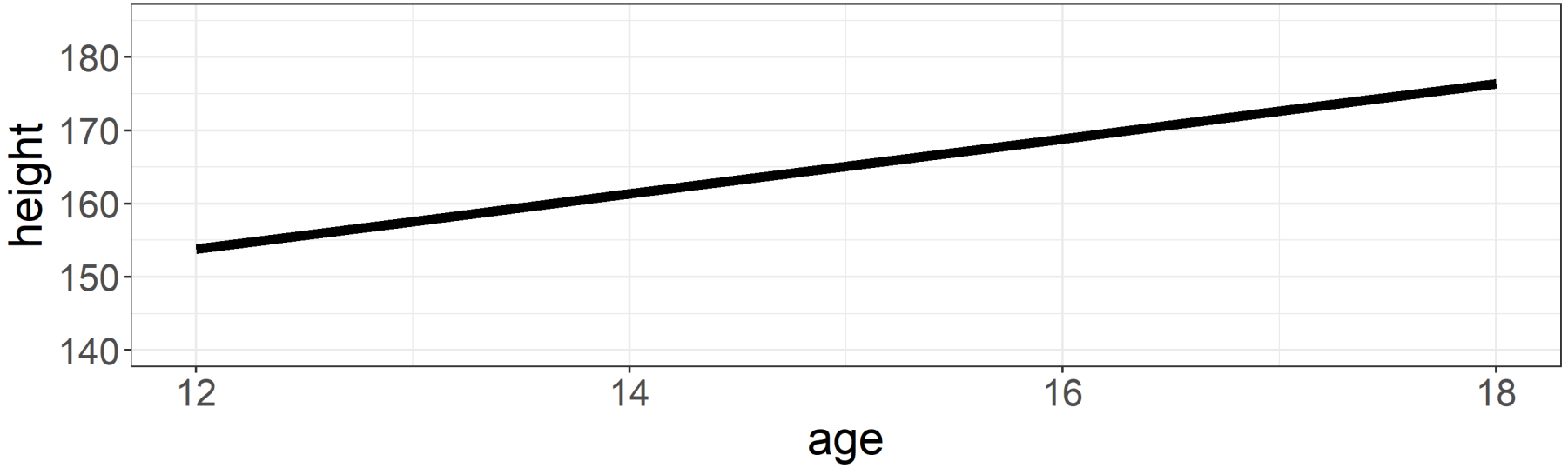
- Stell dir vor, dass du Elternteil von 6 Kindern bist
- Jedes Jahr misst du die Körpergröße deiner Kinder an ihren Geburtstagen

	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)
```

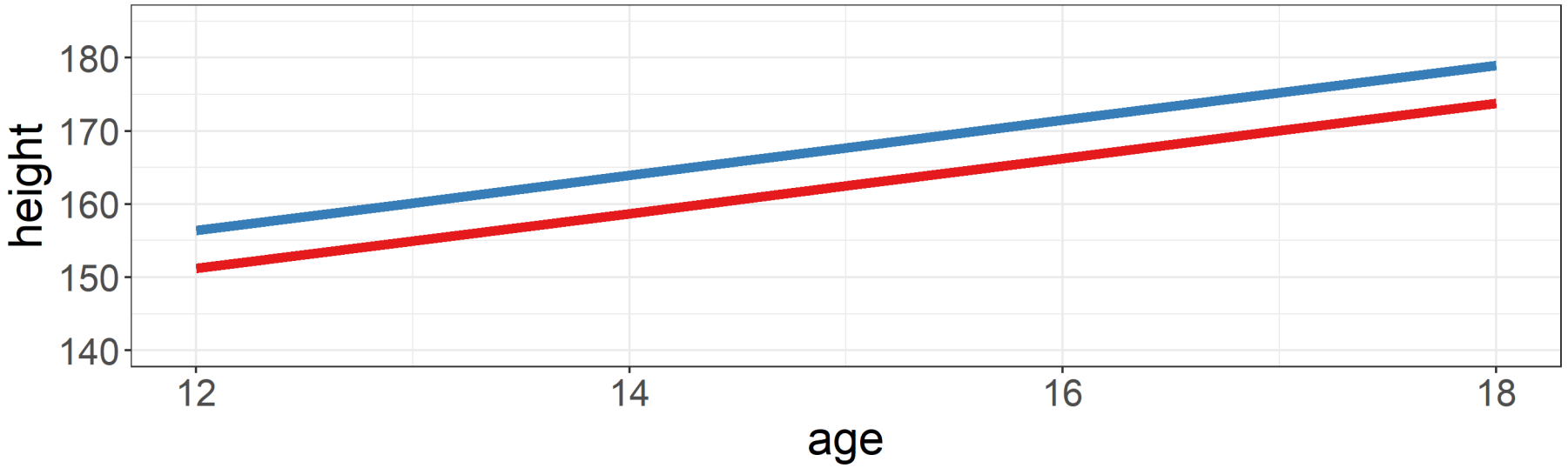


- Laut des Modells wachsen alle Kinder mit gleicher Geschwindigkeit (Steigung)

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 (Steigung), aber Mädchen sind konsistent kleiner als Jungs (Achsenabschnitt)

Gedankenexperiment

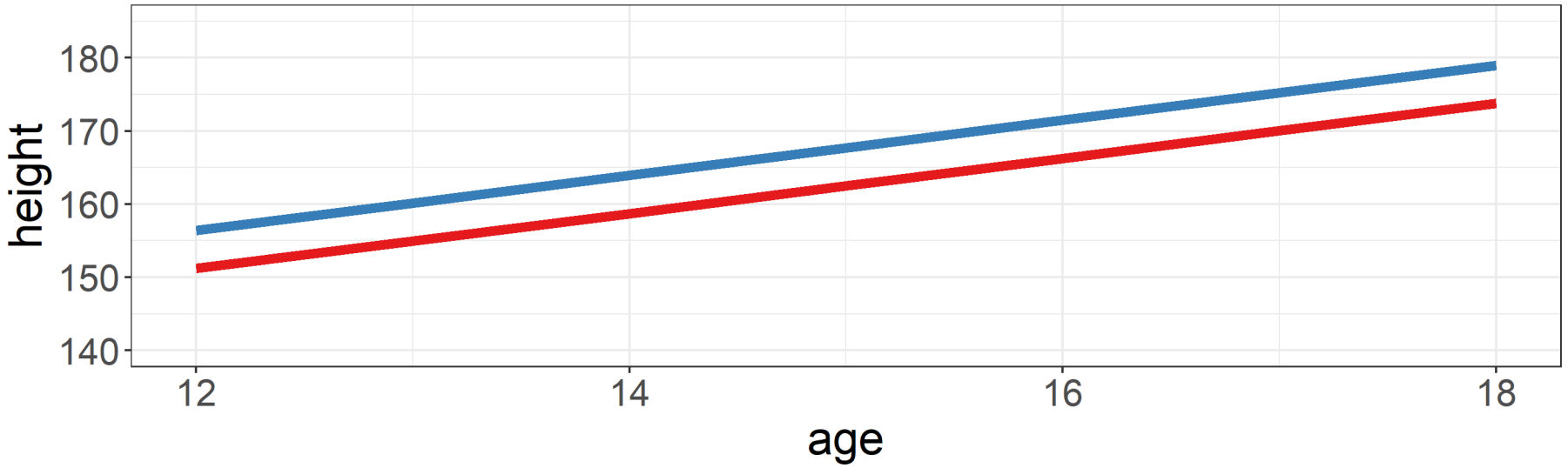
- Aber stimmt das?

	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)
```

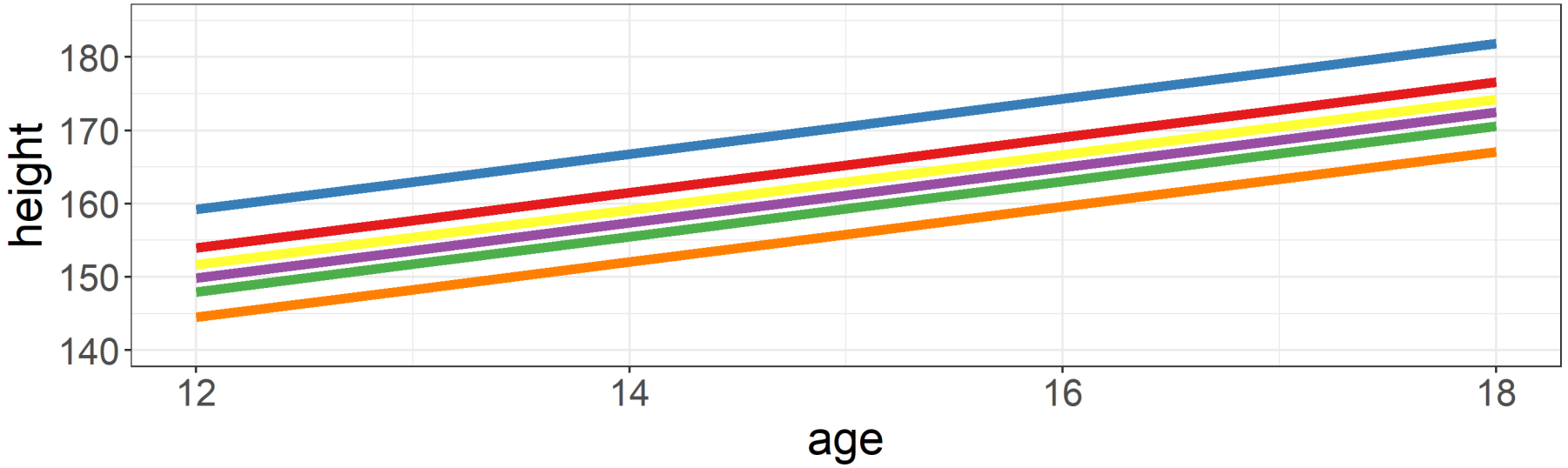


- **Frage**: Was müssen wir tun, damit das Modell realistischer wird?

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 (Achsenabschnitte), während alle Kinder mit gleicher Geschwindigkeit (Steigung) wachsen

Gedankenexperiment

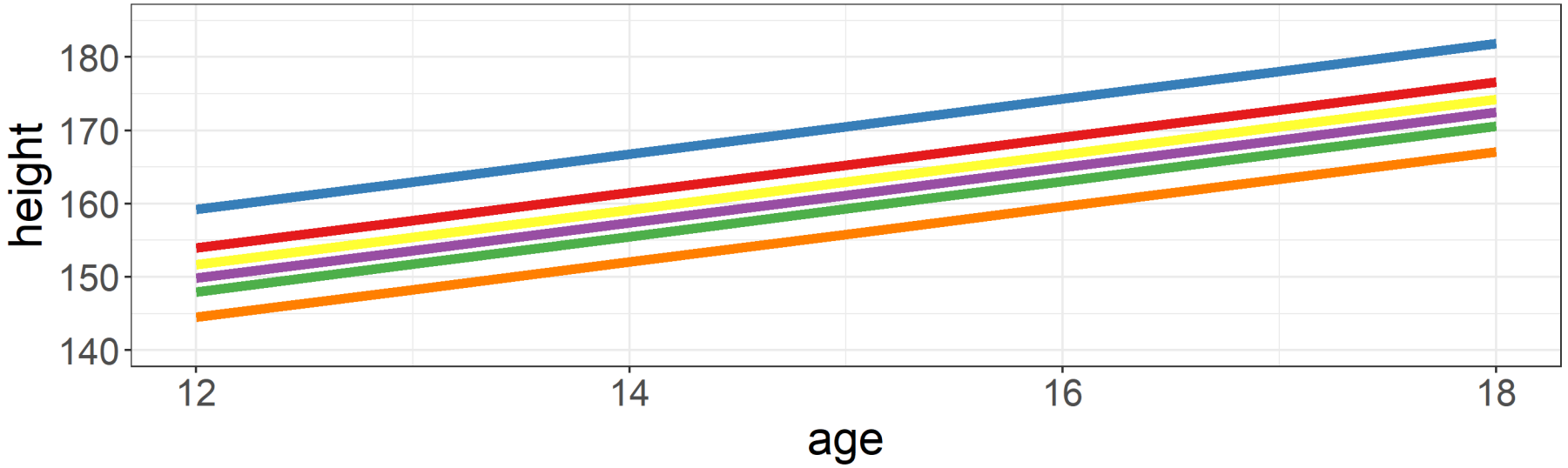
- Aber stimmt das?

	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)
```

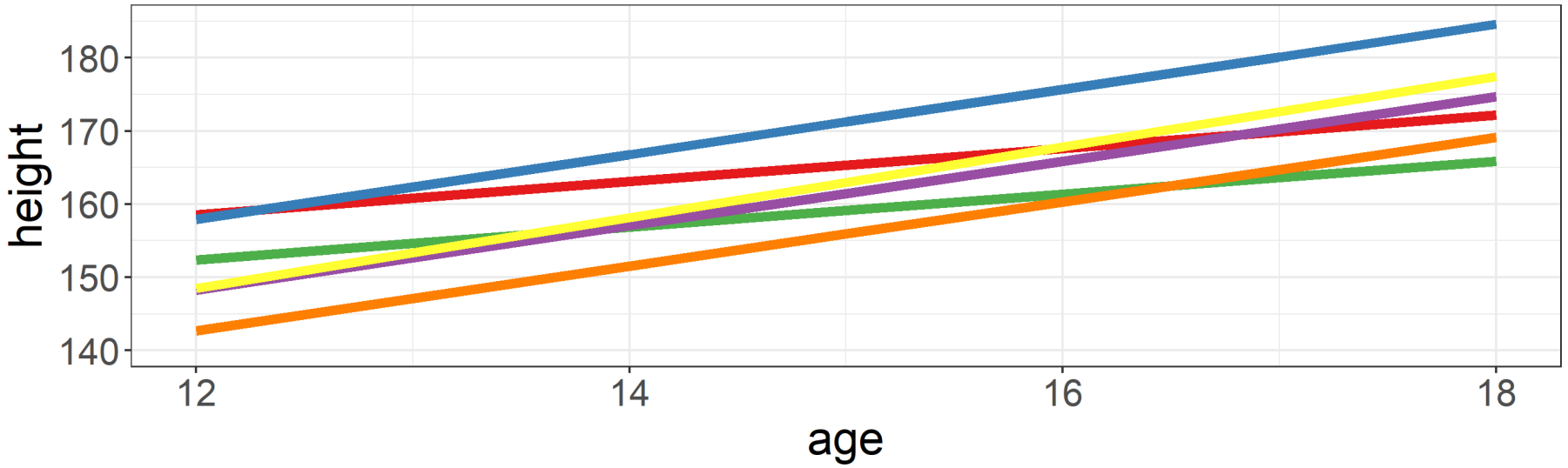


- **Frage**: Was müssen wir tun, damit das Modell realistischer wird?

Gedankenexperiment

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

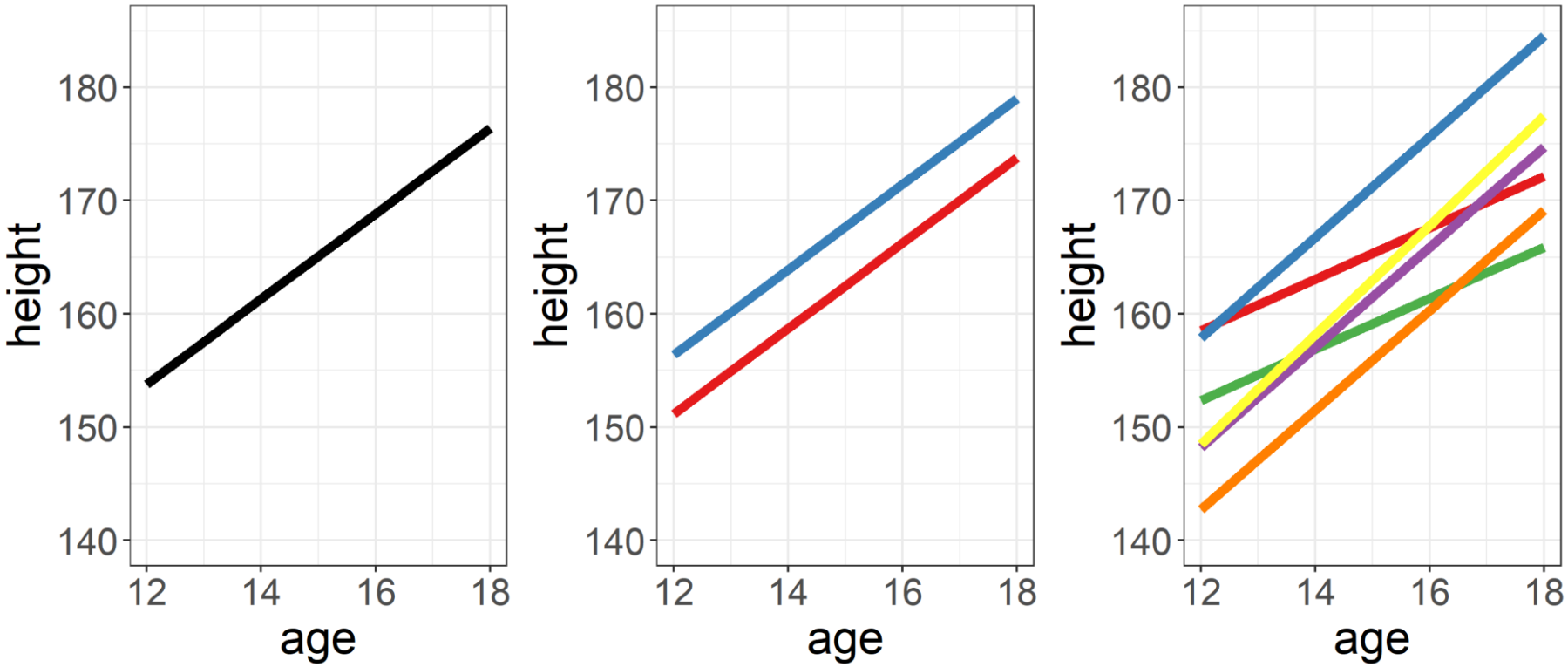
```
lm(height ~ age + bsex + (age | name), data_h)
```



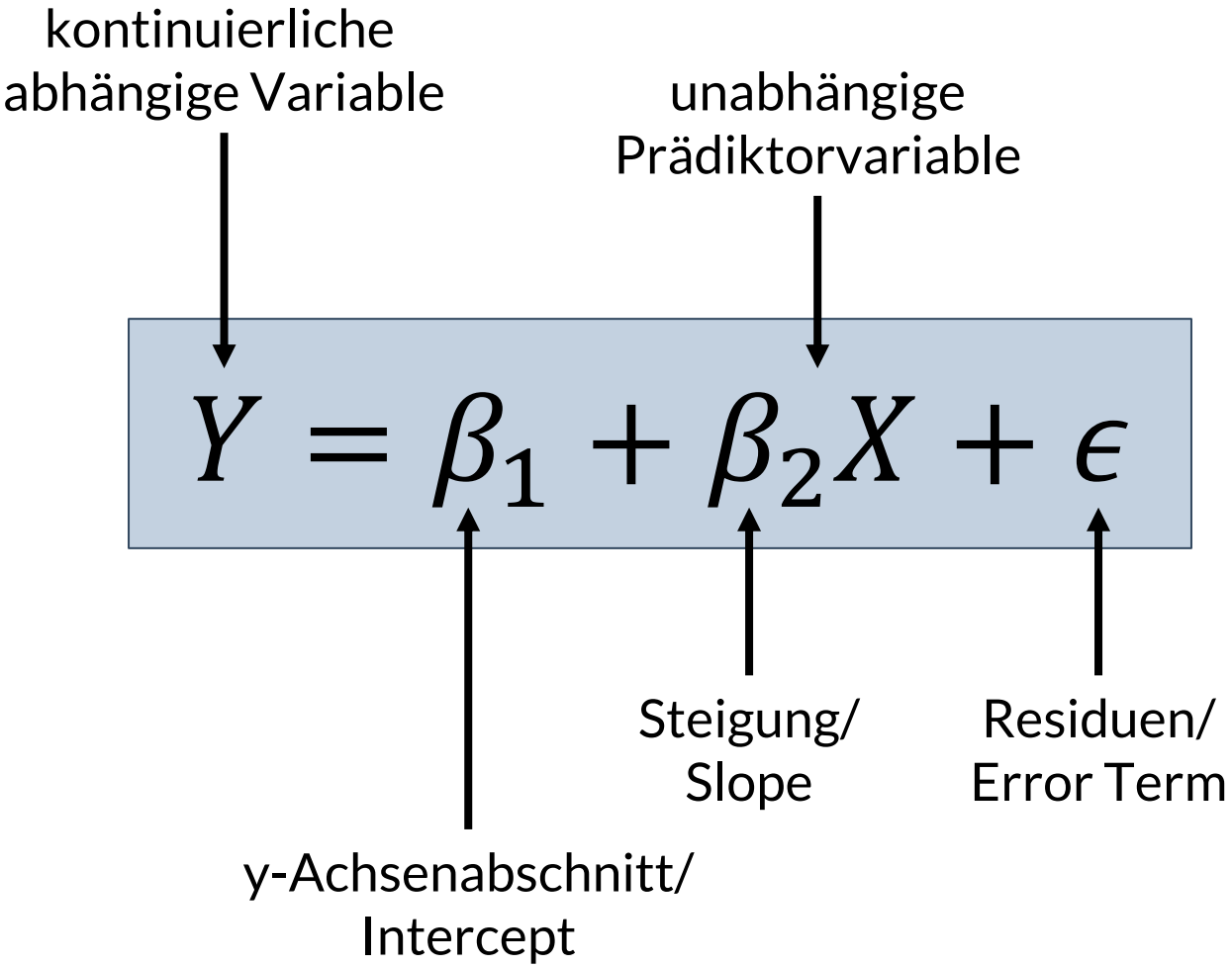
- Laut des Modells startet jedes Kind mit einer eigenen Größe (Achsenabschnitte) und wächst mit einer individuellen Geschwindigkeit (Steigungen)

Gedankenexperiment

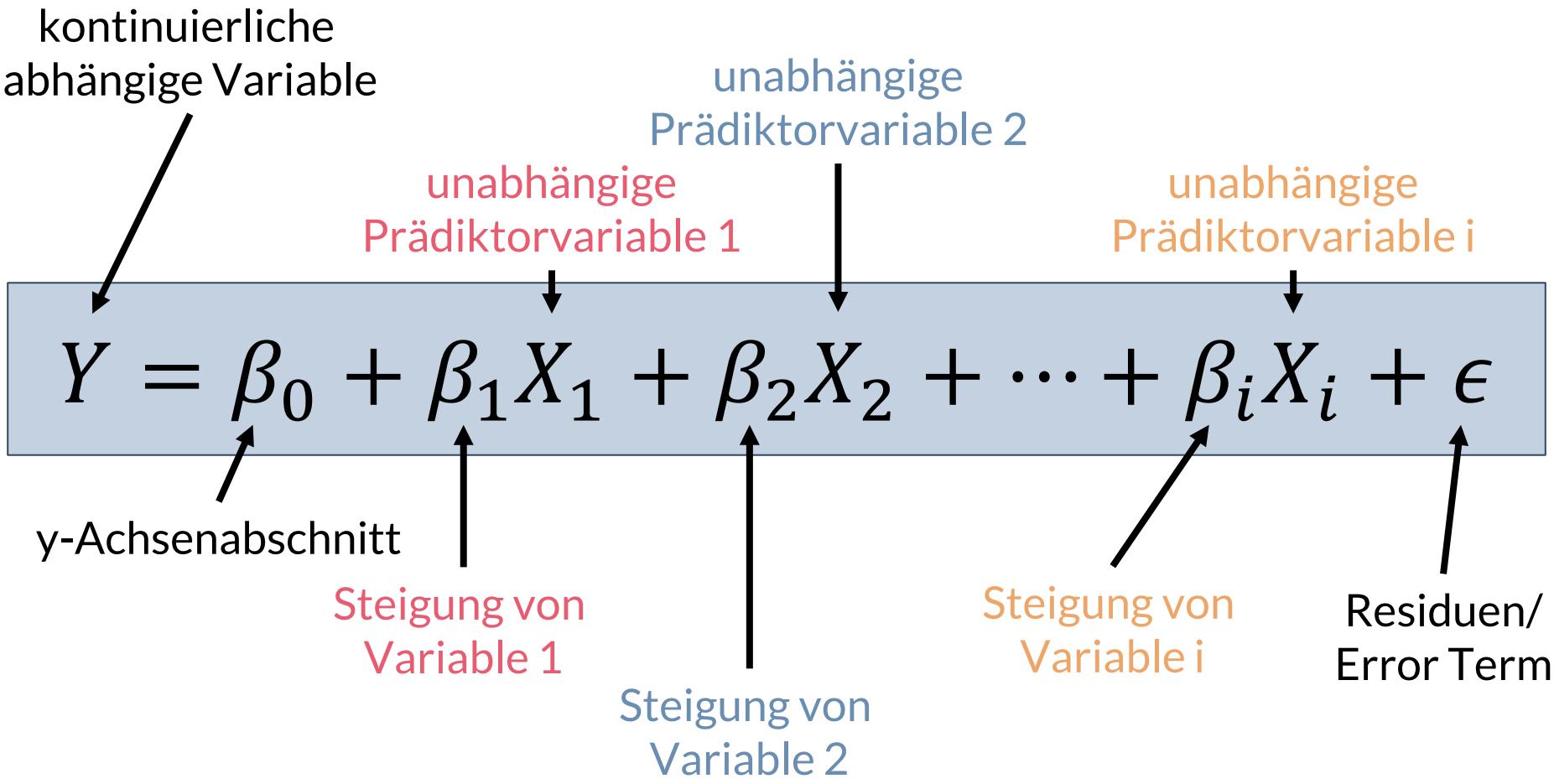
- Offenbar sind Gemischte Modelle besser darin als Einfache oder Multiple Lineare Modelle, die Realität (der Daten) zu erfassen



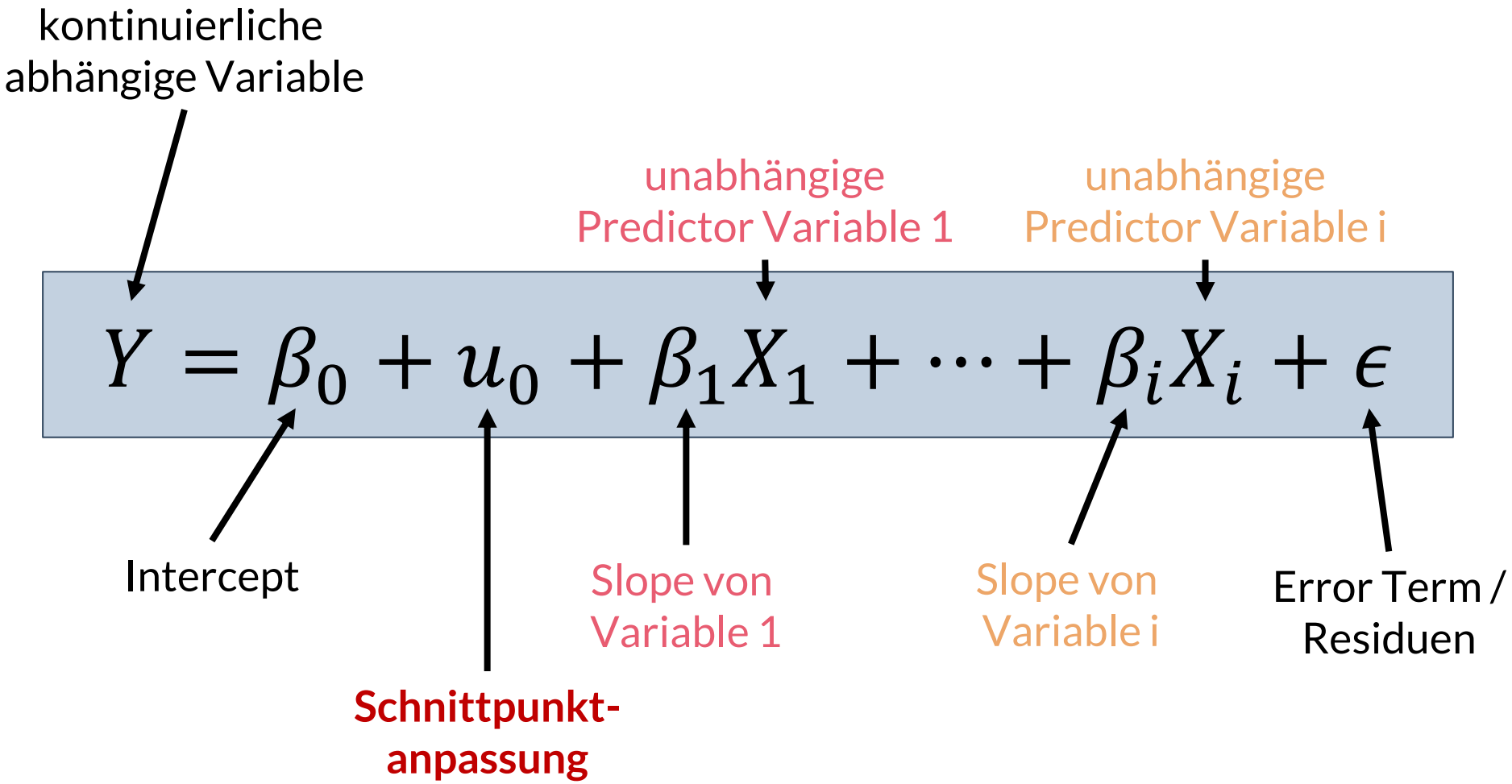
Einfache Lineare Regression



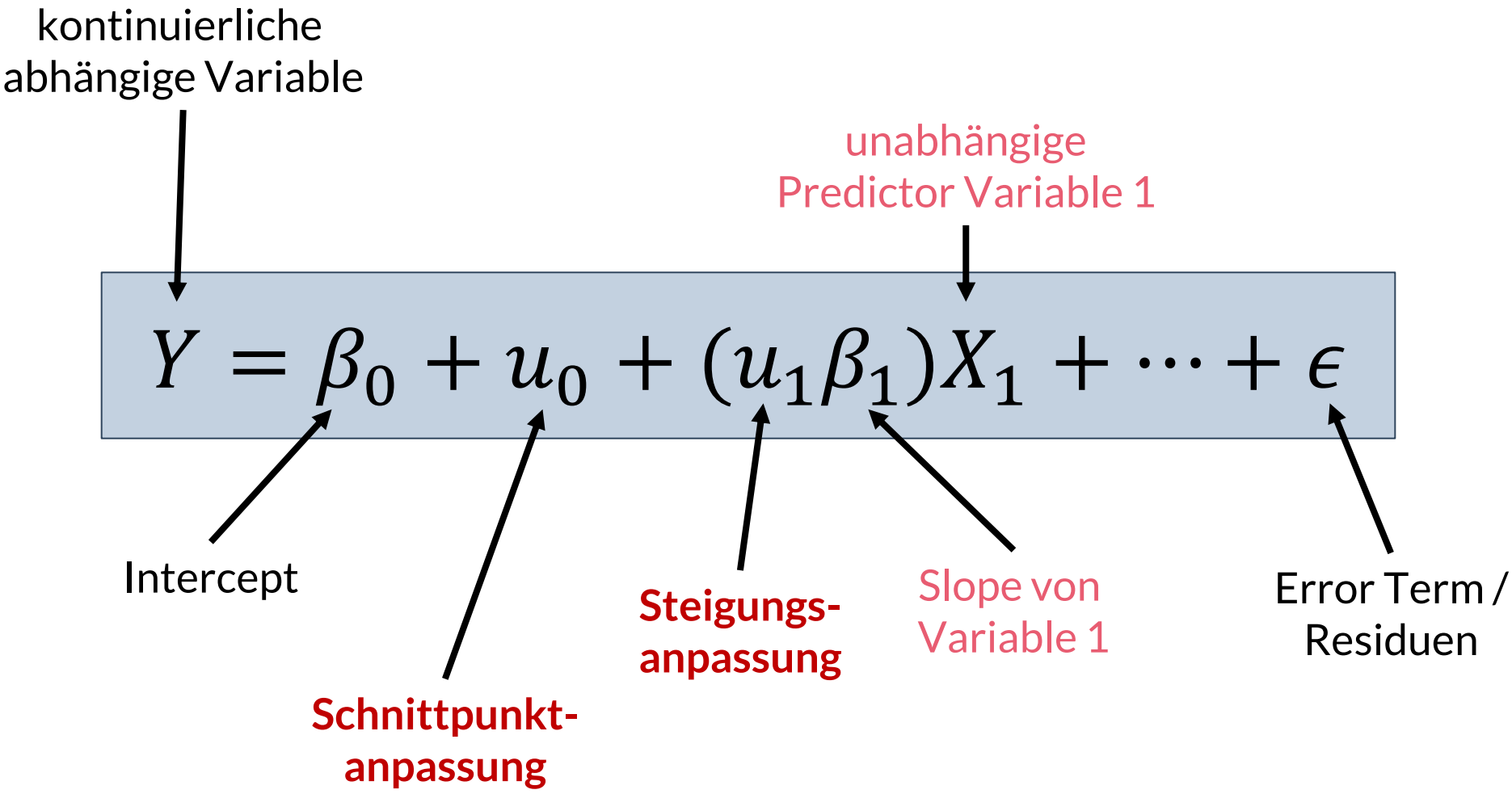
Multiple Lineare Regression



Random Intercept Formula



Random Slope Formula

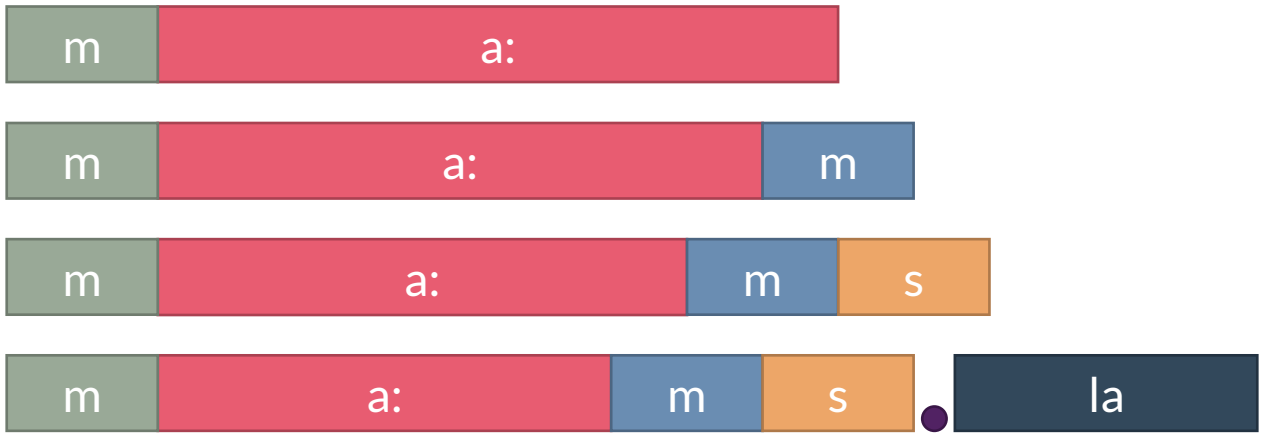


Beispieldaten

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

Compensatory Vowel Shortening in German[1]

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



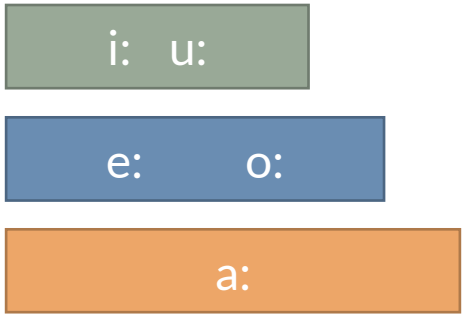
[1] Schmitz, D., Cho, H.-E., & Niemann, H. (2018). Vowel shortening in German as a function of syllable structure. Proceedings 13. Phonetik Und Phonologie Tagung (P&P13), 181–184.

Beispieldaten

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

Compensatory Vowel Shortening in German[1]

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

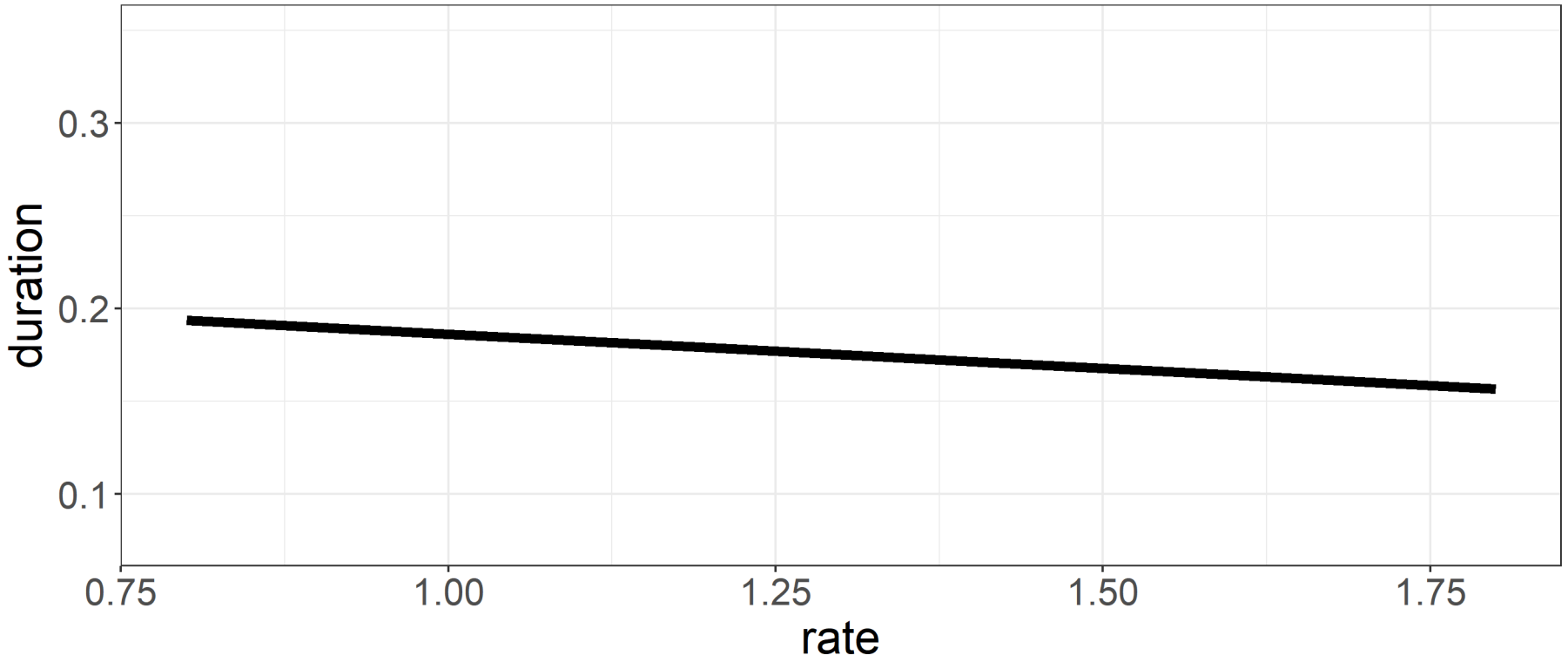


[1] Schmitz, D., Cho, H.-E., & Niemann, H. (2018). Vowel shortening in German as a function of syllable structure. Proceedings 13. Phonetik Und Phonologie Tagung (P&P13), 181–184.

Beispieldaten

- Bisher haben wir Vokaldauer mit Einfacher Linearer Regression…

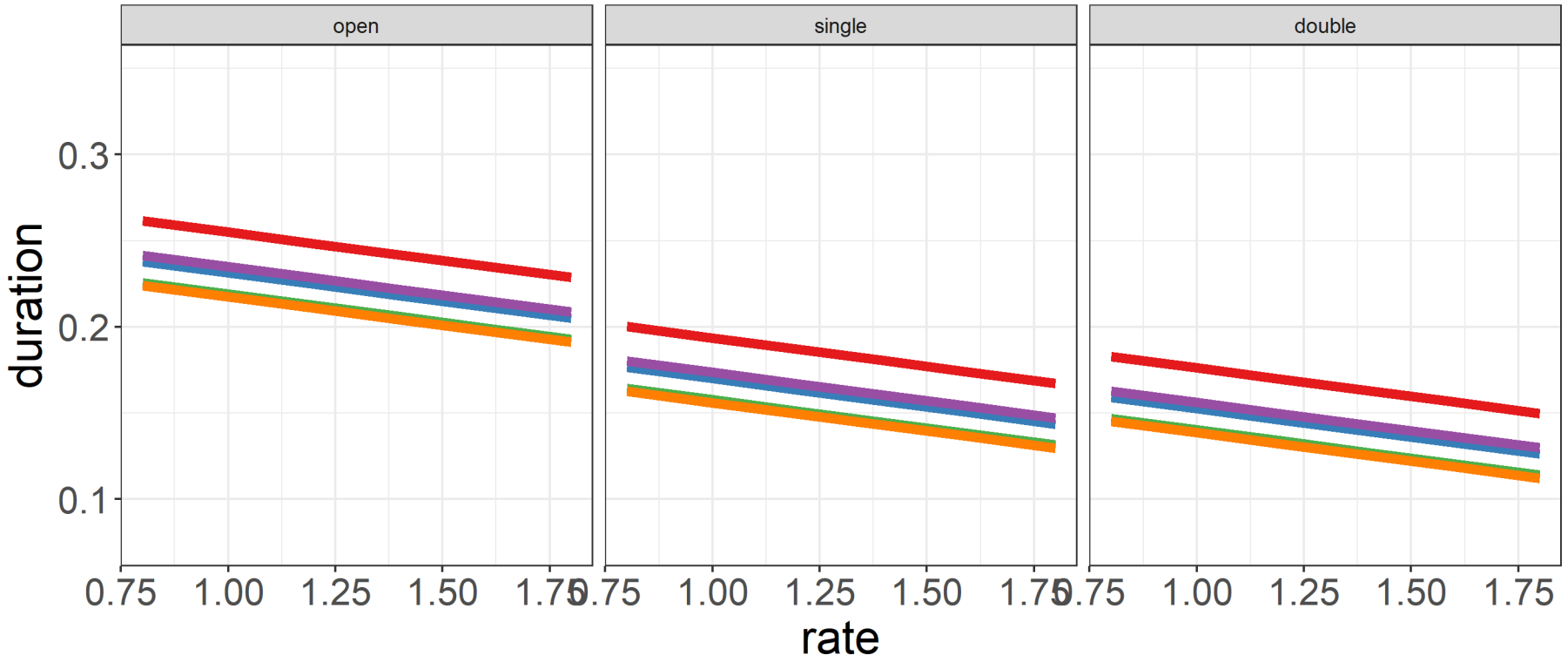
```
lm(duration ~ rate, data_v)
```



Beispieldaten

- … und Multipler Linearer Regression modeliert

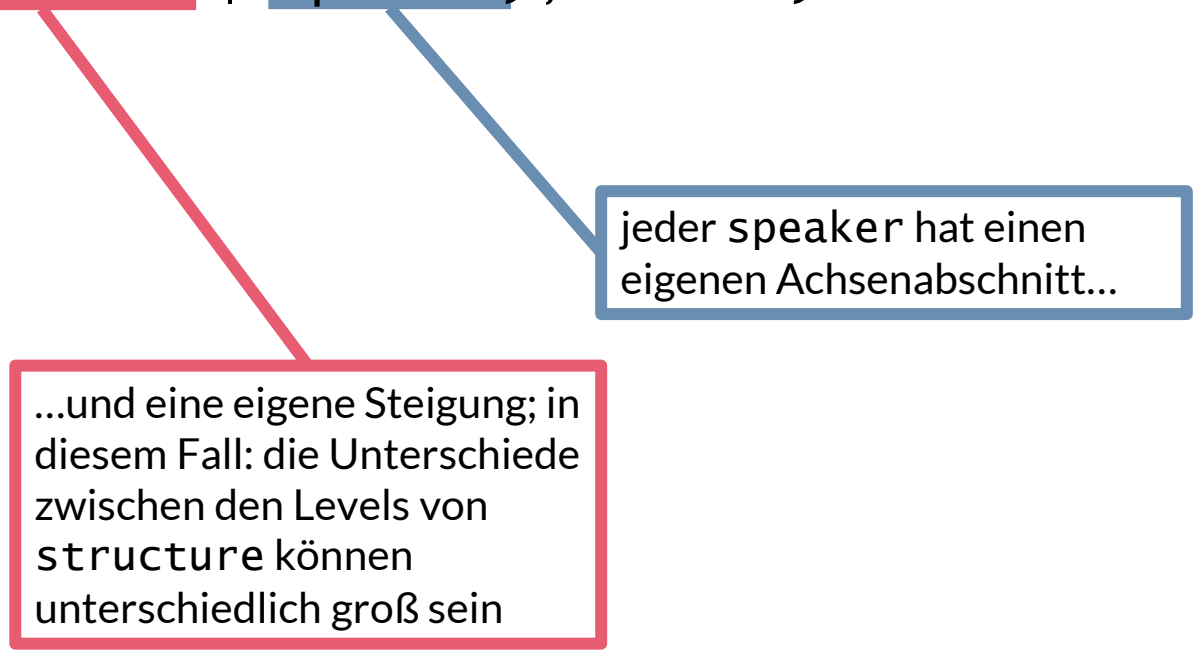
```
lm(duration ~ rate + structure + vowel, data_v)
```



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 Prädiktoren

1. Fixed Effects

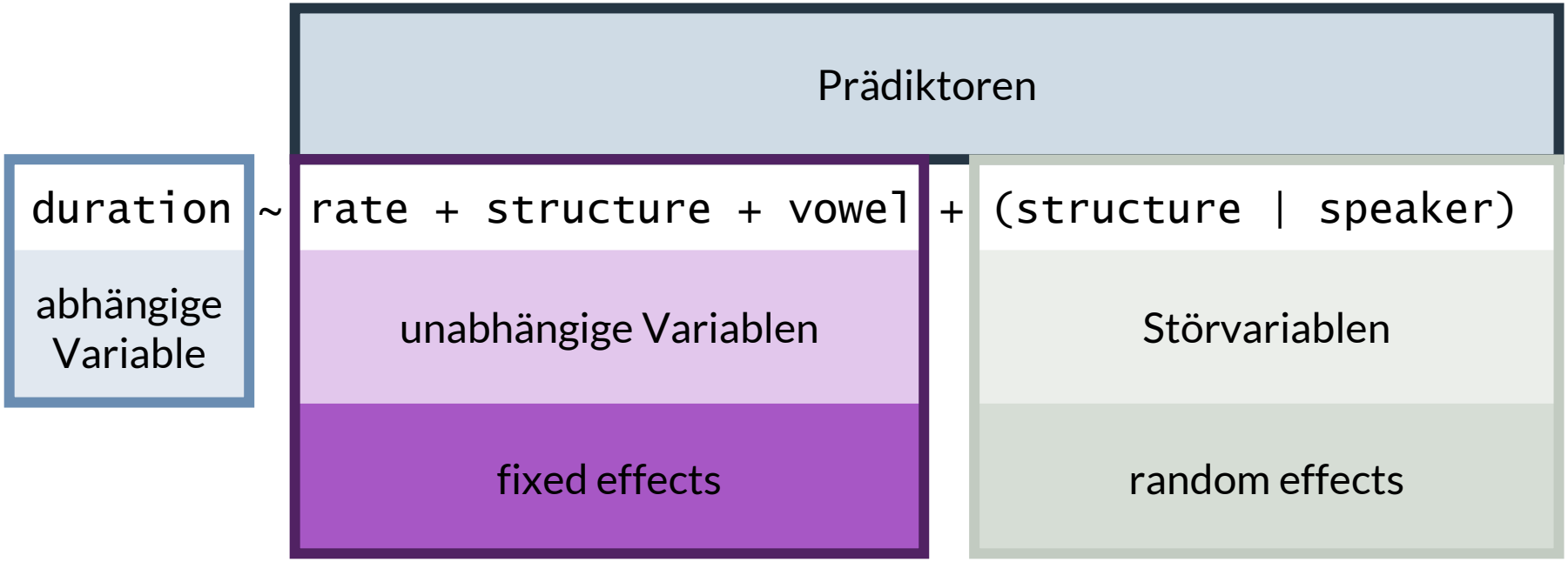
- erklärende Variablen
- Variablen, die im Mittelpunkt stehen
- wiederholbar

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
  - Videospielfrequenz — frequenter = kürzere RT
  - …
- und immer: die Variable, die uns interessiert = **variable of interest**

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

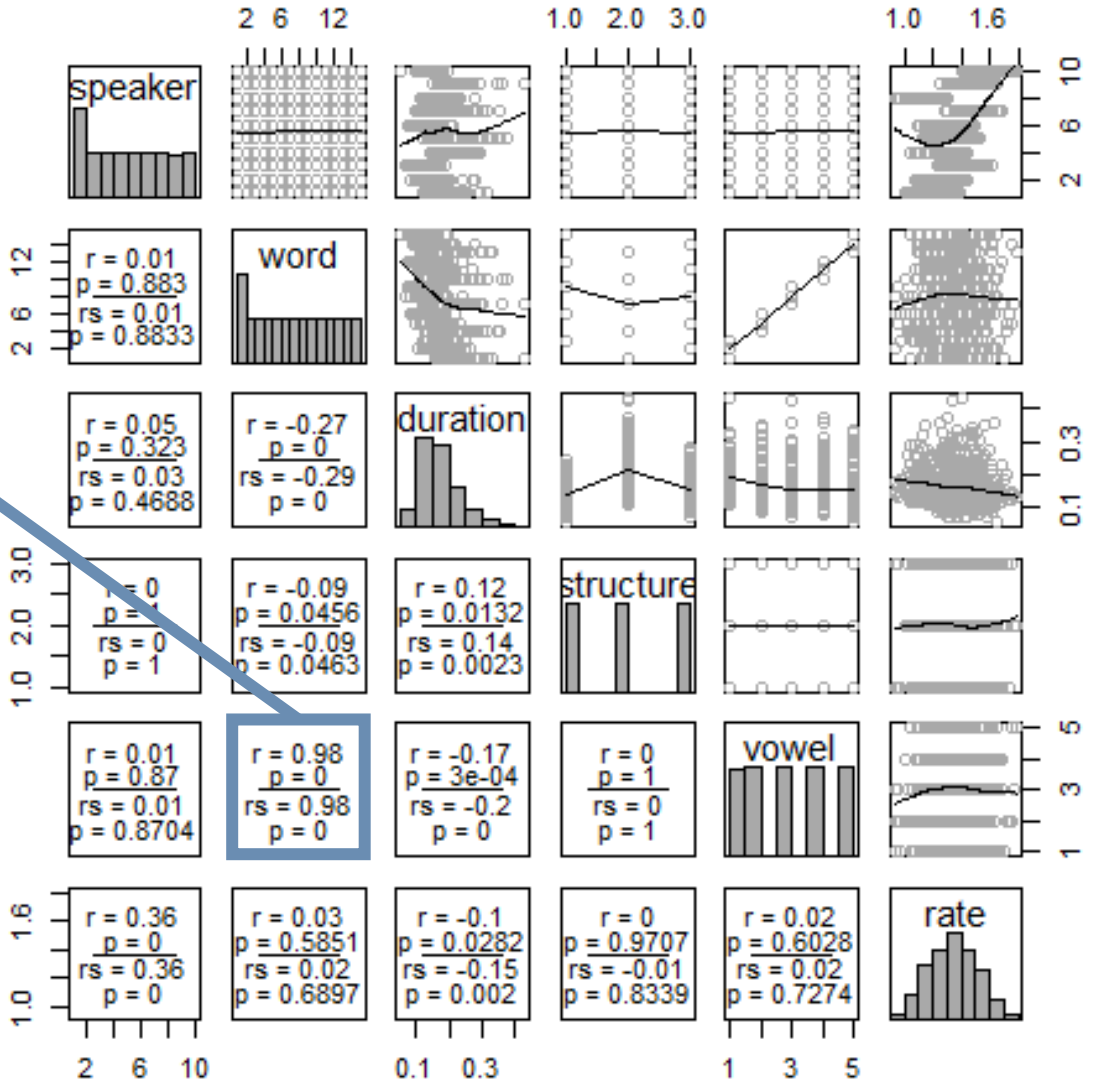
- Wie wir bereits festgestellt haben, bedeuten mehr Variablen auch mehr Arbeitsschritte
- Typische Schritte sind
    1. Verteilung der abhängigen Variable überprüfen
    2. Check der Korrelationen wegen Kollinearität
    3. „volles“ Modell erstellen
    4. „bestes“ Modell finden
    5. Annahmen überprüfen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

Gemischte Modelle in R

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

2. Check der Korrelationen

kein Problem, wenn starke Korrelation für ein Paar aus abhängiger und unabhängiger Variablen gefunden wird



3. „Volles“ Modell

- Erstellen eines vollen Modells:

```
library(lme4)
```

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

4. „Bestes“ Modell

- Finden des „besten“ Modells

```
step(mdl)

...

...

...

Model found:

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

5. Annahmen überprüfen

- Multiple Lineare Regression folgt den gleichen Annahmen, denen auch Einfache und Multiple Lineare Regression folgen
  - Linearität / Linearity
  - Homoskedastizität / Homoscedasticity
  - Normalität / Normality
  - Unabhängigkeit / Independence

- **Hinweis**: Die SfL Datensätze sind i.d.R. zu klein um Gemischte Modelle zu erstellen, die allen Annahmen entsprechen

6. Interpretation

- Generell sind wir an zwei Dingen interessiert:
    1. den ***p*-Werten** der einzelnen Prädiktoren
    2. den **Effekten** der einzelnen Prädiktoren

6. Interpretation – *p*-Werte

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 ***
```

6. Interpretation – Effects

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

6. Interpretation des Modells

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

```
                      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 ***
```

6. Interpretation des Modells

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

```
            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 ***
```

…einmal tief durchatmen

Kollinearität und andere Probleme

- Bisher haben wir relativ naiv Modelle erstellt
  - irgendeine Variable wird vorhergesagt
  - irgendwelche Variablen sagen vorher
- Problem: Potentielle Gefahren
  1. **Nicht-Normalverteilung der Variablen**
  2. **Kollinearität**
  3. Interaktionen

Kollinearität und andere Probleme

- Bisher haben wir relativ naiv Modelle erstellt
  - irgendeine Variable wird vorhergesagt
  - irgendwelche Variablen sagen vorher
- Problem: Potentielle Gefahren
  1. **Nicht-Normalverteilung der Variablen**
  2. **Kollinearität**
  3. **Interaktionen**

Interaktionen

- Variablen können unabhängig voneinander Effekte zeigen – das ist das, was wir bisher modelliert und beobachtet haben
- Variablen können allerdings auch abhängig voneinander Effekte zeigen – dies nennt man **Interaktion**
- Interaktionen sollte man nur dann modellieren, wenn man sie
  1. interpretieren kann
  2. konzeptionell/theoretisch begründen kann
- Eine womöglich signifikante Interaktion, die man nicht interpretieren oder erklären kann, ist für die statistische Analyse und den Erkenntnisgewinn wertlos

Interaktionen

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

Cuteness amplifies effects of size sound symbolism:

A cute /i/ is smaller than an ugly one[1]

- Je niedlicher die Kreatur ist, die zu einem Pseudowort gehört, desto größer wird sie bei Pseudowörtern mit /a/ und desto kleiner wird sie bei Pseudowörtern mit /i/ eingeschätzt

	/a/	/i/
sehr niedlich	größer	kleiner
nicht niedlich	groß	klein

[1] Schmitz, D., Cicek, D., Nguyen, Anh Kim, & Rottleb, D. (2023). Cuteness amplifies effects of size sound symbolism: A cute /i/ is smaller than an ugly one. Manuscript in preparation.

Interaktionen

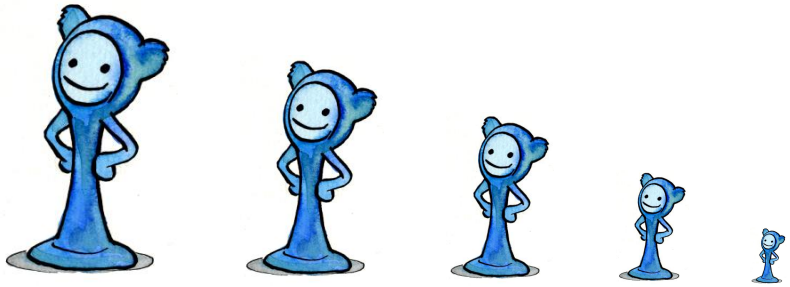
- Für die folgenden Beispiele werden wir Daten folgender Studie nutzen:

Cuteness amplifies effects of size sound symbolism:

A cute /i/ is smaller than an ugly one

Teil 1: Urteile zu Größe

- 5 verschieden große Versionen eines visuellen Stimulus wurden gezeigt



- 1 auditiver Stimulus wurde abgespielt
- Teilnehmende mussten entscheiden zu welchem Bild das gehörte Pseudowort am besten passt

Interaktionen

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

Cuteness amplifies effects of size sound symbolism:

A cute /i/ is smaller than an ugly one

Teil 2: Urteile zur Niedlichkeit

- 1 Version eines visuellen Stimulus wurde gezeigt

- alle visuellen Stimuli wurden in der Gleichen Größe gezeigt
- Teilnehme mussten auf einer Skala (*nicht niedlich* bis *sehr niedlich*) beurteilen wie niedlich sie jedes Alien fanden

Interaktionen

- Unsere Hypothese, dass Niedlichkeit und Größe für zwei bestimmte Vokale zusammenhängen, lässt sich mit einer Interaktion im Modell überprüfen
- Ist die Interaktion signifikant und liefert entsprechende Effekte, ist die Hypothese bestätigt
- Ist die Interaktion nicht signifikant oder liefert Effekte für andere Vokale/in andere Richtungen, ist die Hypothese nicht bestätigt

	/a/	/i/
sehr niedlich	größer	kleiner
nicht niedlich	groß	klein

Interaktionen

- Folgende Variablen nutzen wir für unser Modell
  - `Niedlichkeit` = wie niedlich ein Teilnehmer ein Alien einschätzt
  - `Vokal` = Nucleus des Pseudoworts
  - `phonNachbarschaft` = phonologische Nachbarschaftsdichte
  - `Onset1` = Onset-Konsonant der 1. Silbe
  - `Onset2` = Onset-Konsonant der 2. Silbe

Interaktionen

- Unsere Hypothese, dass Niedlichkeit und Größe für zwei bestimmte Vokale zusammenhängen, lässt sich mit einer Interaktion im Modell überprüfen

Modell ohne Interaktion

```
Größe ~
        Niedlichkeit +
        Vokal +
        phonNachbarschaft +
        Onset1 +
        Onset2
```

Modell mit Interaktion

```
Größe ~
        Niedlichkeit *
        Vokal +
        phonNachbarschaft +
        Onset1 +
        Onset2
```

Interaktionen

- Das Modell sagt also Größe u.a. anhand einer Interaktion von Niedlichkeit und Vokal voraus
- Das Modell räumt für verschiedene Stufen von Niedlichkeit verschiedene Effekte für verschiedene Vokale ein

Modell mit Interaktion

```
Größe ~
        Niedlichkeit *
        Vokal +
        phonNachbarschaft +
        Onset1 +
        Onset2
```

Interaktionen

- Die Summary des erstellten Modells sieht wie folgt aus:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.174	0.168	18.946	0.000	***
phonNachbarschaft	0.029	0.036	0.797	0.426	
Onset1f	0.006	0.180	0.031	0.975	
...	…	…	…	…	
Onset2r	-0.151	0.158	-0.956	0.340	
Niedlich:Vokala	0.177	0.049	3.631	0.000	***
Niedlich:VokalA	0.083	0.055	1.500	0.134	
Niedlich:Vokale	0.040	0.051	0.796	0.426	
Niedlich:Vokali	-0.194	0.056	-3.447	0.001	**
Niedlich:Vokalo	0.053	0.053	0.999	0.318	
Niedlich:VokalO	0.148	0.057	2.568	0.010	*
Niedlich:Vokalu	0.069	0.059	1.165	0.244	
Niedlich:Vokaly	-0.047	0.051	-0.927	0.355	

Interaktionen

- Die Summary des erstellten Modells sieht wie folgt aus:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.174	0.168	18.946	0.000	***
phonNachbarschaft	0.029	0.036	0.797	0.426	
Onset1f	0.006	0.180	0.031	0.975	
...	…	…	…	…	
Onset2r	-0.151	0.158	-0.956	0.340	
Niedlich:Vokala	0.177	0.049	3.631	0.000	***
Niedlich:VokalA	0.083	0.055	1.500	0.134	
Niedlich:Vokale	0.040	0.051	0.796	0.426	
Niedlich:Vokali	-0.194	0.056	-3.447	0.001	**
Niedlich:Vokalo	0.053	0.053	0.999	0.318	
Niedlich:VokalO	0.148	0.057	2.568	0.010	*
Niedlich:Vokalu	0.069	0.059	1.165	0.244	
Niedlich:Vokaly	-0.047	0.051	-0.927	0.355	

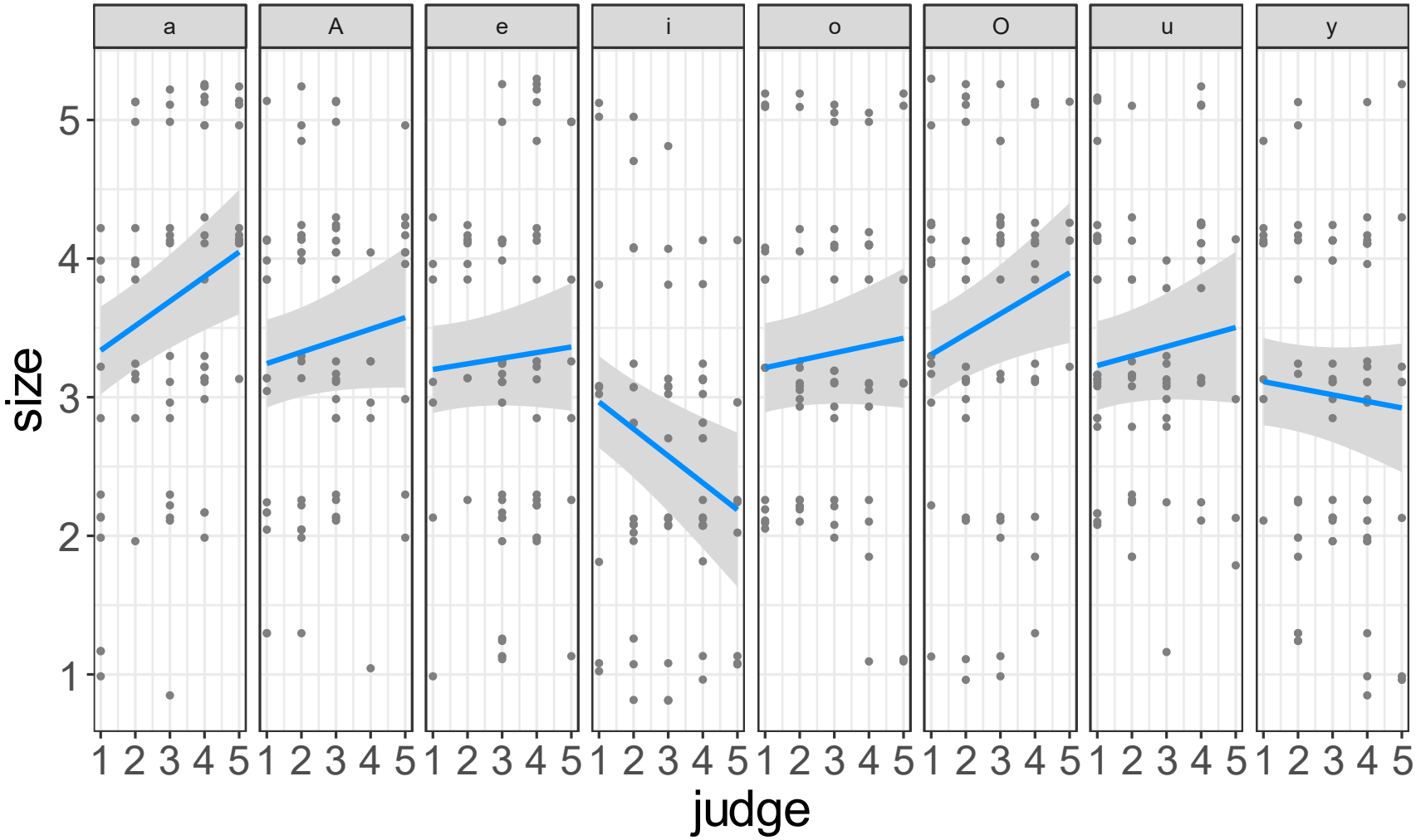
Interaktionen

- Die Summary des erstellten Modells sieht wie folgt aus:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	3.174	0.168	18.946	0.000	***
phonNachbarschaft	0.029	0.036	0.797	0.426	
Onset1f	0.006	0.180	0.031	0.975	
...	…	…	…	…	
Onset2r	-0.151	0.158	-0.956	0.340	
Niedlich:Vokala	0.177	0.049	3.631	0.000	***
Niedlich:VokalA	0.083	0.055	1.500	0.134	
Niedlich:Vokale	0.040	0.051	0.796	0.426	
Niedlich:Vokali	-0.194	0.056	-3.447	0.001	**
Niedlich:Vokalo	0.053	0.053	0.999	0.318	
Niedlich:VokalO	0.148	0.057	2.568	0.010	*
Niedlich:Vokalu	0.069	0.059	1.165	0.244	
Niedlich:Vokaly	-0.047	0.051	-0.927	0.355	

Interaktionen

- So sieht die Interaktion mit `visreg` visualisiert aus:



Kollinearität und andere Probleme

- Bisher haben wir relativ naiv Modelle erstellt
  - irgendeine Variable wird vorhergesagt
  - irgendwelche Variablen sagen vorher
- Problem: Potentielle Gefahren
  1. **Nicht-Normalverteilung der Variablen**
  2. **Kollinearität**
  3. **Interaktionen**