08
Einfache Lineare Regression

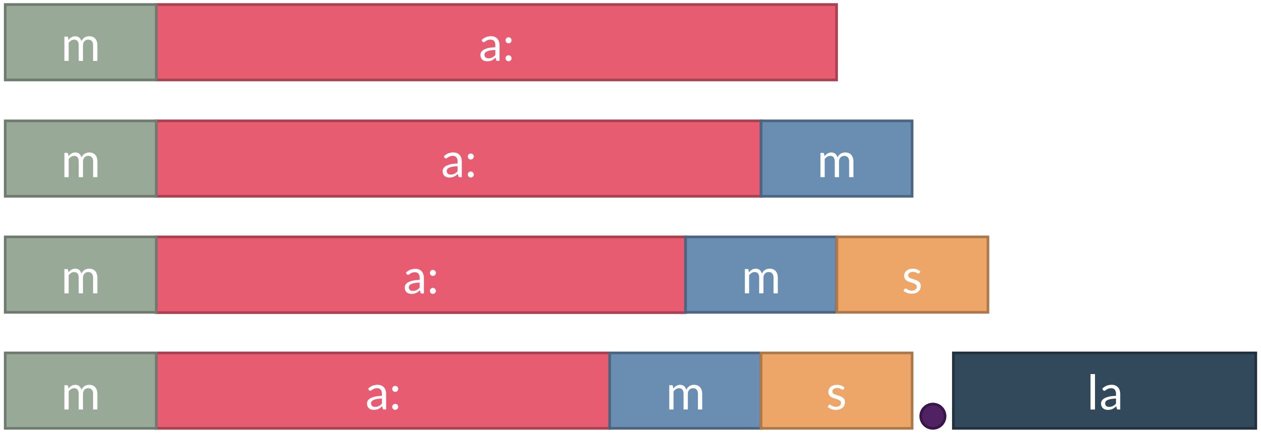
Dominic Schmitz & Janina Esser

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:



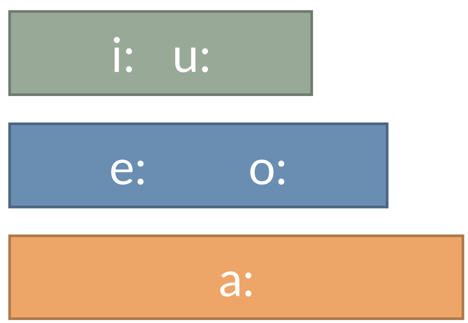
[1] Schmitz, D., Cho, H.-E., & Niemann, H. (2018). Vowel shortening in German as a function of syllable structure. Proceedings 13. Phonetik Und Phonologie Tagung (P&P13), 181–184.

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.

Einfache Lineare Regression: Formel

kontinuierliche abhängige Variable

unabhängige Prädiktorvariable

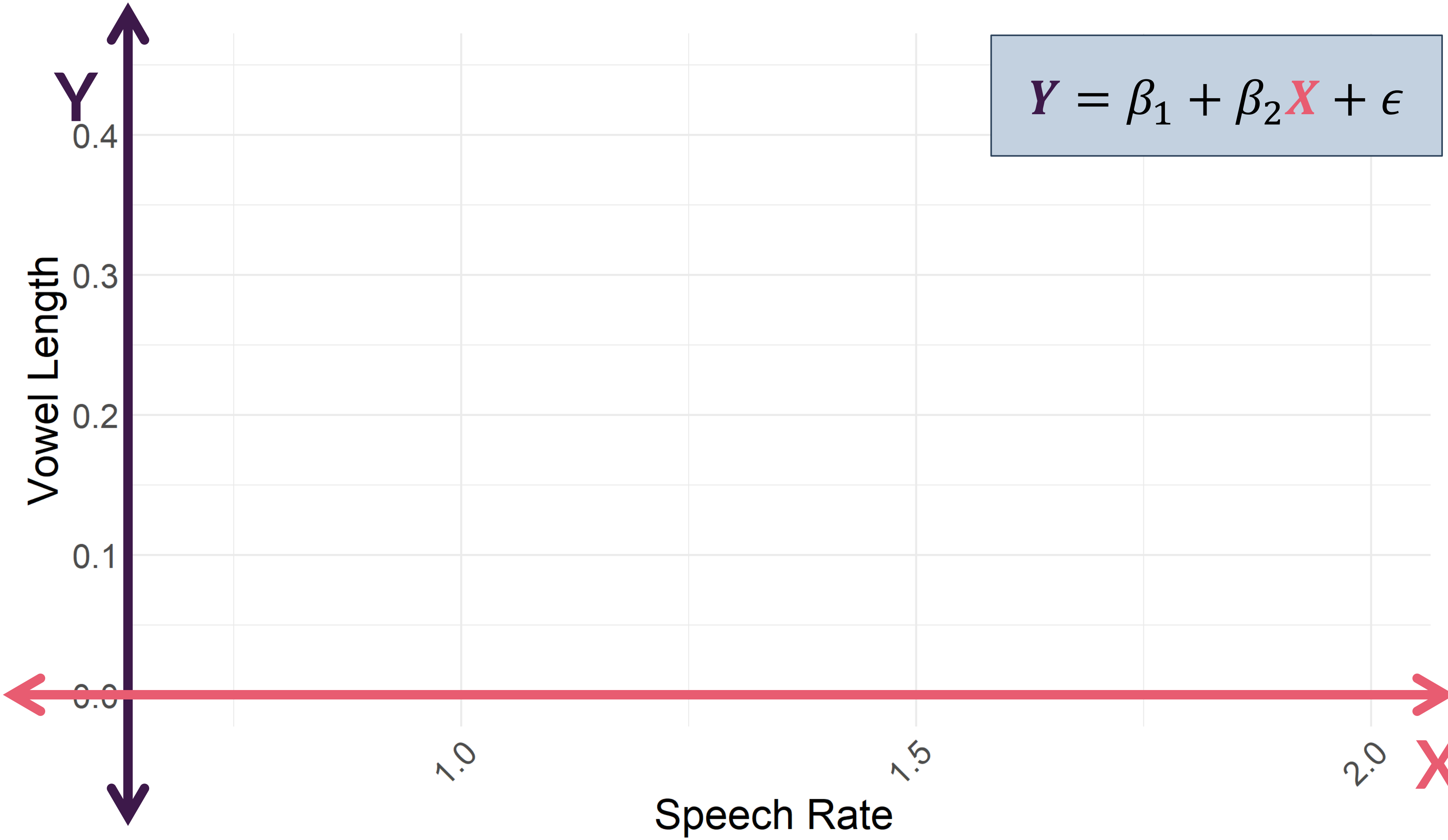
$$Y = \beta_1 + \beta_2 X + \epsilon$$

Steigung/ Slope

Residuen/ Error Term

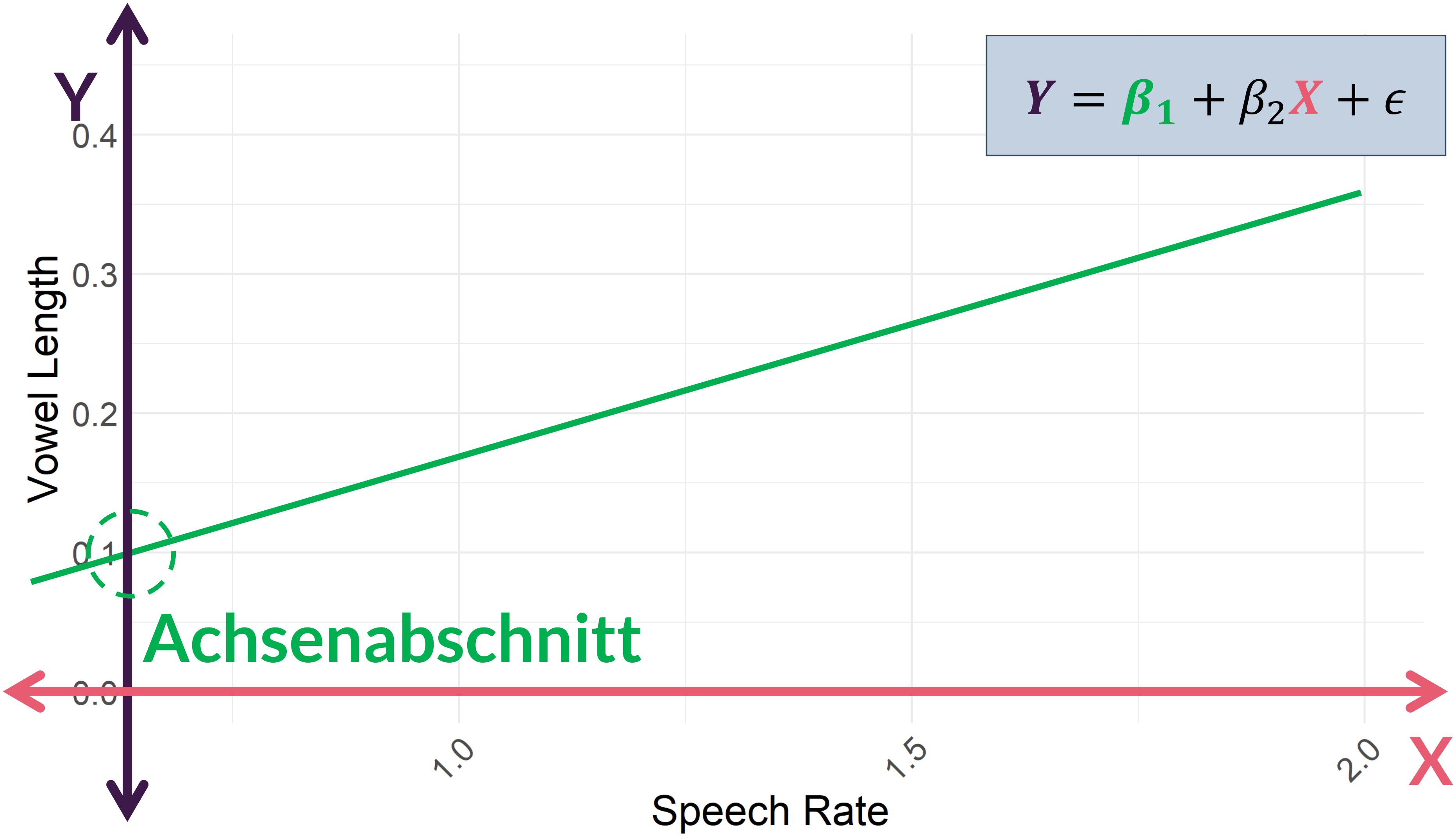
y-Achsenabschnitt/ Intercept

Einfache Lineare Regression: Formel

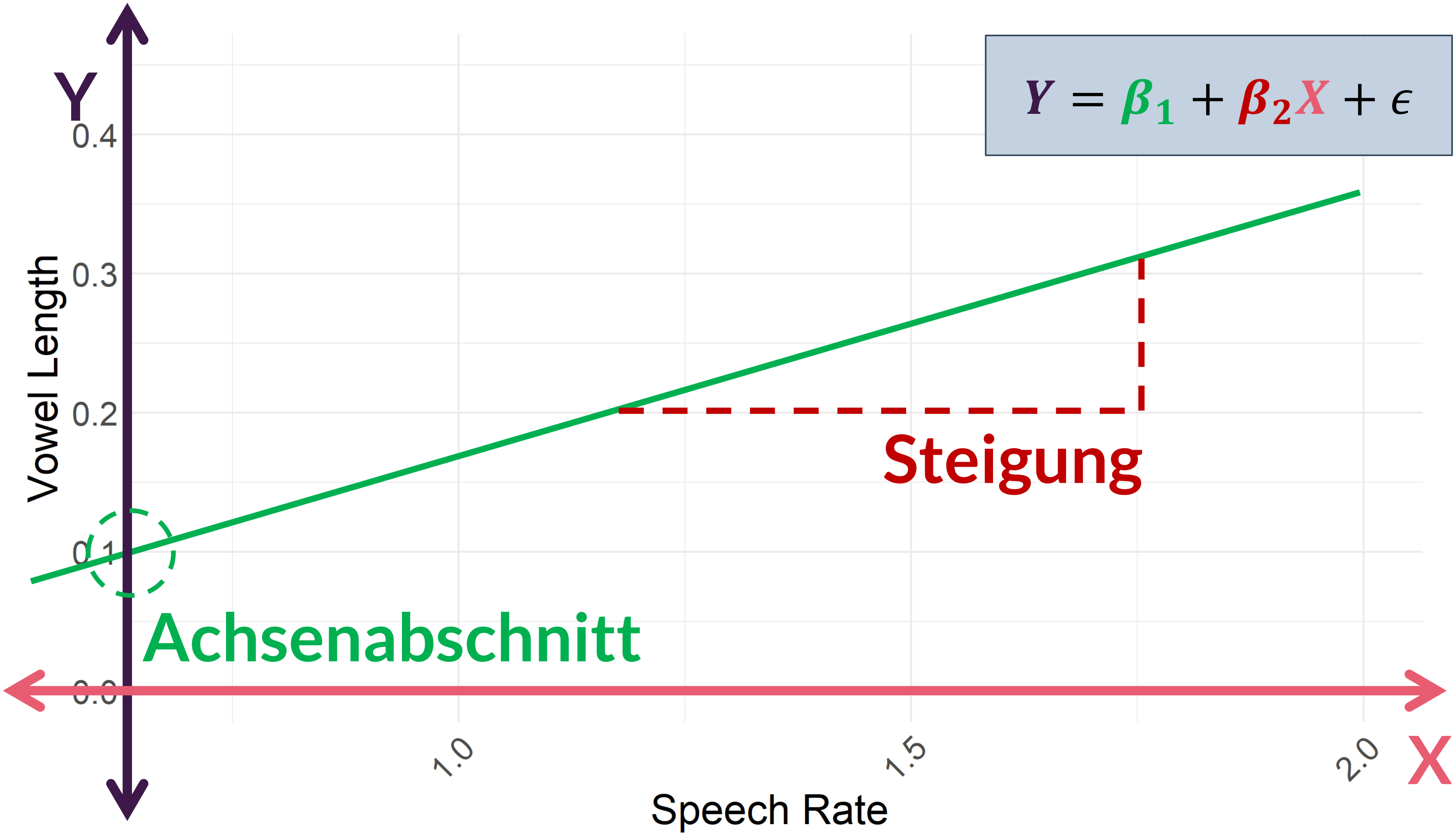


$$Y = \beta_1 + \beta_2 X + \epsilon$$

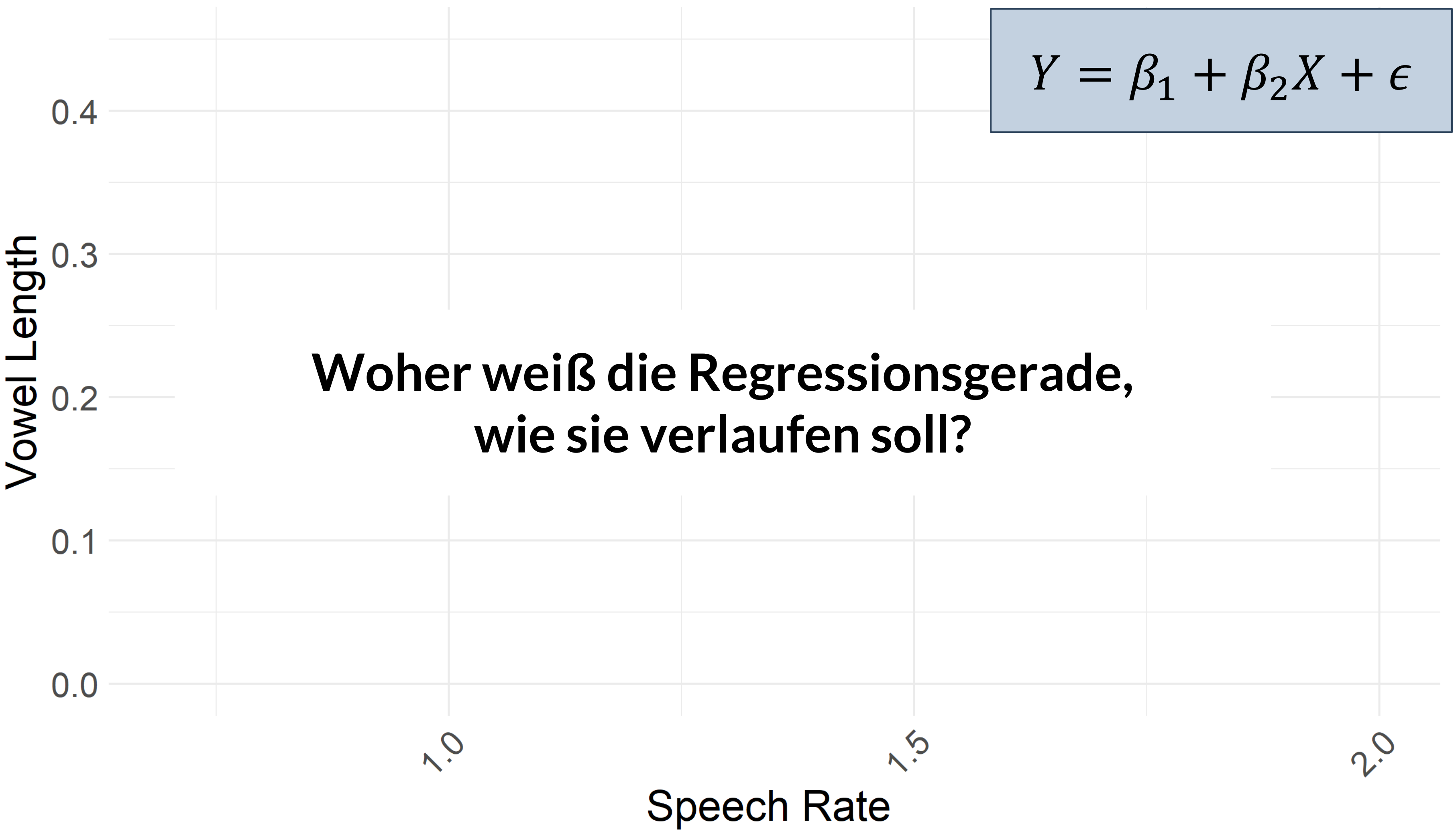
Einfache Lineare Regression: Formel



Einfache Lineare Regression: Formel



Einfache Lineare Regression: Formel

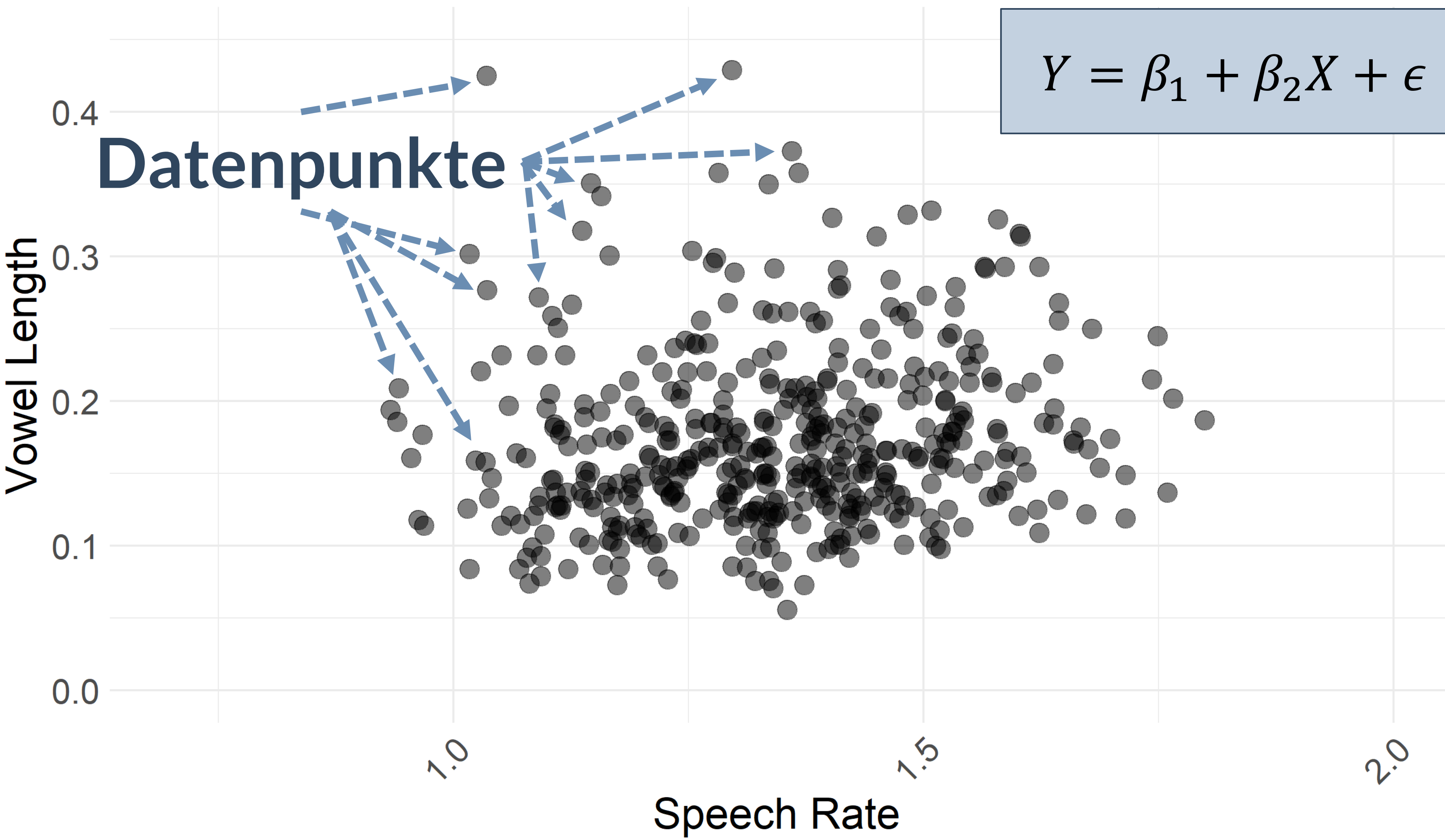


$$Y = \beta_1 + \beta_2 X + \epsilon$$

Woher weiß die Regressionsgerade,
wie sie verlaufen soll?

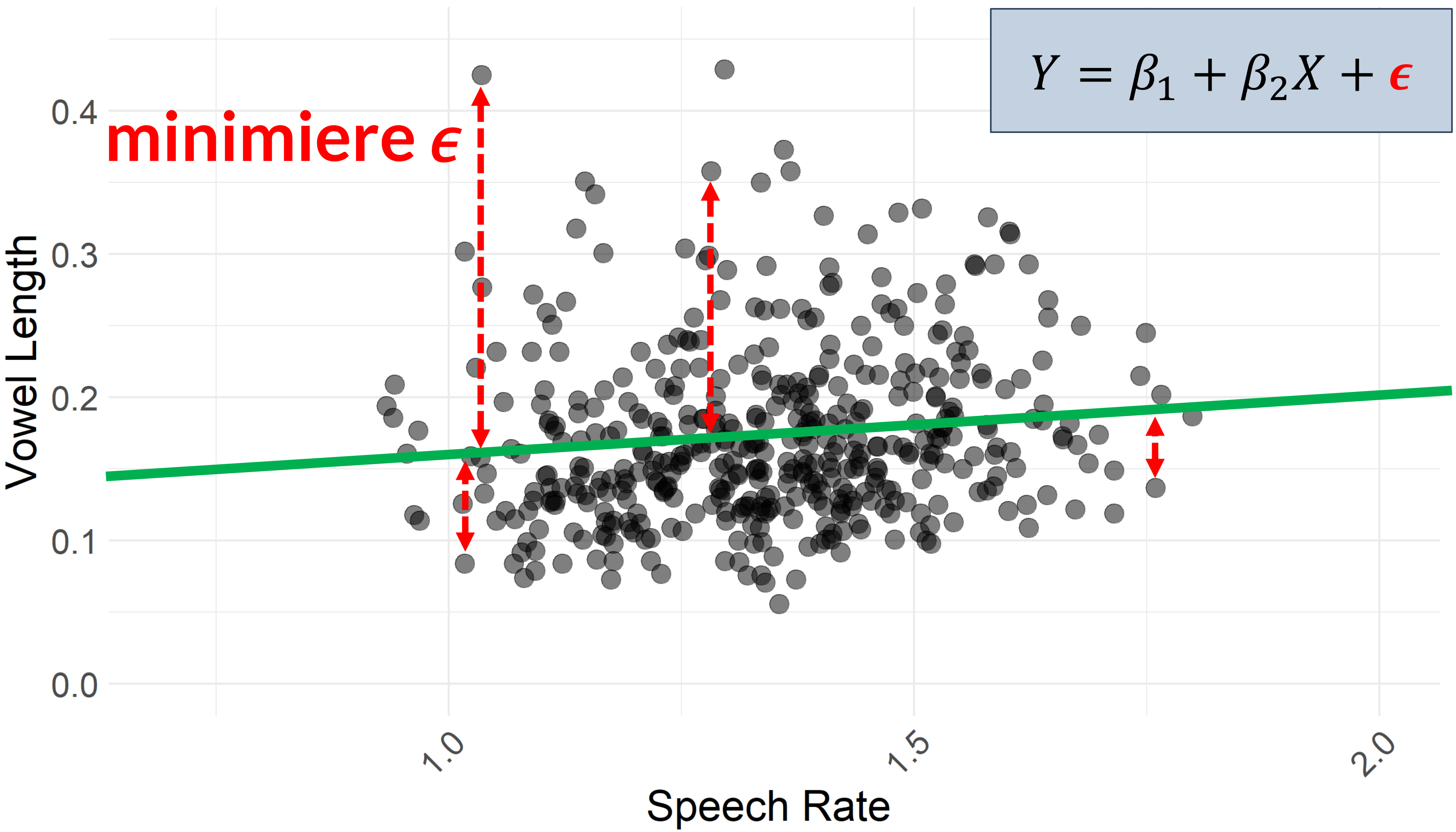
Einfache Lineare Regression: Formel

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



Einfache Lineare Regression: Formel

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



Einfache Lineare Regression in R

- In R erstellt man ein Einfaches lineares Regressionsmodell

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

- mit folgendem Befehl und folgender Syntax:

```
lm(Y ~ X, data)
```

- y-Achsenabschnitt und Steigung berechnet R, indem es die Residuen zwischen tatsächlichen Datenpunkten und der Regressionsgeraden minimiert

Einfache Lineare Regression in R

- Beispiel: vowel duration modelliert durch speech rate

```
model = lm(duration ~ rate, data)
```

- Nach der Berechnung erhalten wir folgende Information zum Modell:

```
Call:
lm(formula = duration ~ rate, data = data)

Coefficients:
(Intercept)       rate
0.22301       -0.03687
```

Einfache Lineare Regression in R

- Beispiel: vowel duration modelliert durch speech rate

```
model = lm(duration ~ rate, data)
```

- Nach der Berechnung erhalten wir folgende Information zum Modell:

```
Call:
lm(formula = duration ~ rate, data = data)

Coefficients:
(Intercept)        rate
0.22301       -0.03687
```

Achsenabschnitt Steigung

Einfache Lineare Regression in R

- Einen *p*-Wert erhalten wir mit der `anova()` Funktion:

```
anova(model)
Analysis of Variance Table

Response: duration
           Df  Sum Sq   Mean Sq F value  Pr(>F)
rate        1 0.01787 0.0178734  4.8468 0.02821 *
Residuals 446 1.64468 0.0036876
```

Einfache Linear Regression in R

```
           Df  Sum Sq   Mean Sq F value  Pr(>F)
rate        1 0.01787 0.0178734  4.8468 0.02821 *
Residuals 446 1.64468 0.0036876
```

Freiheitsgrade

Die Anzahl der unabhängigen Informationen, die in die Berechnung der Schätzung der jeweiligen Variable einfließen.

Einfache Linear Regression in R

```
             Df  Sum Sq   Mean Sq F value  Pr(>F)
rate          1 0.01787 0.0178734  4.8468 0.02821 *
Residuals 446 1.64468 0.0036876
```

Quadratsumme / Squared Sum

Je höher der Wert, desto wichtiger ist die Variable für das Modell.

Einfache Linear Regression in R

```
           Df  Sum Sq   Mean Sq F value  Pr(>F)
rate        1 0.01787 0.0178734  4.8468 0.02821 *
Residuals 446 1.64468 0.0036876
```

Quadratisches Mittel / Squared Mean

Je höher der Wert, desto wichtiger ist die Variable für das Modell.

Einfache Linear Regression in R

```
           Df  Sum Sq   Mean Sq F value  Pr(>F)
rate        1 0.01787 0.0178734  4.8468 0.02821 *
Residuals 446 1.64468 0.0036876
```

Fisher-Wert

Je höher der Wert ist, desto mehr Einfluss hat die Variable auf die abhängige Variable.

Einfache Linear Regression in R

```
             Df  Sum Sq   Mean Sq F value  Pr(>F)
rate          1 0.01787 0.0178734  4.8468 0.02821 *
Residuals 446 1.64468 0.0036876
```

p-Wert / Probability Value

Zeigt an, ob die jeweilige Variable einen signifikanten Einfluss auf die abhängige Variable hat.

Einfache Linear Regression in R

```
             Df  Sum Sq   Mean Sq F value  Pr(>F)
rate          1 0.01787 0.0178734  4.8468 0.02821 *
Residuals 446 1.64468 0.0036876
```

Residuen

Der Fehler, der nicht durch die unabhängigen Variablen/Faktoren erklärt wird. $\rightarrow \epsilon$

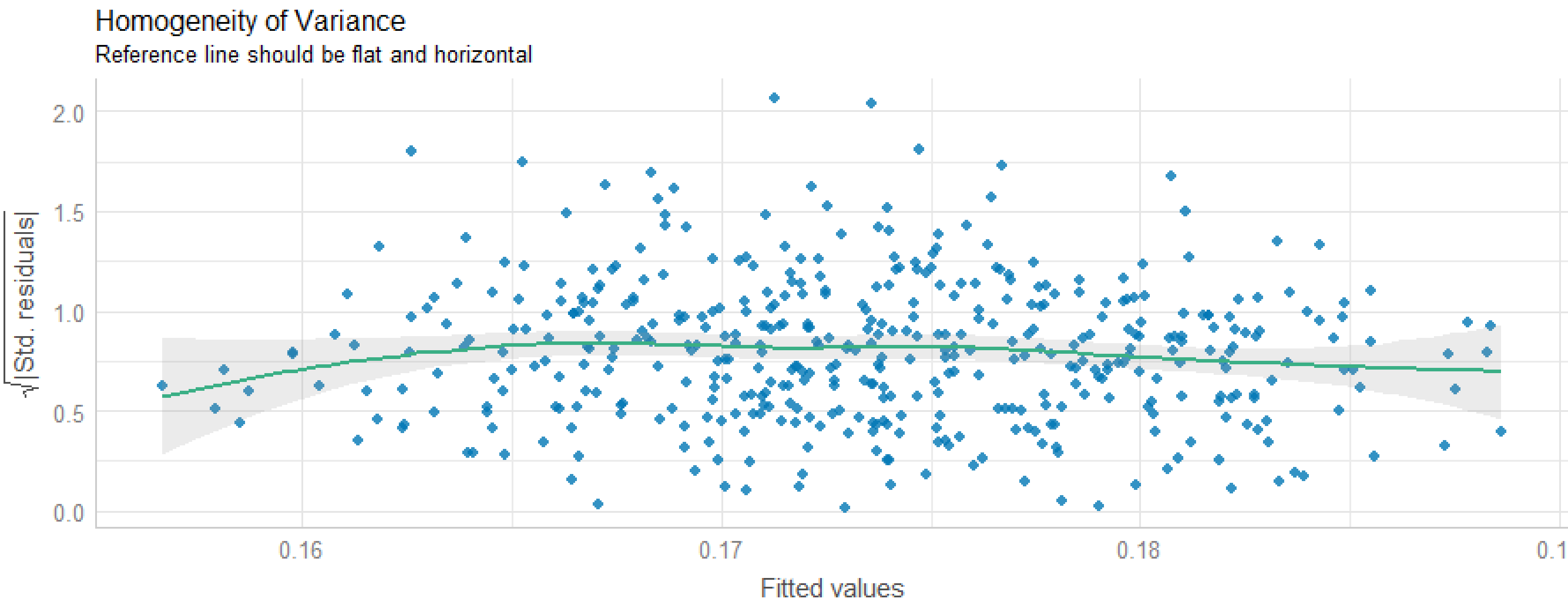
Annahmen

- Laut unserem Modell sinkt die Vowel Duration signifikant, wenn die Speaking Rate ansteigt
- Allerdings wissen wir gar nicht, ob unser Modell zuverlässig ist – wir haben nicht überprüft, ob es den **Annahmen** linearer Regression folgt:
  - Linearität / Linearity
  - Homoskedastizität / Homoscedasticity
  - Normalität / Normality
  - Unabhängigkeit / Independence

Annahmen: Linearität

- Annahme:

  Die Beziehung zwischen X und dem Mittelwert von Y ist linear.

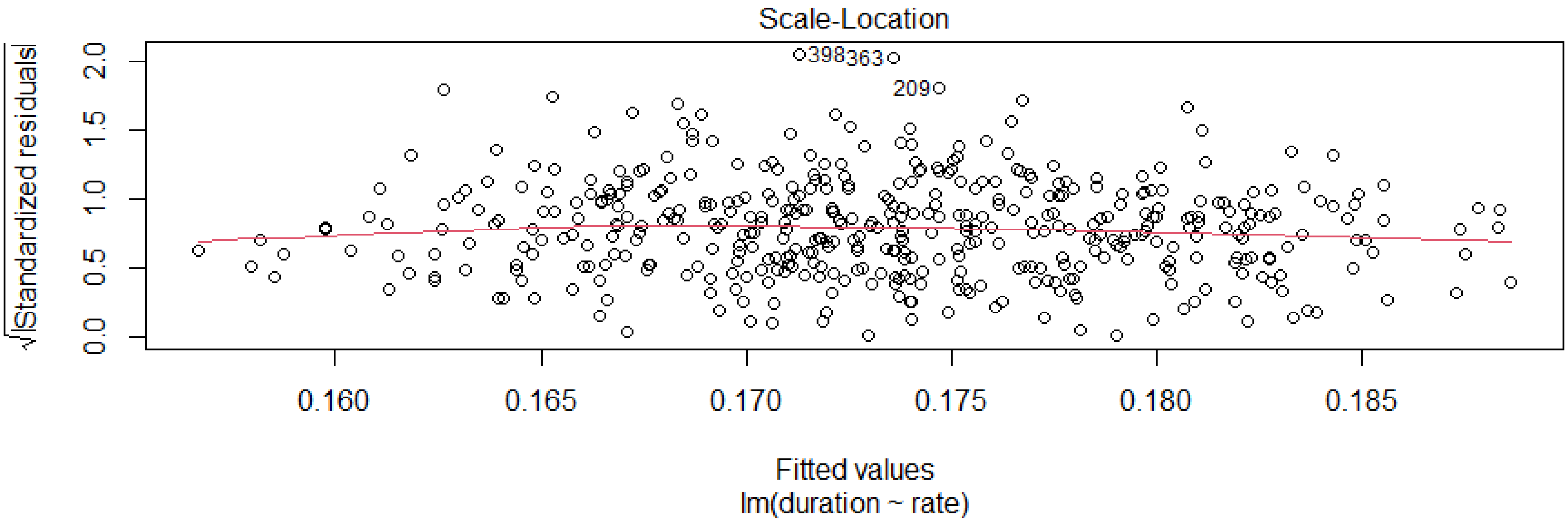


- Die Linie sollte horizontal und flach verlaufen.

Annahmen: Homoskedastizität

- Annahme:

  Die Varianz der Residuen ist für jeden Wert von X gleich.

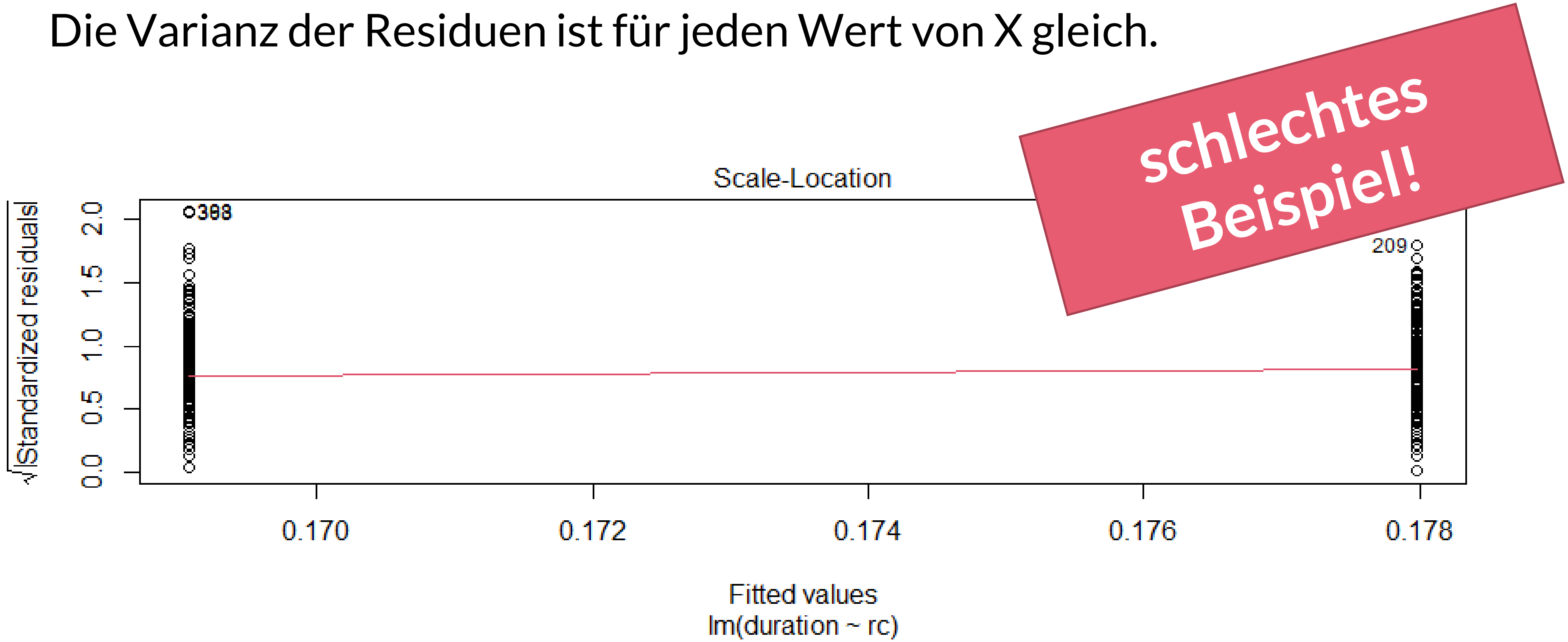


- Die Daten sollten gleichmäßig über die Linie verteilt sein und keine offensichtlichen Muster aufweisen.

Annahmen: Homoskedastizität

- Annahme:

  Die Varianz der Residuen ist für jeden Wert von X gleich.

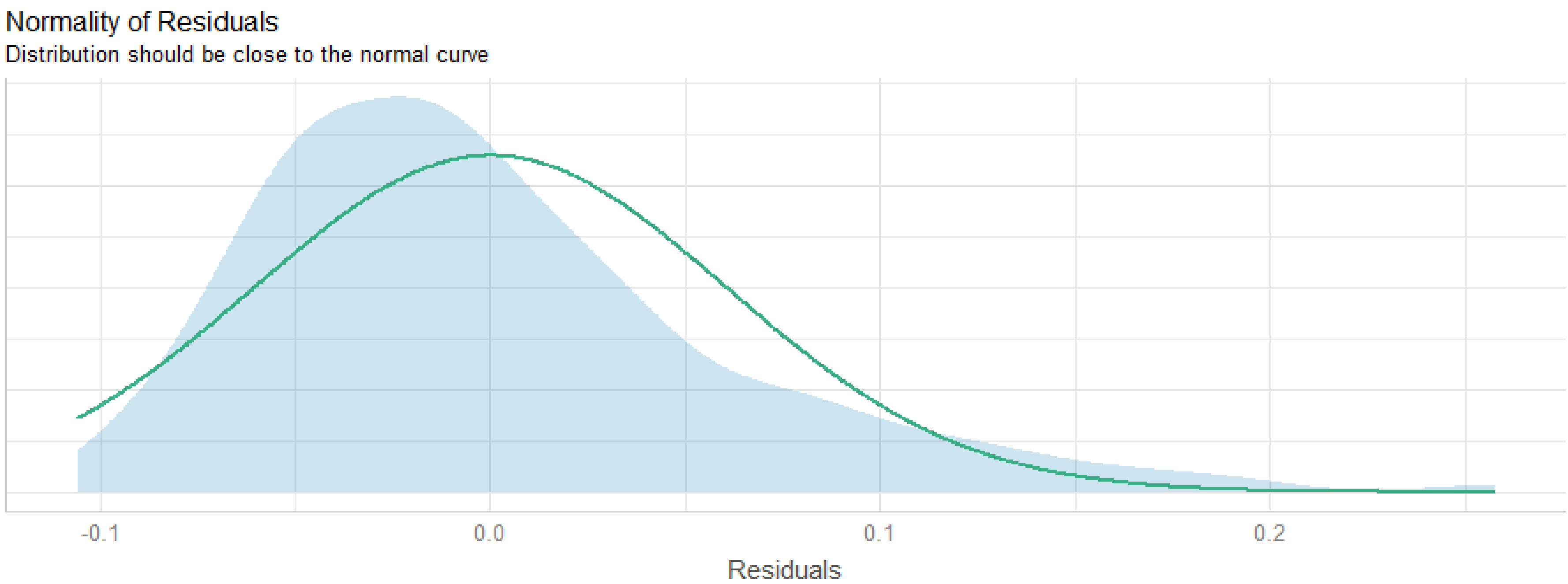


- Die Daten sollten gleichmäßig über die Linie verteilt sein und keine offensichtlichen Muster aufweisen.

Annahmen: Normalität

- Annahme:

  Für jeden festen Wert von X ist Y normalverteilt.

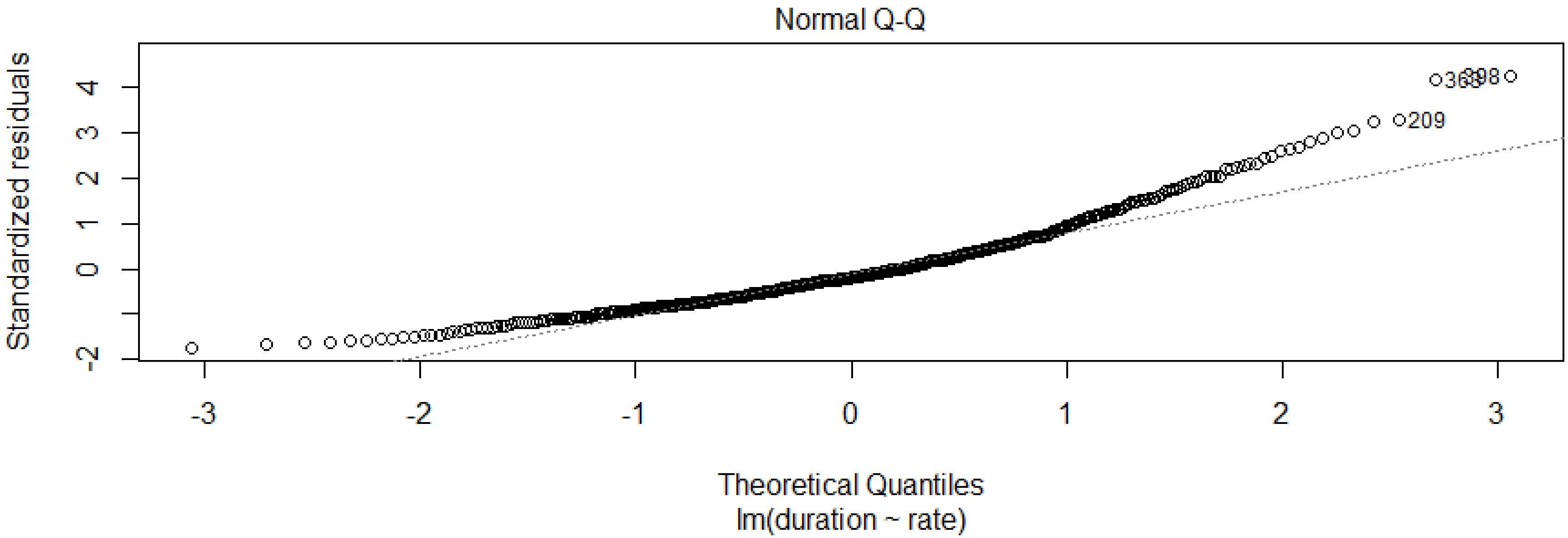


- Die Verteilung der Residuen eines linearen Modells sollte einer Normalverteilung folgen.

Annahmen: Normalität

- Annahme:

  Für jeden festen Wert von X ist Y normalverteilt.



- Die einzelnen Residualwerte sollten auf der Linie liegen.

Annahmen: Unabhängigkeit

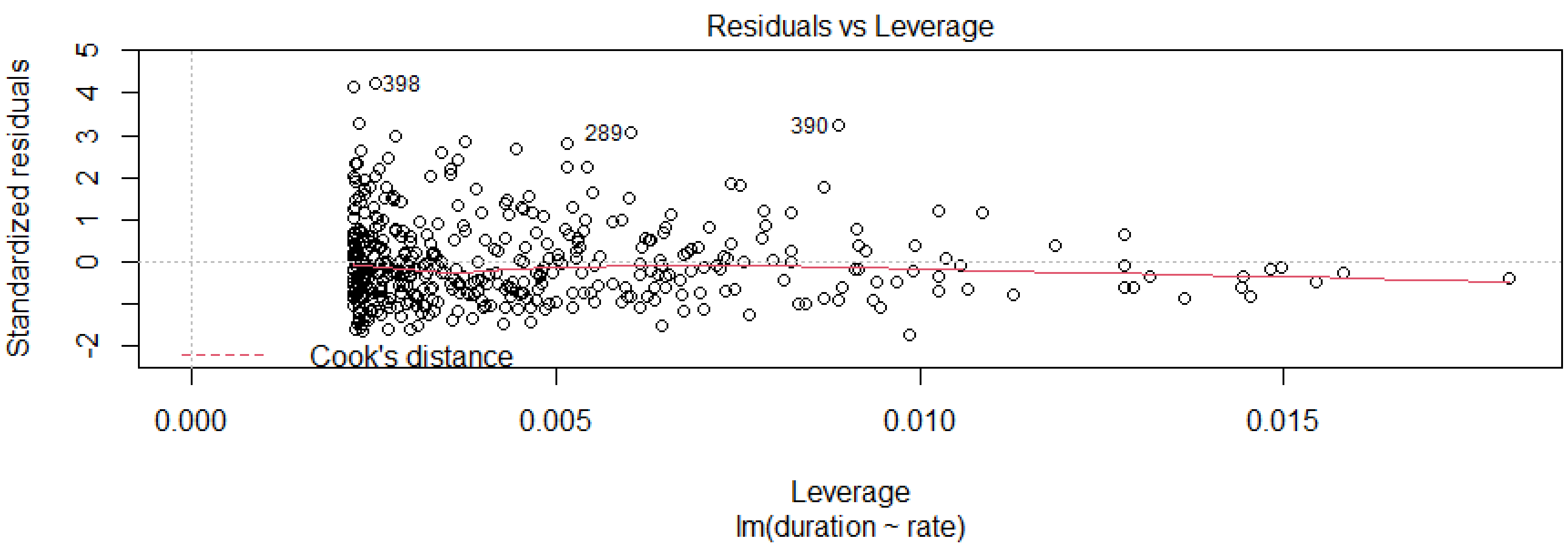
- Annahmen:

  Beobachtungen sind unabhängig voneinander.

- Unabhängigkeit kann nicht visuell überprüft werden
- Es handelt sich um eine Annahme, die durch Untersuchung des Studiendesigns getestet wird

Extra: Beeinflussende Datenpunkte

- Cook-Abstand:
  - Ein Maß für den Einfluss der einzelnen Beobachtungen auf die Regressionskoeffizienten
  - Jede Beobachtung, für die der Cook-Abstand nahe bei 1 liegt oder die wesentlich größer ist als andere Cook-Abstände, muss untersucht werden.



Überprüfung der Verteilung

- Lineare Regressionsmodelle sind zuverlässiger, wenn ihre abhängige Variable einer Normalverteilung folgt
- Daher sollte man vor dem Erstellen von Modellen überprüfen, ob die abhängige Variable diese Voraussetzung erfüllt
- Falls die Verteilung fernab einer Normalverteilung liegt, ist es ratsam die Variable zu **transformieren**
- In seltenen Fällen hilft keine Transformation dabei, die Variable näher an eine Normalverteilung zu bringen – hier kann Lineare Regression dennoch genutzt werden

Überprüfung der Verteilung

- Wie wir bereits gelernt haben, kann man die Verteilung einer Variable mit einem **Shapiro-Wilk Test** überprüfen
- Je höher der *p*-Wert, desto normaler verteilt ist die Variable

```
shapiro.test(data$duration)

      Shapiro-Wilk normality test

data:  data$duration
W = 0.93844, p-value = 1.171e-12
```

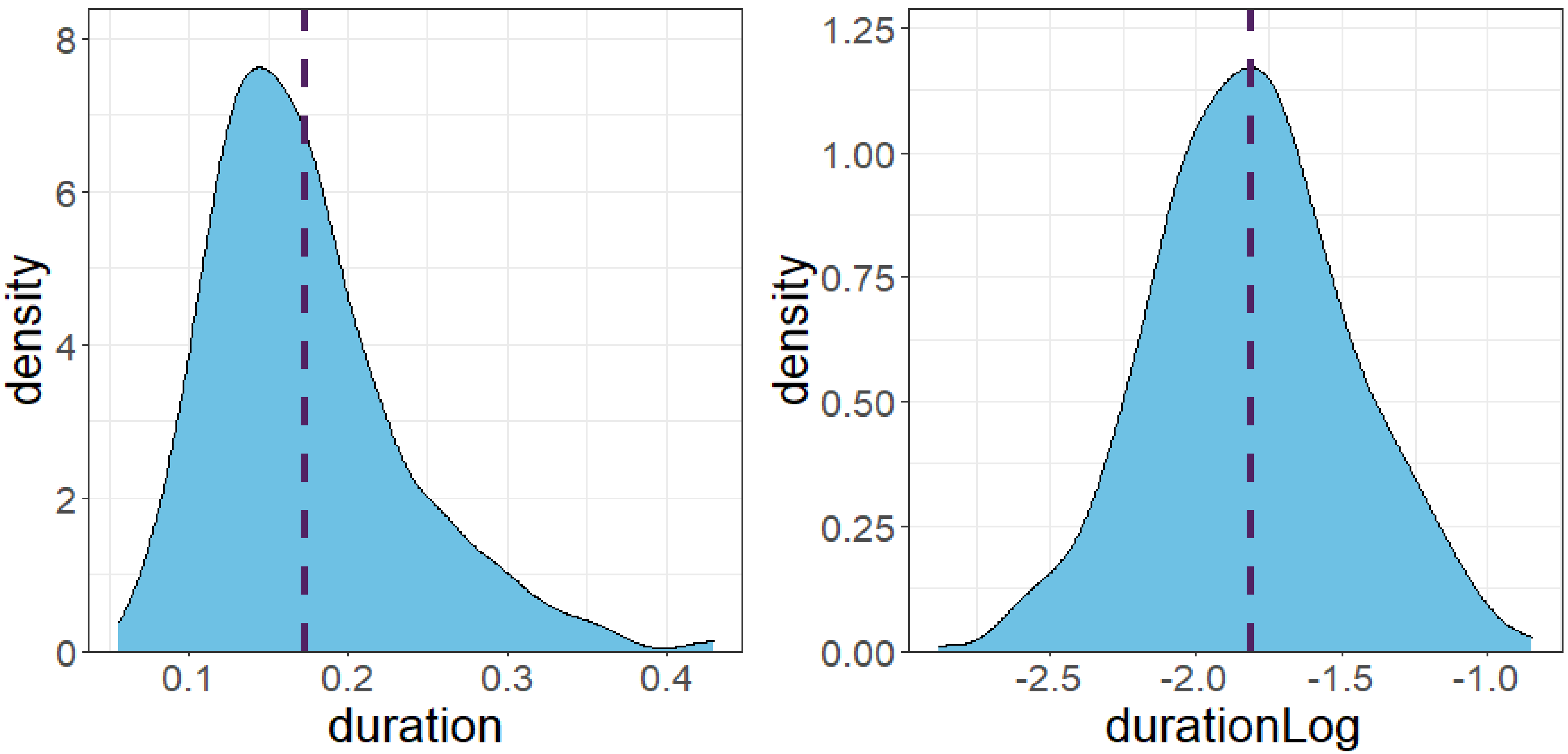
Überprüfung der Verteilung

- Duration ist **nicht normal verteilt**; der *p*-Wert ist extrem niedrig
- Daher erstellen wir eine **log-transformierte** (= logarithmierte) **Version**

```
data$durationLog = log(data$duration)

shapiro.test(data$durationLog)

      Shapiro-Wilk normality test

data:  data$duration
W = 0.99762, p-value = 0.7798
```

Überprüfung der Verteilung

- Eine Visualisierung zeigt deutlich, dass die transformierte Variable normalverteilter ist



Überprüfung der Verteilung

- Wenn wir das zuvor erstellte Modell nun mit der log-transformierten Duration-Variable erneut erstellen, finden wir eine Verbesserung für die Normalitäts-Annahme