Statistics for Linguistics

Session 04

Simple Linear Regression

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)

Simple Linear Regression Formula

continuous
dependent variable

predictor variable

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

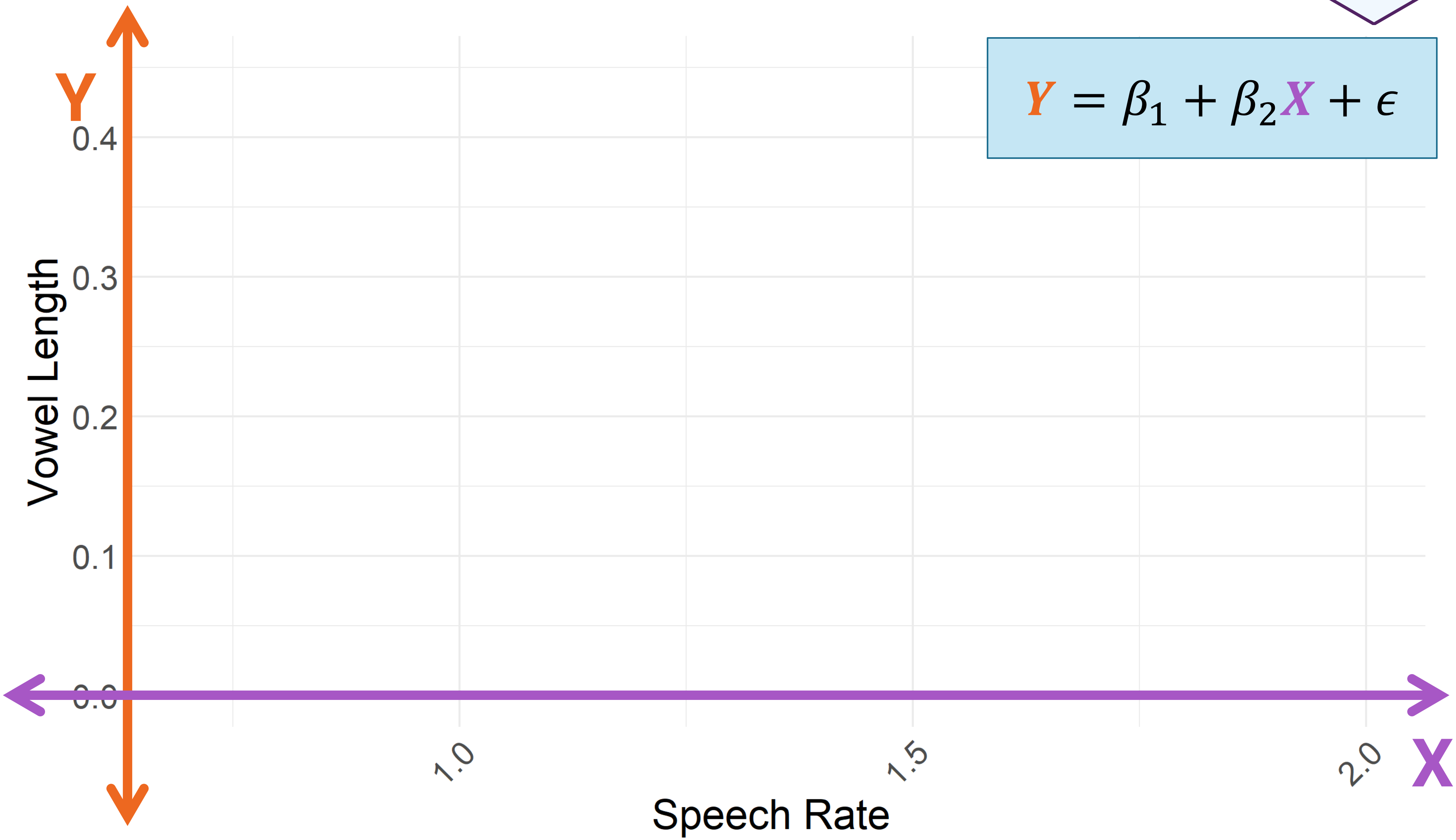
intercept

slope

error term /
residuals

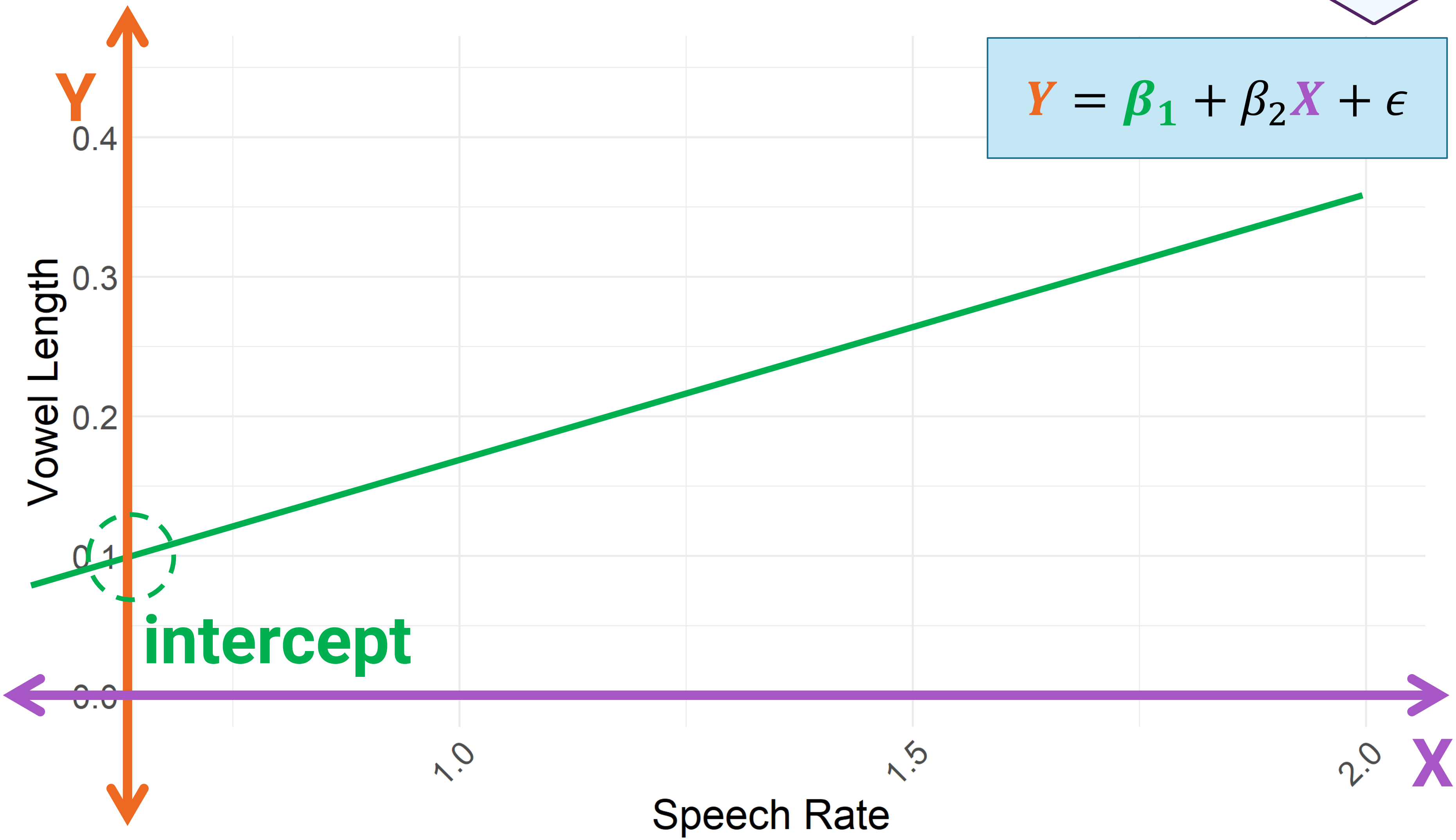
Simple Linear Regression Formula

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



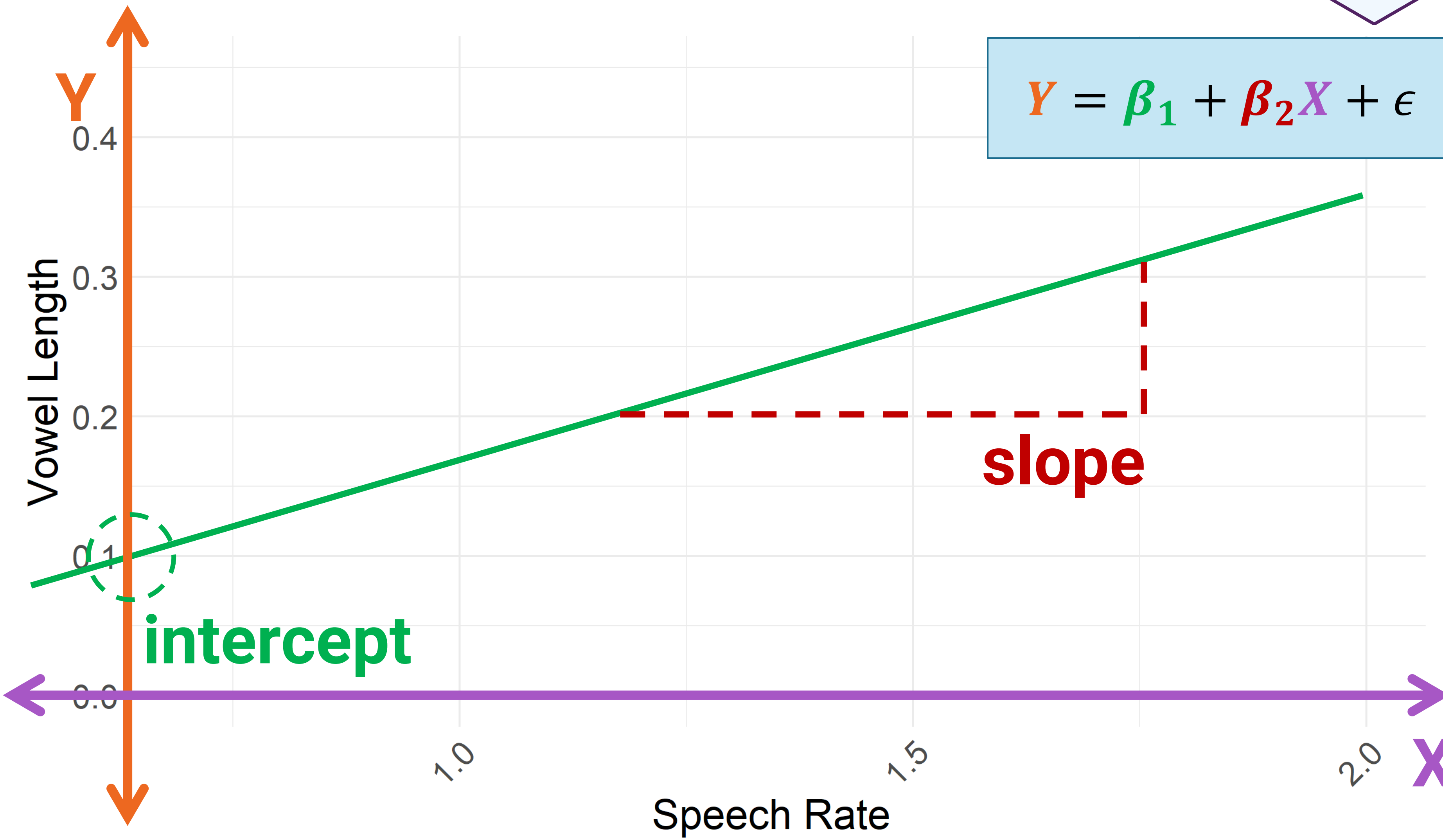
Simple Linear Regression Formula

$$Y = \boldsymbol{\beta_1} + \beta_2 X + \epsilon$$



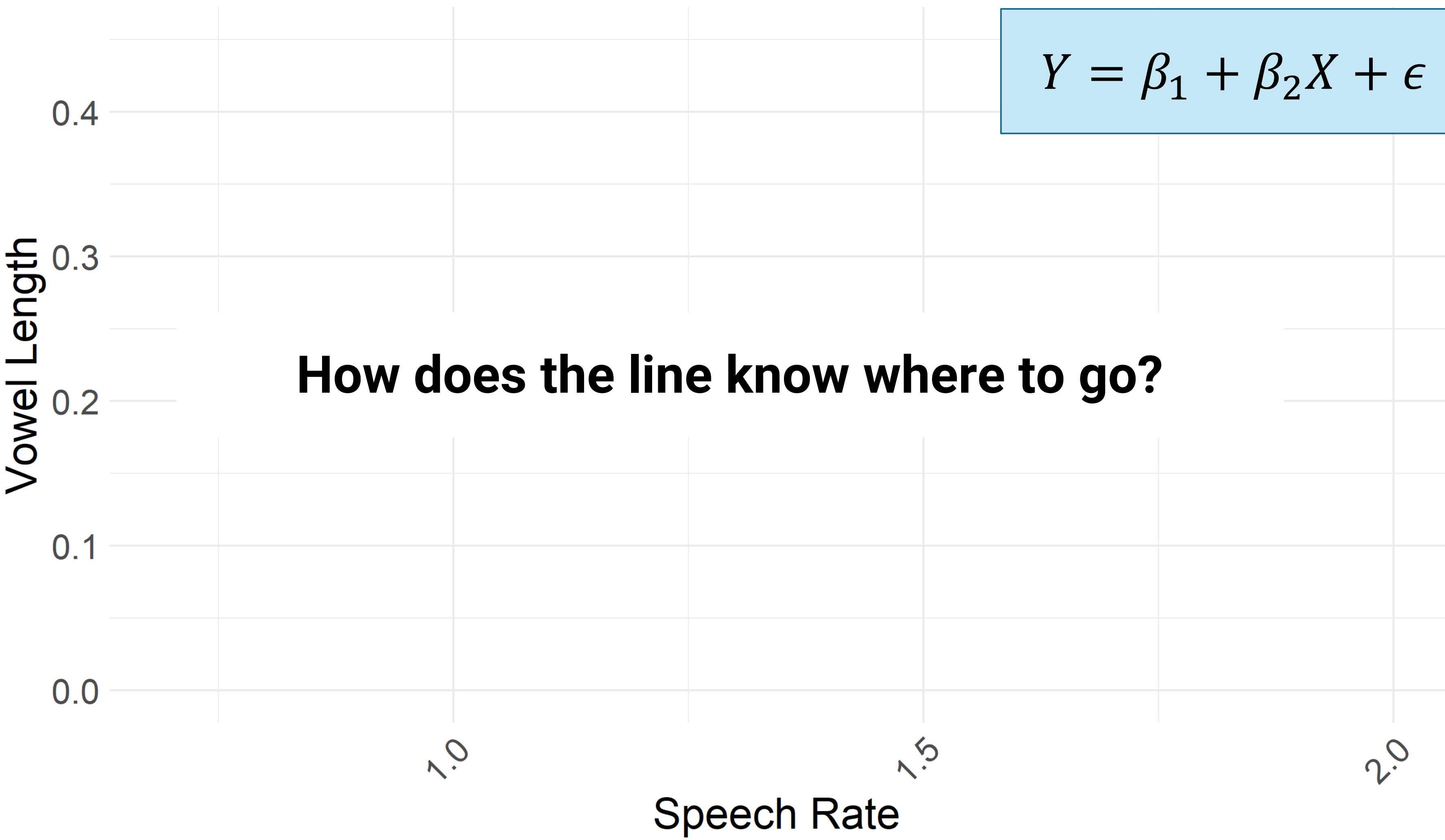
Simple Linear Regression Formula

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



Simple Linear Regression Formula

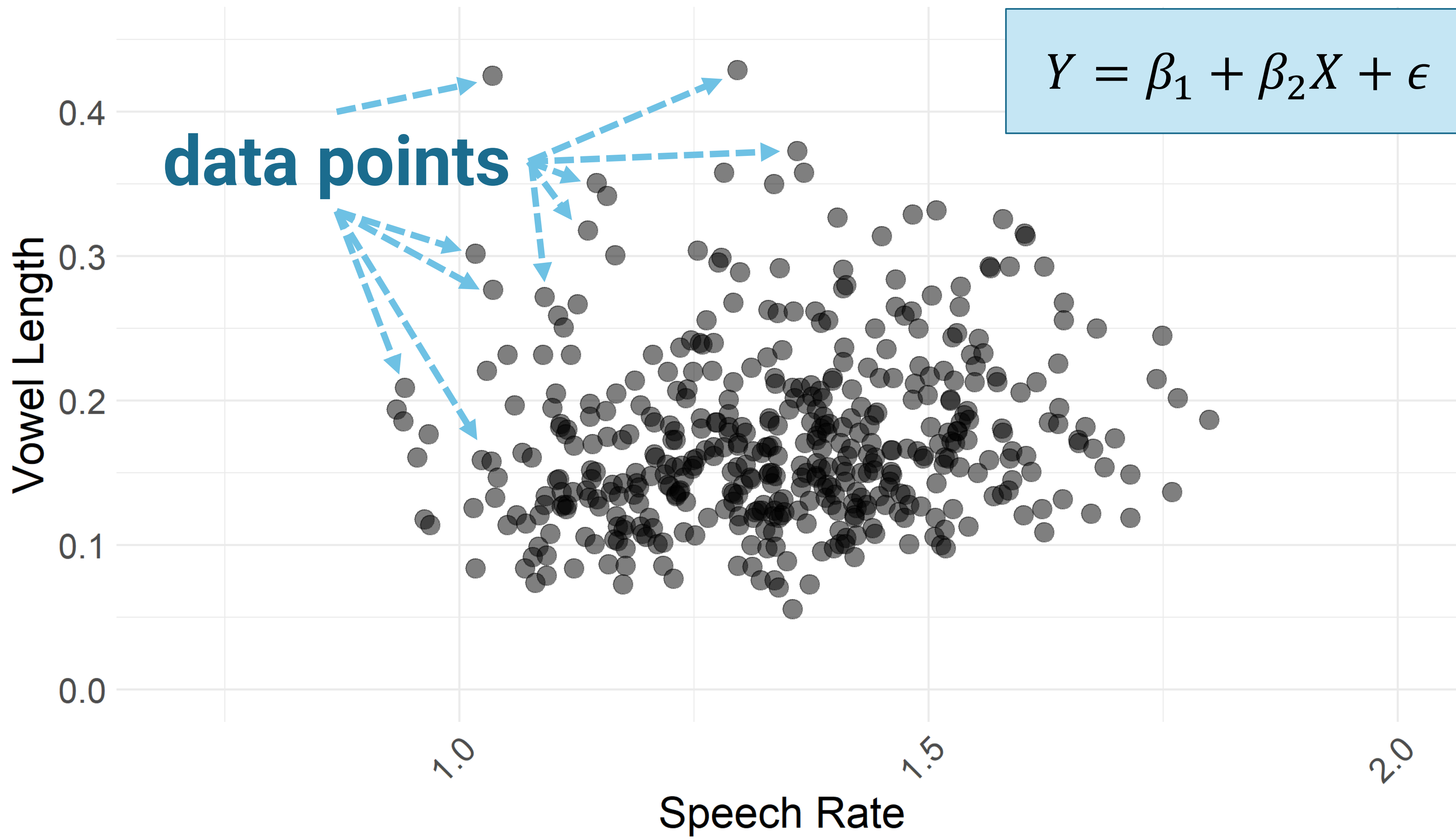
$$Y = \beta_1 + \beta_2 X + \epsilon$$



How does the line know where to go?

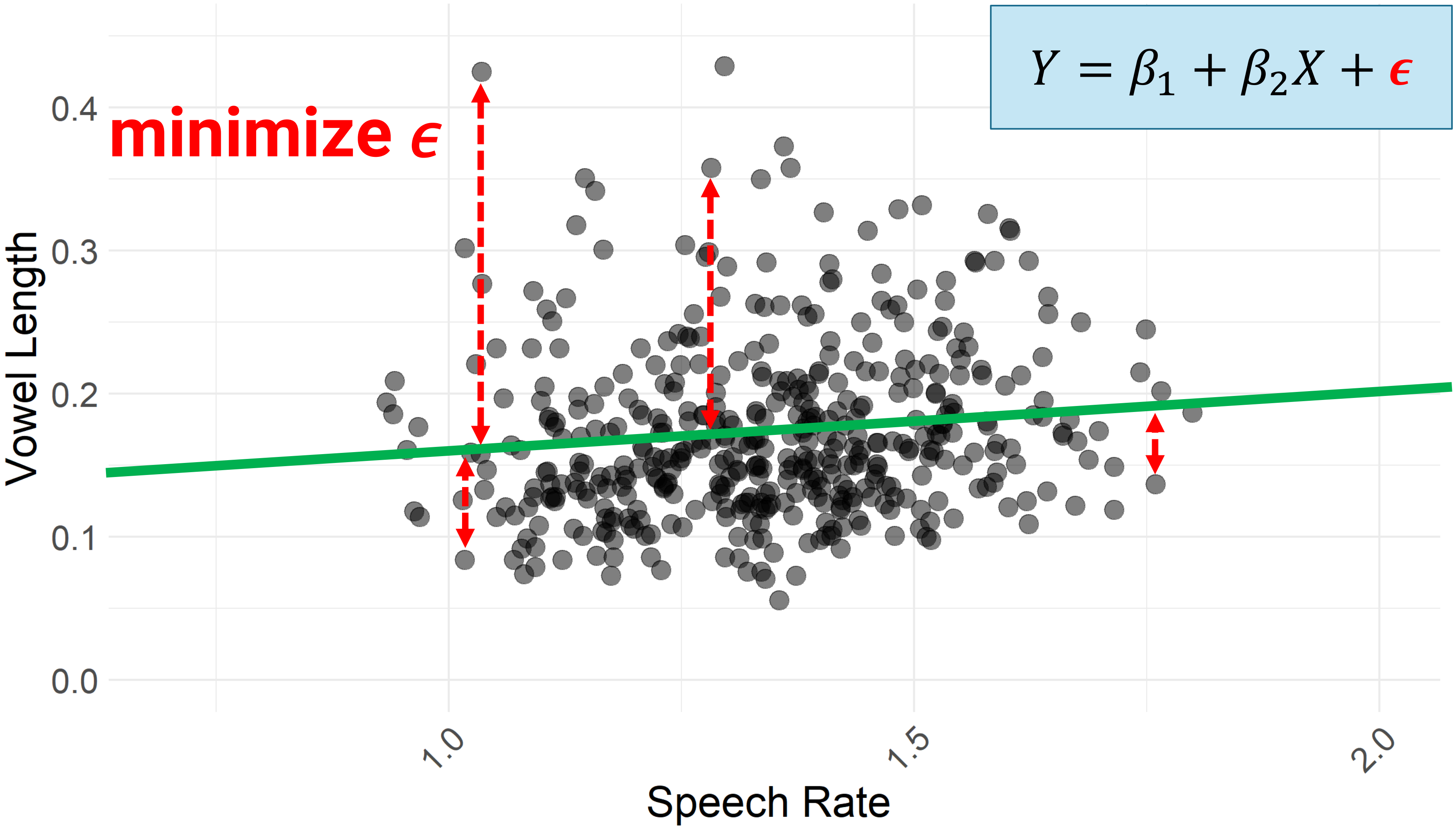
Simple Linear Regression Formula

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



Simple Linear Regression Formula

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



Simple Linear Regression in R

- In R, a simple linear regression model

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

- is created using the following syntax:

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

- Intercept and slope are calculated by R minimizing the residual error between measured data points and estimated regression line

Simple Linear Regression in R

- As an example, we model vowel duration by speech rate

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

- After creating the model, printing it yields the following output:

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

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

Simple Linear Regression in R

- As an example, we model vowel duration by speech rate

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

- After creating the model, printing it yields the following output:

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

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

intercept **slope**

Simple Linear Regression in R

- A *p*-value can be found using the `anova()` function

```
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
```

Simple 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
```

- **Degrees of Freedom**

  The number of independent pieces of information that went into calculating the estimate of said factor.

Simple 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
```

- **Squared Sum**

  The higher the value, the more important the factor is to the model.

Simple 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			

- **Squared Mean**

  The higher the value, the more important the factor is to the model.

Simple 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 Value**

  The higher the value, the more influence the factor has on the dependent variable.

Simple 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
```

- **Probability Value**

  Indicates whether an included factor has a significant influence on the dependent variable.

Simple 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
```

- **Residuals**

  The deviation/error not explained by the independent variables/factors. $\rightarrow \epsilon$

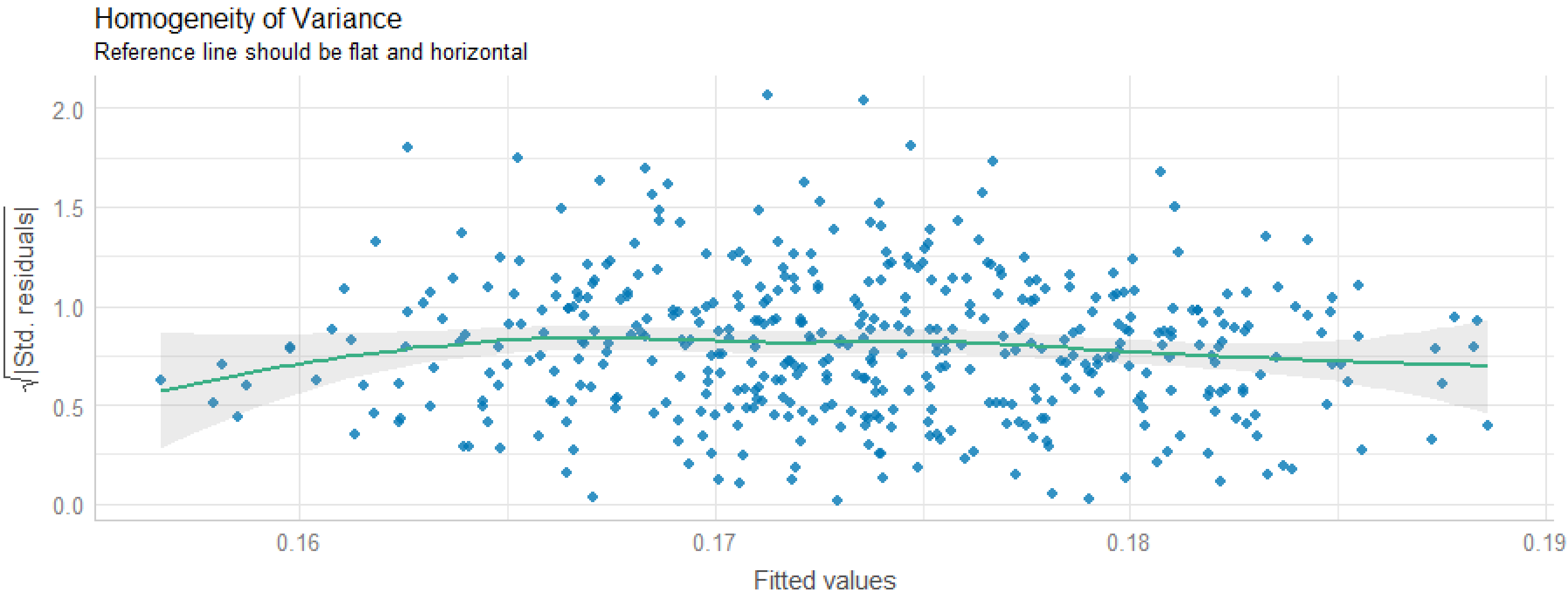
Assumptions

- According to our model, vowel duration decreases significantly with increasing speaking rate
- However, we do not know whether our model relies on valid calculations as we still have to check whether it follows the **assumptions** of a linear regression model
  - Linearity
  - Homoscedasticity
  - Normality
  - Independence

Assumptions: Linearity

- Assumption:

  The relationship between X and the mean of Y is linear.

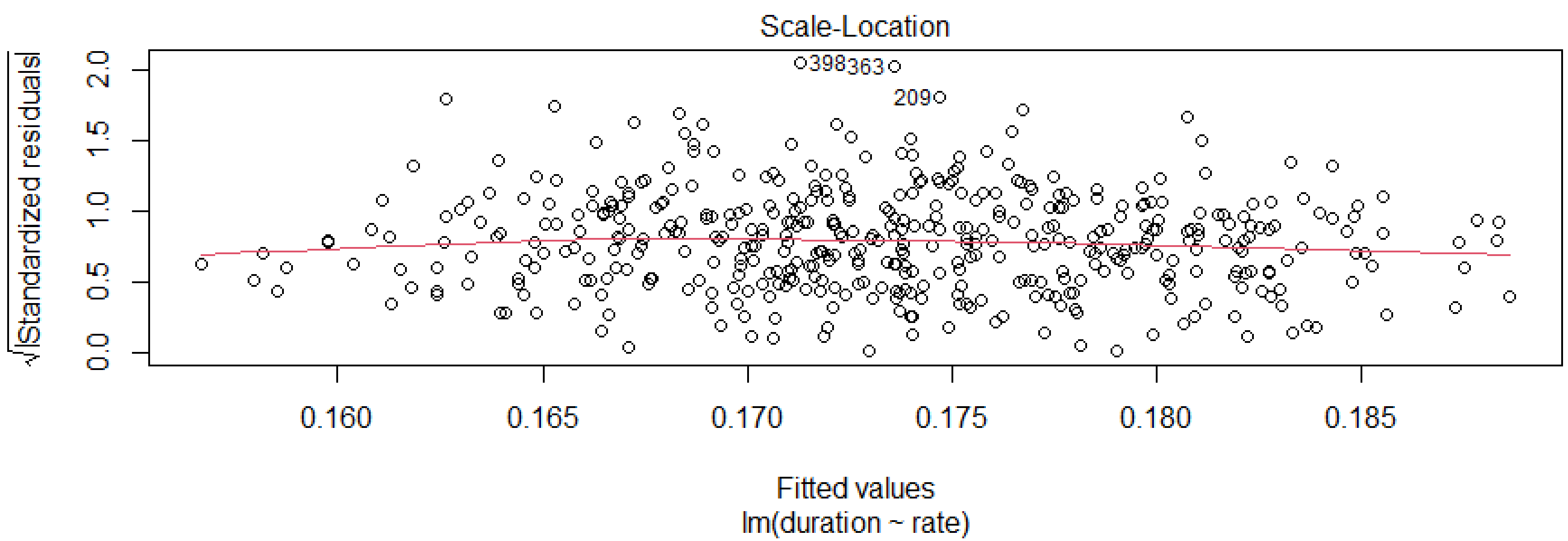


- The line should be horizontal and flat.

Assumptions: Homoscedasticity

- Assumption:

  The variance of residuals is the same for any value of X.



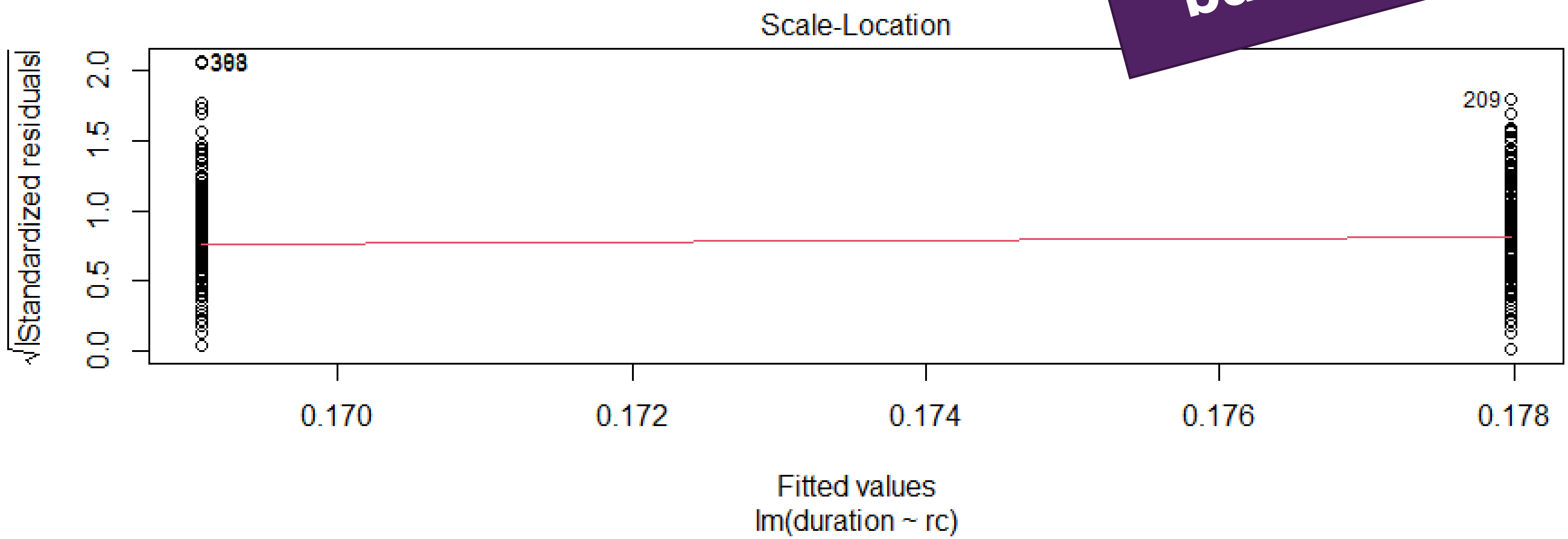
- Data should be spread equally around the line, with no obvious patterns visible.

Assumptions: Homoscedasticity

- Assumption:

  The variance of residuals is the same for any value of X.

bad example!

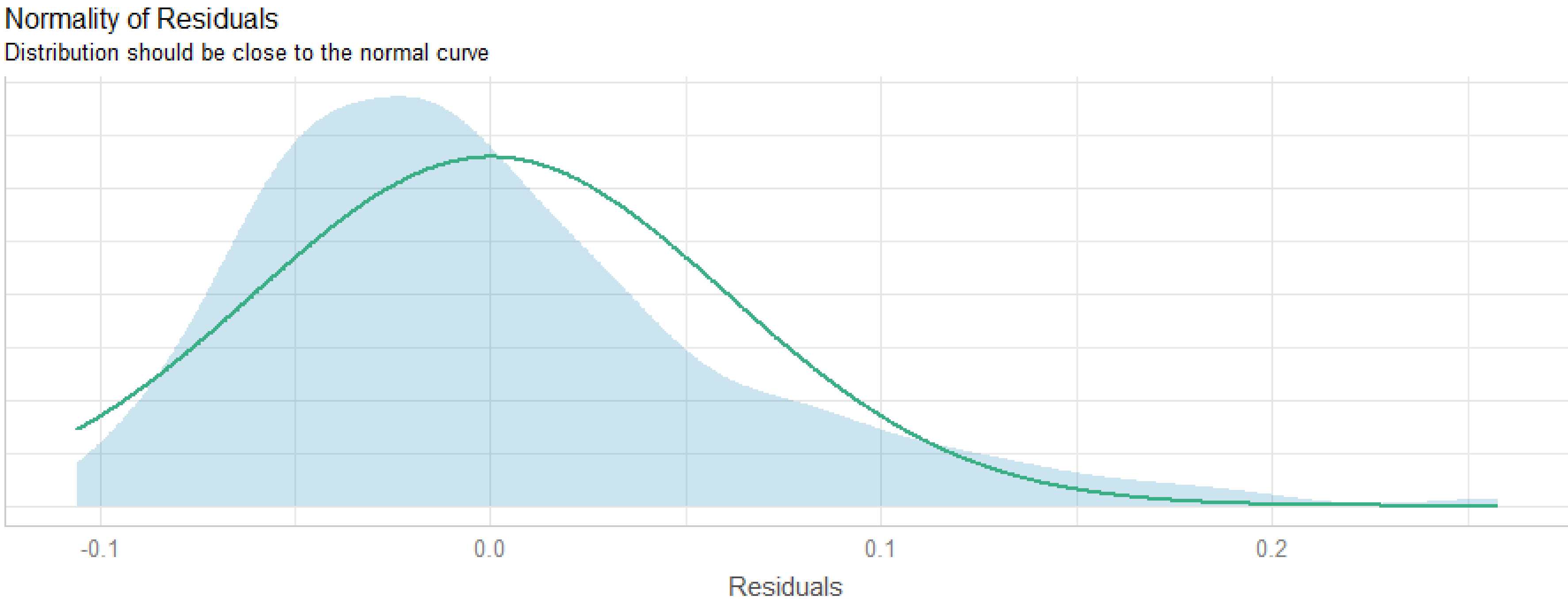


- Data should be spread equally around the line, with no obvious patterns visible.

Assumptions: Normality

- Assumption:

  For any fixed value of X, Y is normally distributed.

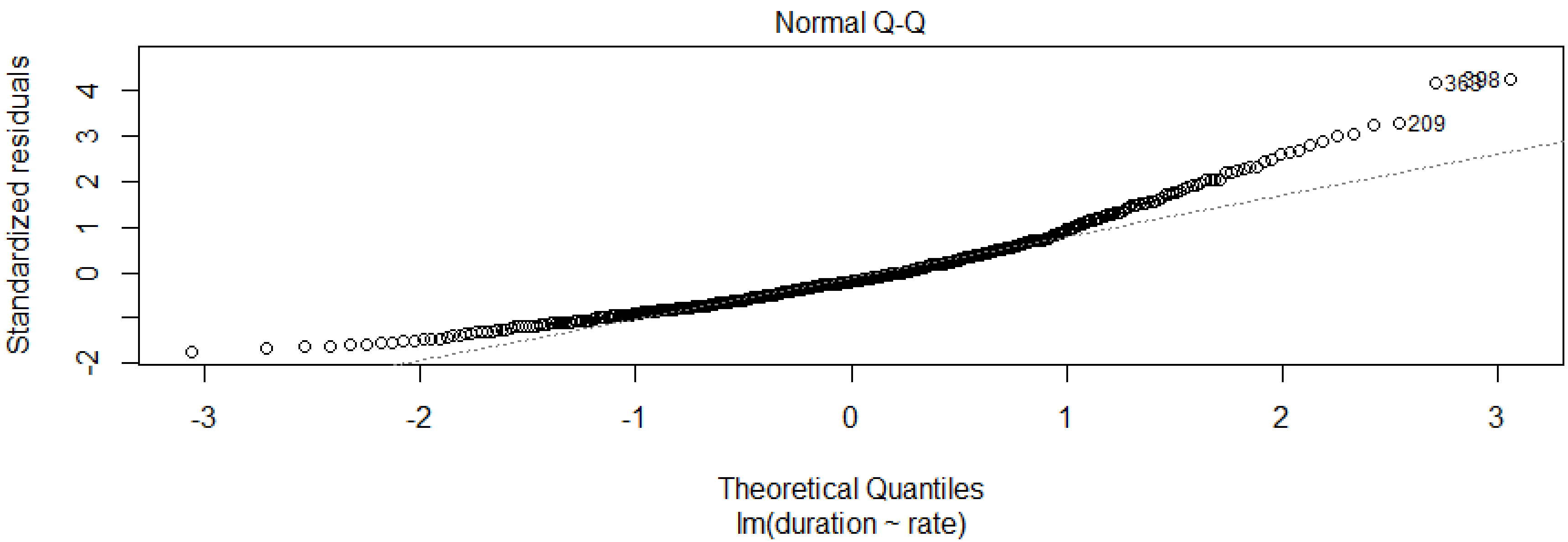


- The distribution of a linear model's residuals should follow a normal distribution.

Assumptions: Normality

- Assumption:

  For any fixed value of X, Y is normally distributed.



- Residual points should follow the line.

Assumptions: Independence

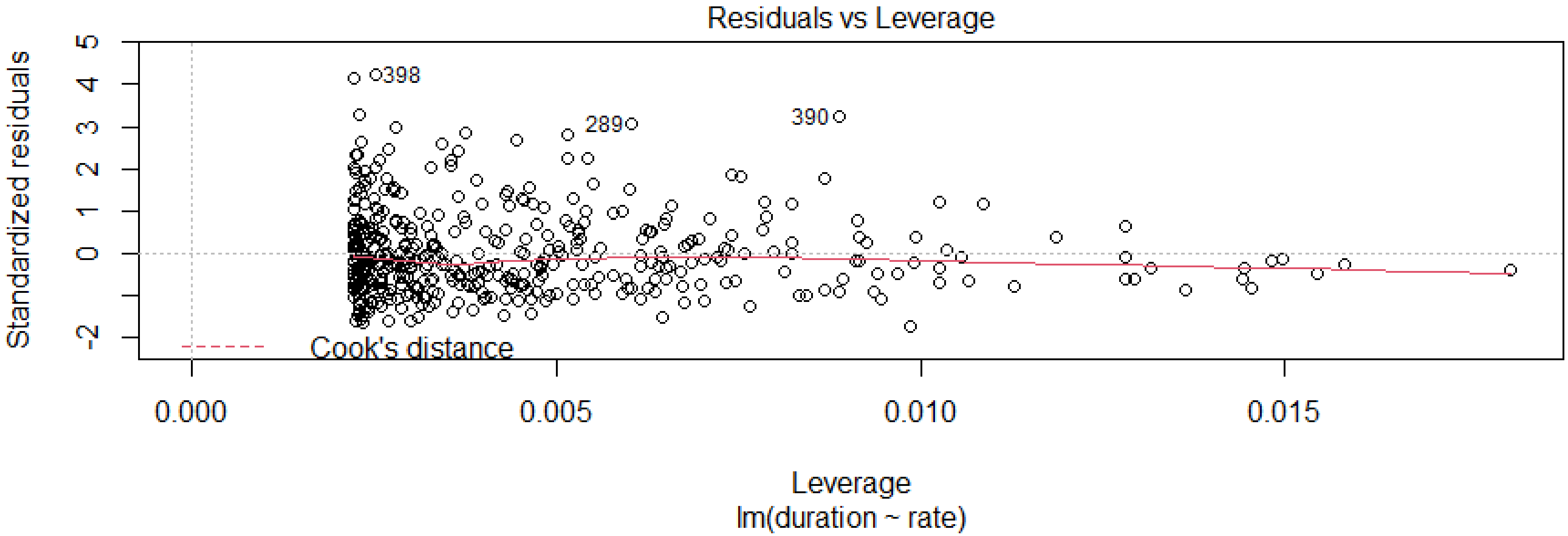
- Assumption:

  Observations are independent of each other.

- Independence cannot be checked visually
- It is an assumption that you can test by examining the study design

Extra: Influential Data Points

- Cook‘s Distance:
  - A measure of the influence of each observation on the regression coefficients
  - Any observation for which the Cook's distance is close to 1, or that is substantially larger than other Cook's distances requires investigation.



Dependent Variable Distribution Check

- Results of linear regression are more reliable for dependent variables following the normal distribution
- Thus, one should check the dependent variable‘s distribution before running models
- In case the dependent variable is not normally distributed, data transformation may be advisable
- However, in the rare case that no transformation technique brings the dependent variable closer to a normal distribution, linear regression can still be used

Dependent Variable Distribution Check

- You can check whether a variable is normally distributed using a Shapiro-Wilk Test
- Here, higher p-values indicate a normal distribution

```
shapiro.test(data$duration)

      Shapiro-Wilk normality test

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

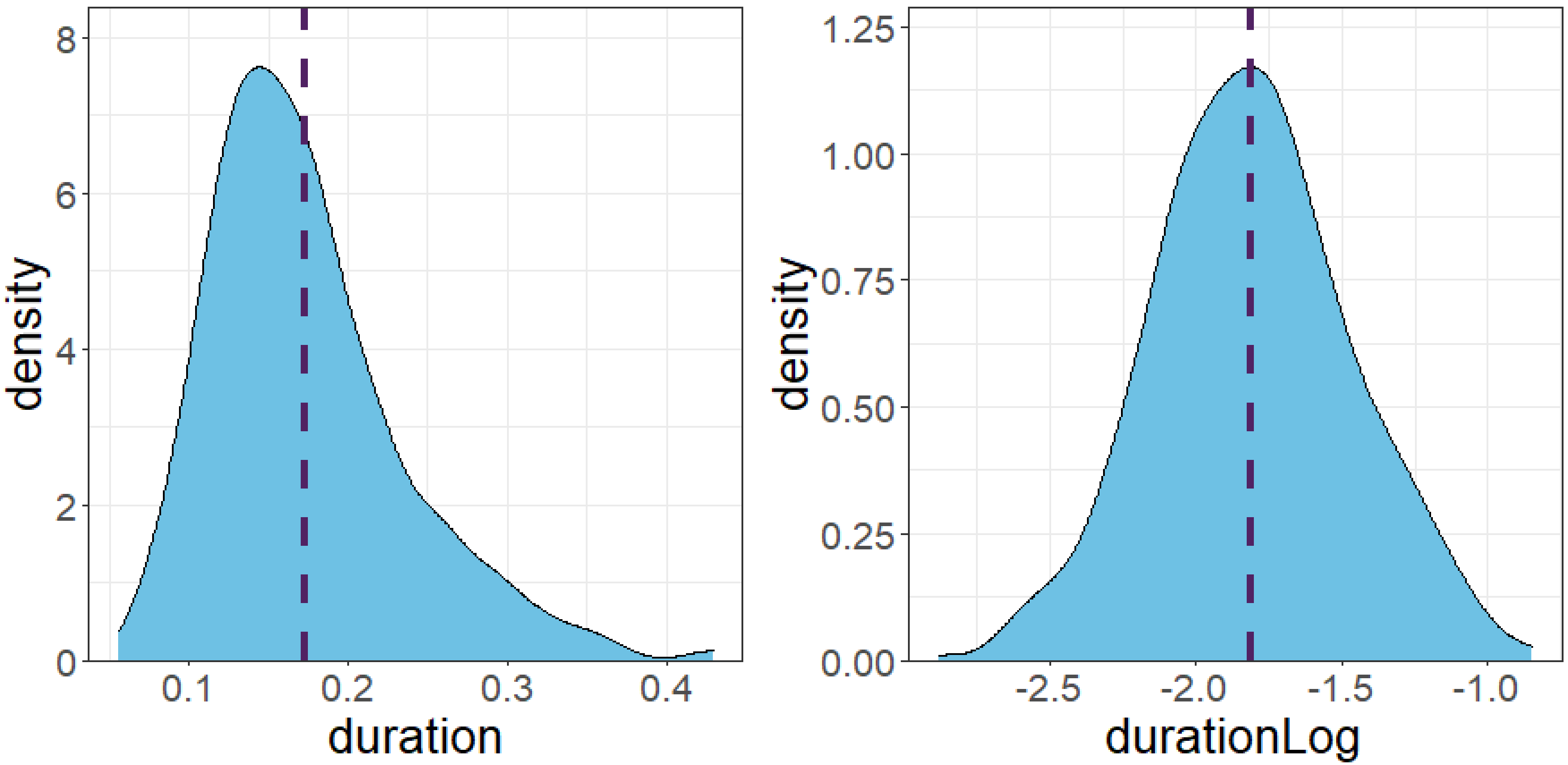
Dependent Variable Distribution Check

- As `duration` is no way near a normal distribution, we create a log-transformed version

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

shapiro.test(data$durationLog)

      Shapiro-Wilk normality test

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

Dependent Variable Distribution Check

- Visual inspection clearly shows that the newly created variable is closer to a normal distribution



Dependent Variable Distribution Check

- If we now redo our previous model, using the log-transformed dependent variable, we find that it fulfils the normality of residuals assumption much better