Simple Linear Regression

Partitioning variability

Prof. Maria Tackett

STA 210

[Click here for PDF of slides]

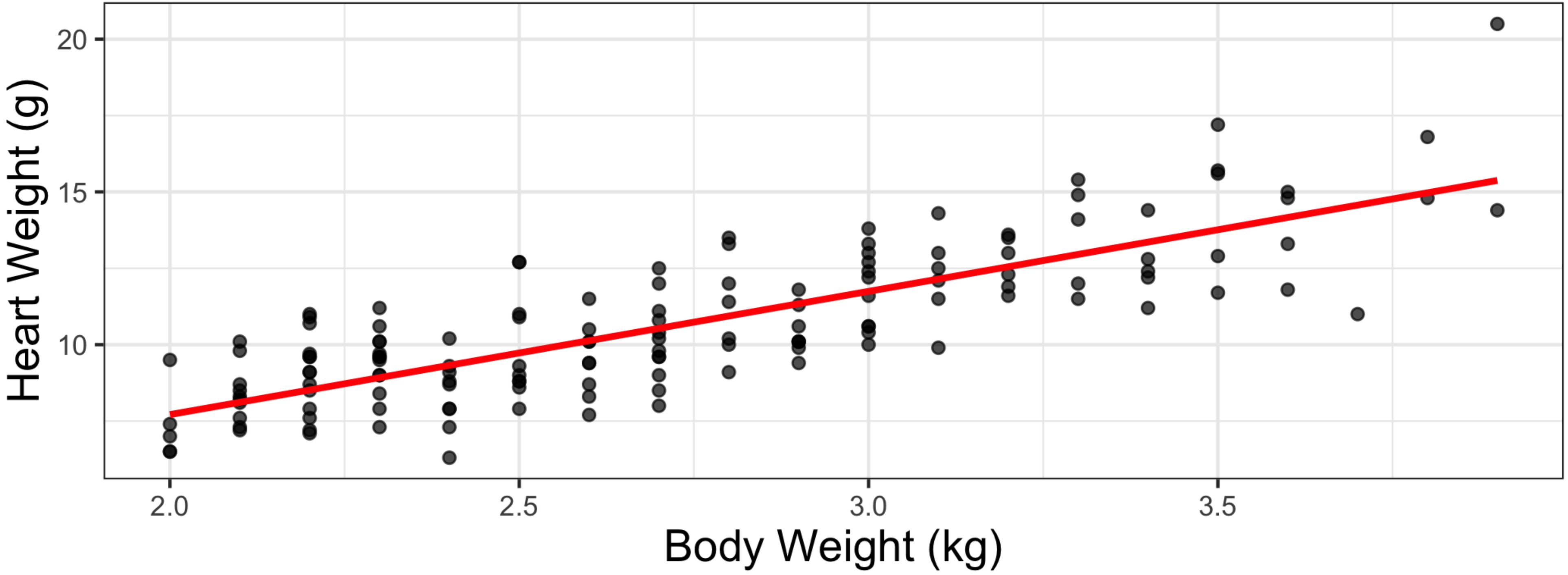
Topics

- Use analysis of variance to partition variability in the response variable
- Define and calculate R^2
- Use ANOVA to test the hypothesis

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

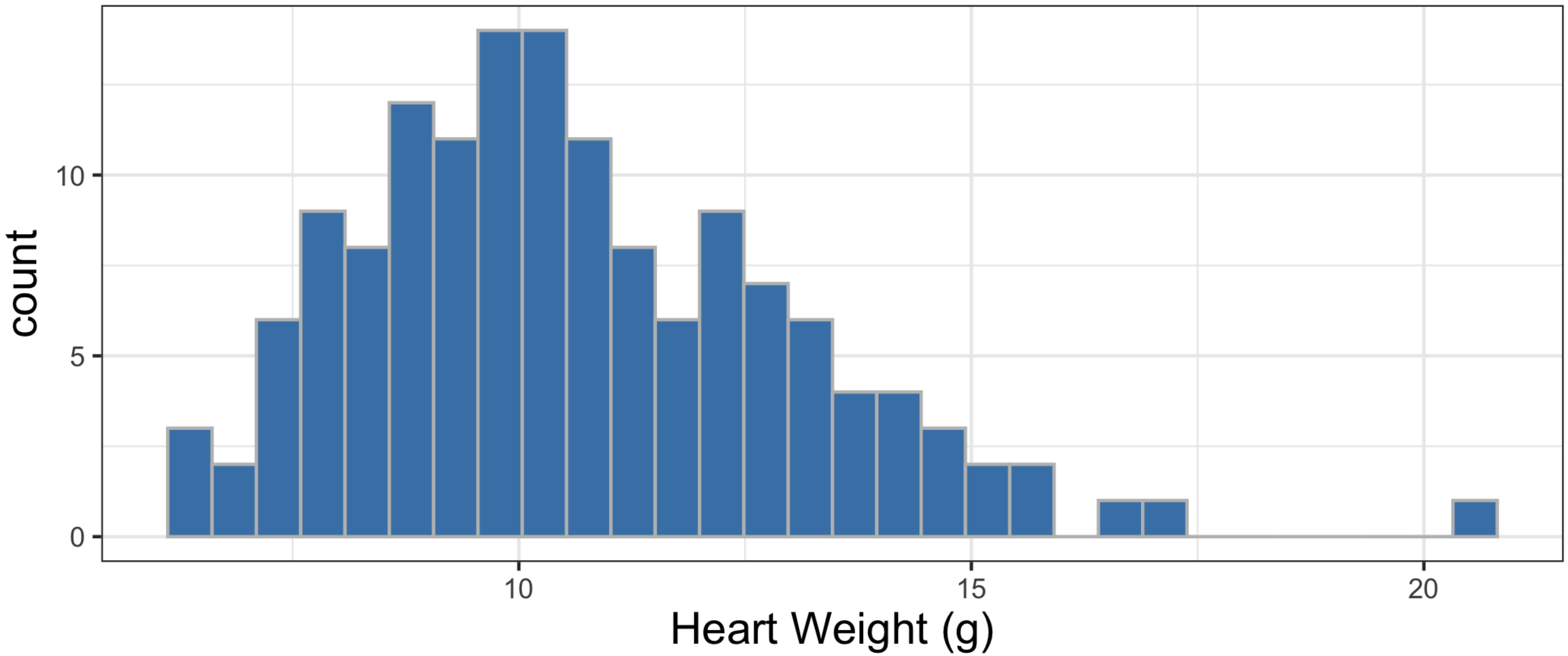
Cats data

The data set contains the **heart weight** (`Hwt`) and **body weight** (`Bwt`) for 144 domestic cats.



STA 210

Distribution of response

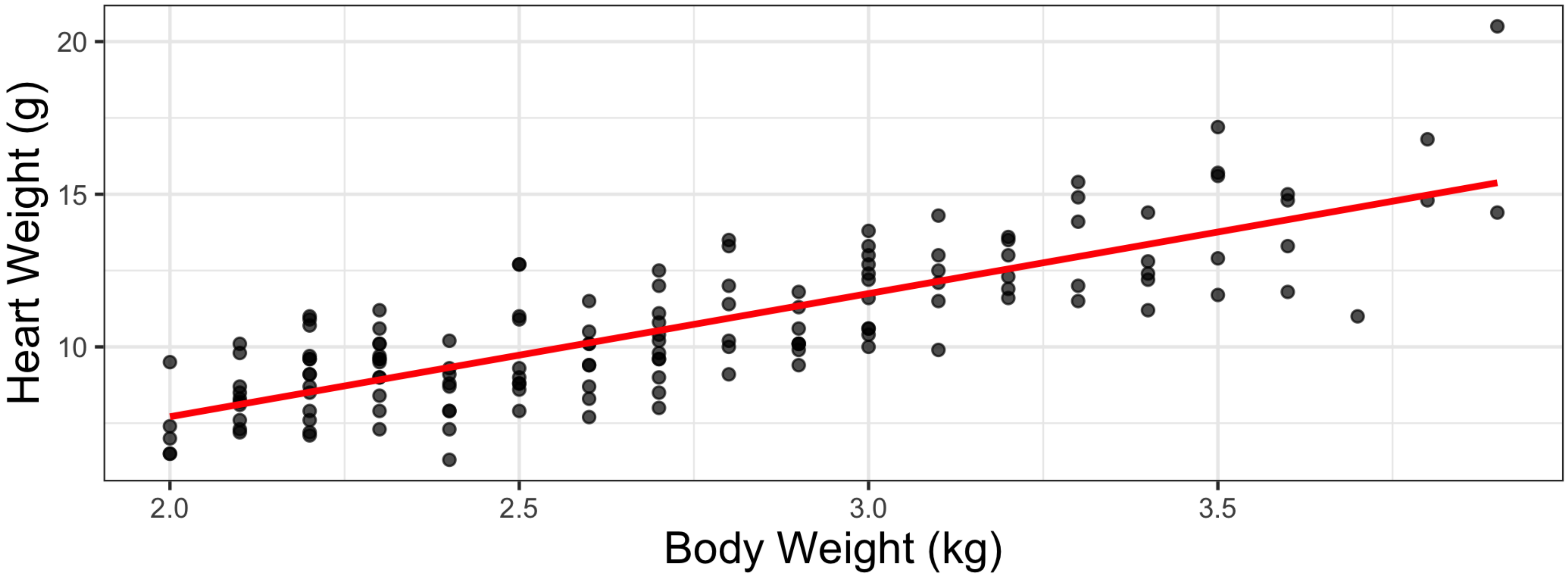


Mean	Std. Dev.	IQR
10.631	2.435	3.175

STA 210

The model

$$\widehat{\text{Hwt}} = -0.357 + 4.034 \times \text{Bwt}$$



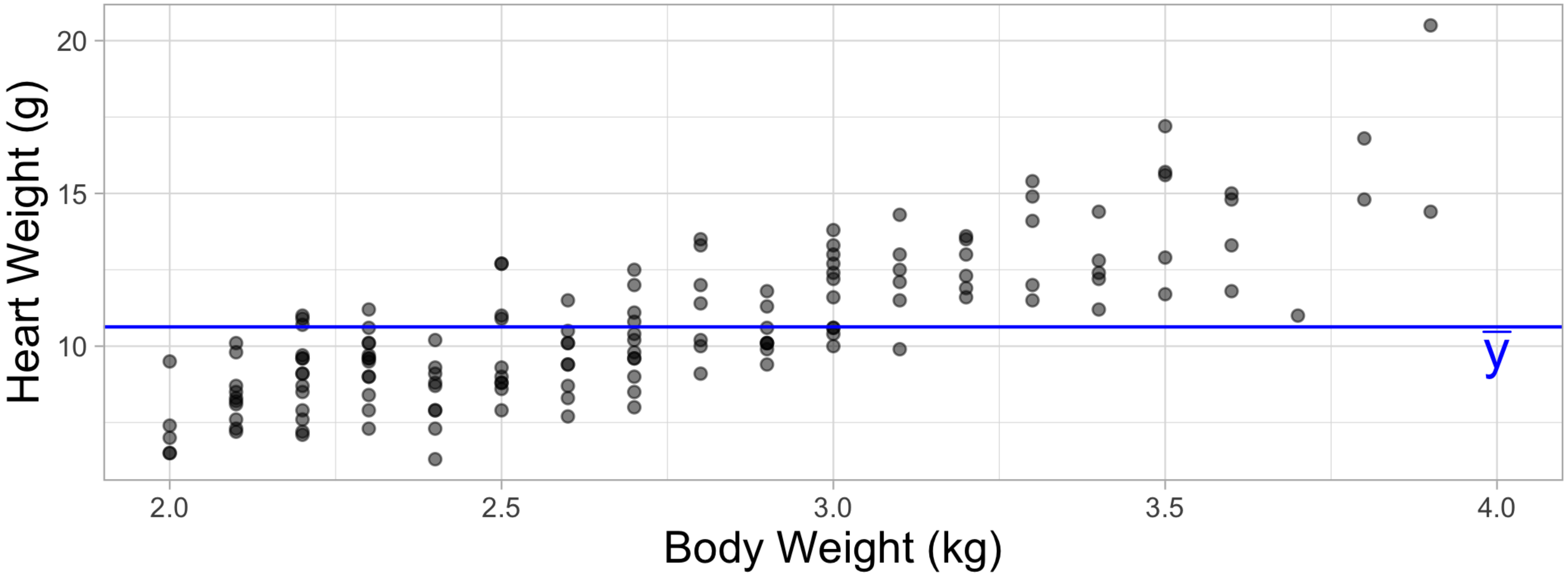
How much of the variation in cats' heart weights can be explained by knowing their body weights?

ANOVA

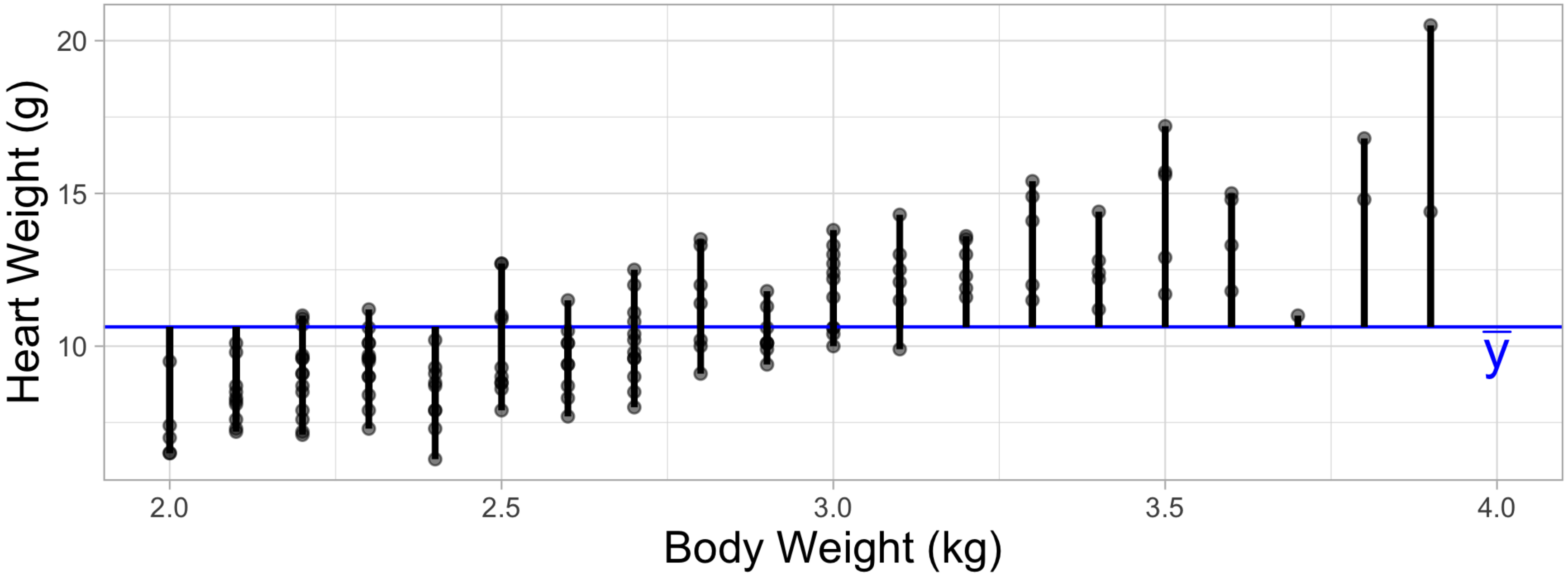
We will use **Analysis of Variance (ANOVA)** to partition the variation in the response variable Y.

Total variation in Y = Explained variation (Model) + Unexplained variation (Residuals)

Response variable, Y



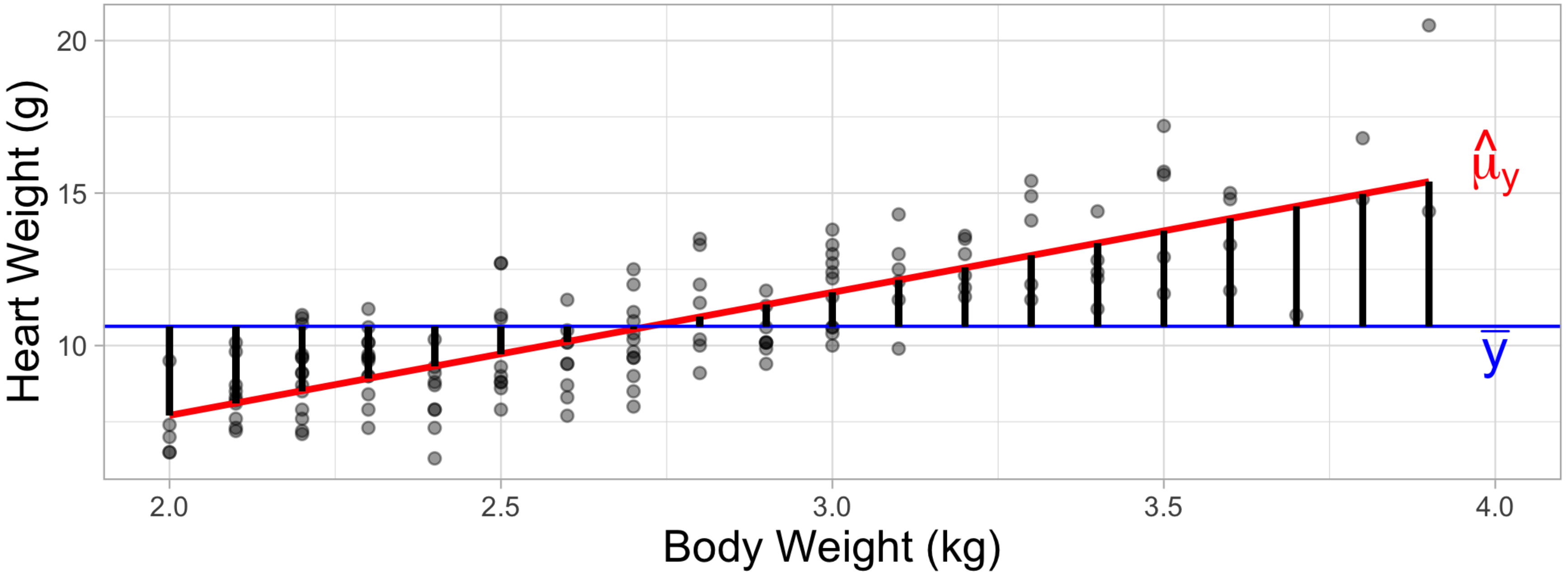
Total variation



$$SS_{Total} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = (n-1)s_y^2$$

STA 210

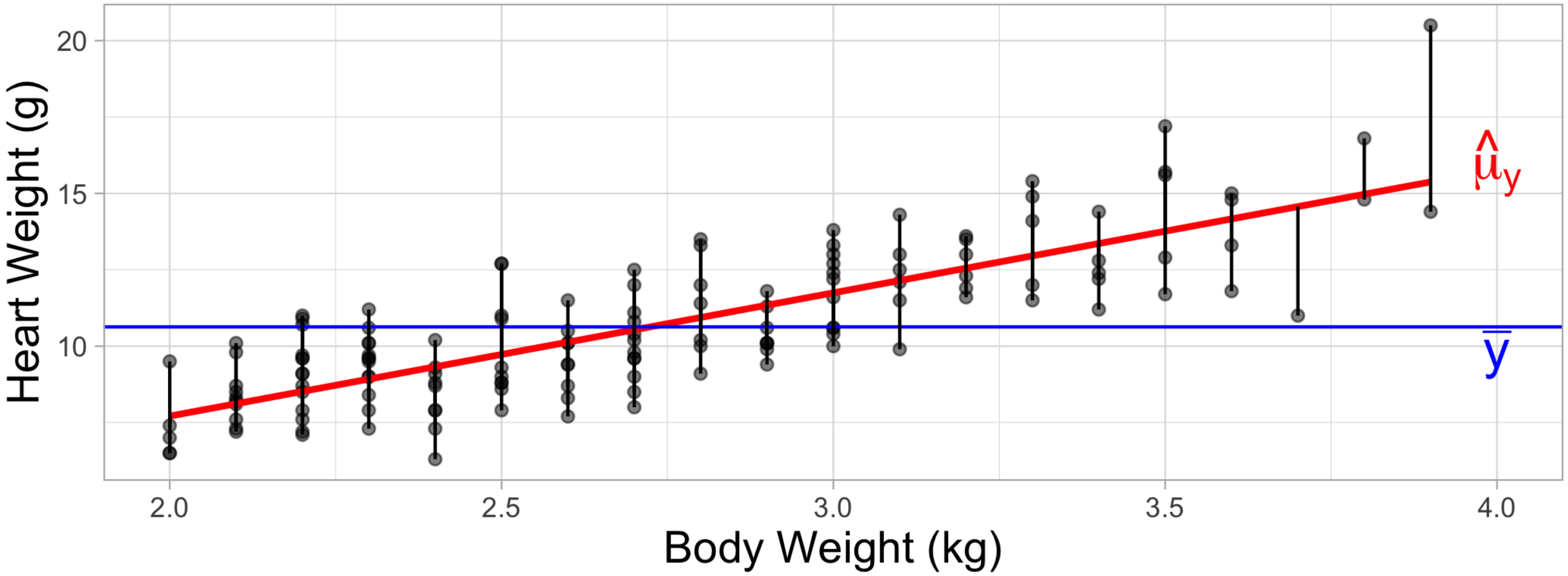
Explained variation (Model)



$$SS_{Model} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

STA 210

Unexplained variation (Residuals)



$$SS_{Error} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

STA 210

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2} + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

STA 210

R^2

The **coefficient of determination, R^2**, is the proportion of variation in the response, Y, that is explained by the regression model

$$R^2 = \frac{SS_{Model}}{SS_{Total}} = 1 - \frac{SS_{Error}}{SS_{Total}}$$

STA 210

R^2 for our model

$$SS_{Model} = 548.092$$

$$SS_{Error} = 299.533$$

$$SS_{Total} = 847.625$$

$$R^2 = \frac{548.092}{847.625}$$

$$= \mathbf{0.647}$$

About 64.7% of the variation in the heart weight of cats can be explained by variation in body weight.

STA 210

ANOVA table

Source	Df	Sum Sq	Mean Sq	F Stat	Pr(> F)
Model	1	548.092	548.092	259.835	0
Residuals	142	299.533	2.109		
Total	143	847.625			

STA 210

ANOVA table

Source	Df	Sum Sq	Mean Sq	F Stat	Pr(> F)
Model	1	548.092	548.092	259.835	0
Residuals	142	299.533	2.109		
Total	143	847.625			

Sum of squares

$$SS_{Total} = 847.625 = 548.092 + 299.533$$

$$SS_{Model} = 548.092$$

$$SS_{Error} = 299.533$$

ANOVA Test

Source	Df	Sum Sq	Mean Sq	F Stat	Pr(> F)
Model	1	548.092	548.092	259.835	0
Residuals	142	299.533	2.109		
Total	143	847.625			

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

STA 210

ANOVA Test

Source	Df	Sum Sq	Mean Sq	F Stat	Pr(> F)
Model	1	548.092	548.092	259.835	0
Residuals	142	299.533	2.109		
Total	143	847.625			

Degrees of freedom

$$df_{Total} = 144 - 1 = 143$$

$$df_{Model} = 1$$

$$df_{Error} = 143 - 1 = 142$$

STA 210

ANOVA Test

Source	Df	Sum Sq	Mean Sq	F Stat	Pr(> F)
Model	1	548.092	548.092	259.835	0
Residuals	142	299.533	2.109		
Total	143	847.625			

Mean squares

$$MS_{Model} = \frac{548.092}{1} = 548.092$$

$$MS_{Error} = \frac{299.533}{142} = 2.109$$

STA 210

ANOVA Test

Source	Df	Sum Sq	Mean Sq	F Stat	Pr(> F)
Model	1	548.092	548.092	259.835	0
Residuals	142	299.533	2.109		
Total	143	847.625			

F test statistic: ratio of explained to unexplained variability

$$F = \frac{MS_{Model}}{MS_{Error}} = \frac{548.092}{2.109} = 259.835$$

STA 210

F distribution

ANOVA test

Source	Df	Sum Sq	Mean Sq	F Stat	Pr(> F)
Model	1	548.092	548.092	259.835	0
Residuals	142	299.533	2.109		
Total	143	847.625			

P-value: Probability of observing a test statistic at least as extreme as *F Stat* given the population slope β_1 is 0

The p-value is calculated using an F distribution with 1 and $n-2$ degrees of freedom

STA 210

Calculating p-value

ANOVA

Source	Df	Sum Sq	Mean Sq	F Stat	Pr(> F)
Model	1	548.092	548.092	259.835	0
Residuals	142	299.533	2.109		
Total	143	847.625			

The p-value is very small (≈ 0), so we reject H_0.

The data provide strong evidence that population slope, β_1, is different from 0.

The data provide sufficient evidence that there is a linear relationship between a cat's heart weight and body weight.

STA 210

Recap

- Used analysis of variance to partition variability in the response variable

- Defined and calculated R^2

- Used ANOVA to test the hypothesis

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$