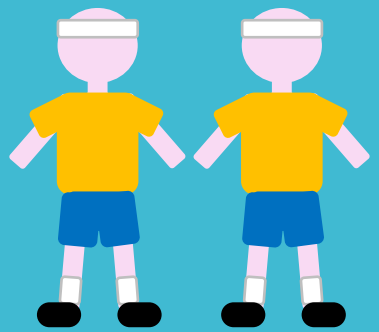
Linear regression: Part 1



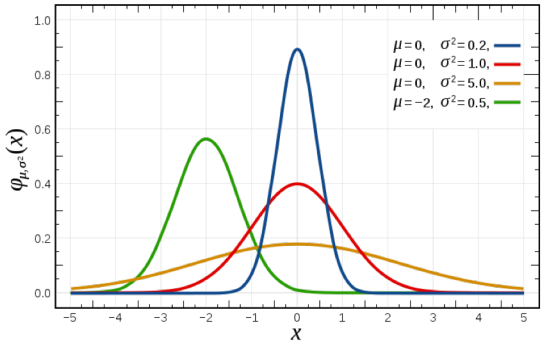
Recap

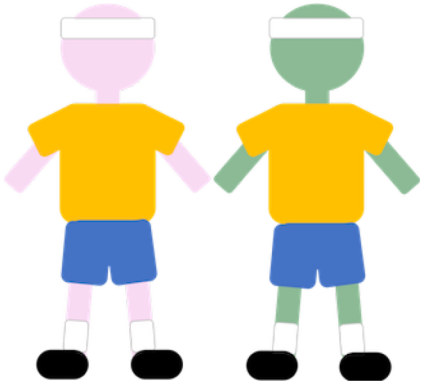
Normal distribution

Maximum likelihood estimation

Ground hogs

Zombies





Lecture Outline

What are linear models?

What is linear regression?

How to find the best line?

Maximum likelihood and regression

Lecture Outline

What are linear models?

What is linear regression?

- EX1: Why and when for regression
- EX2: What is a best line?

How to find the best line?

- EX3: Trying fitting a line 1

Maximum likelihood and regression

- EX4: Trying fitting a line in R

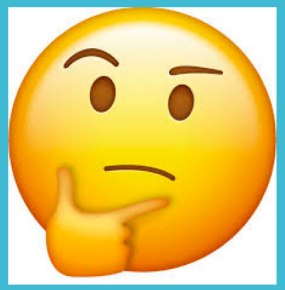
Reading

Chapter 4 – The New Statistics with R

Chapter 17 – The Analysis of Biological Data (2nd Edition)

Chapter 1.6 – Generalized linear models with examples in R

What are linear models?



Definition

Linear models:

Models with a **continuous response** variable as a function of one or more **explanatory** variable. Variables are connected by **linear equations**.

Definition

Linear models:

Models with a **continuous response** variable as a function of one or more **explanatory** variable. Variables are connected by **linear equations**.

We want to explain variable Y with variable X.

Definition

Linear models:

Models with a **continuous response** variable as a function of one or more **explanatory** variable. Variables are connected by **linear equations**.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Definition

Linear models:

Models with a continuous **response** variable as a function of one or more **explanatory** variable. Variables are connected by **linear equations**.

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

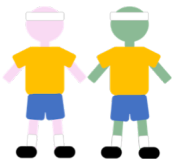
parameters

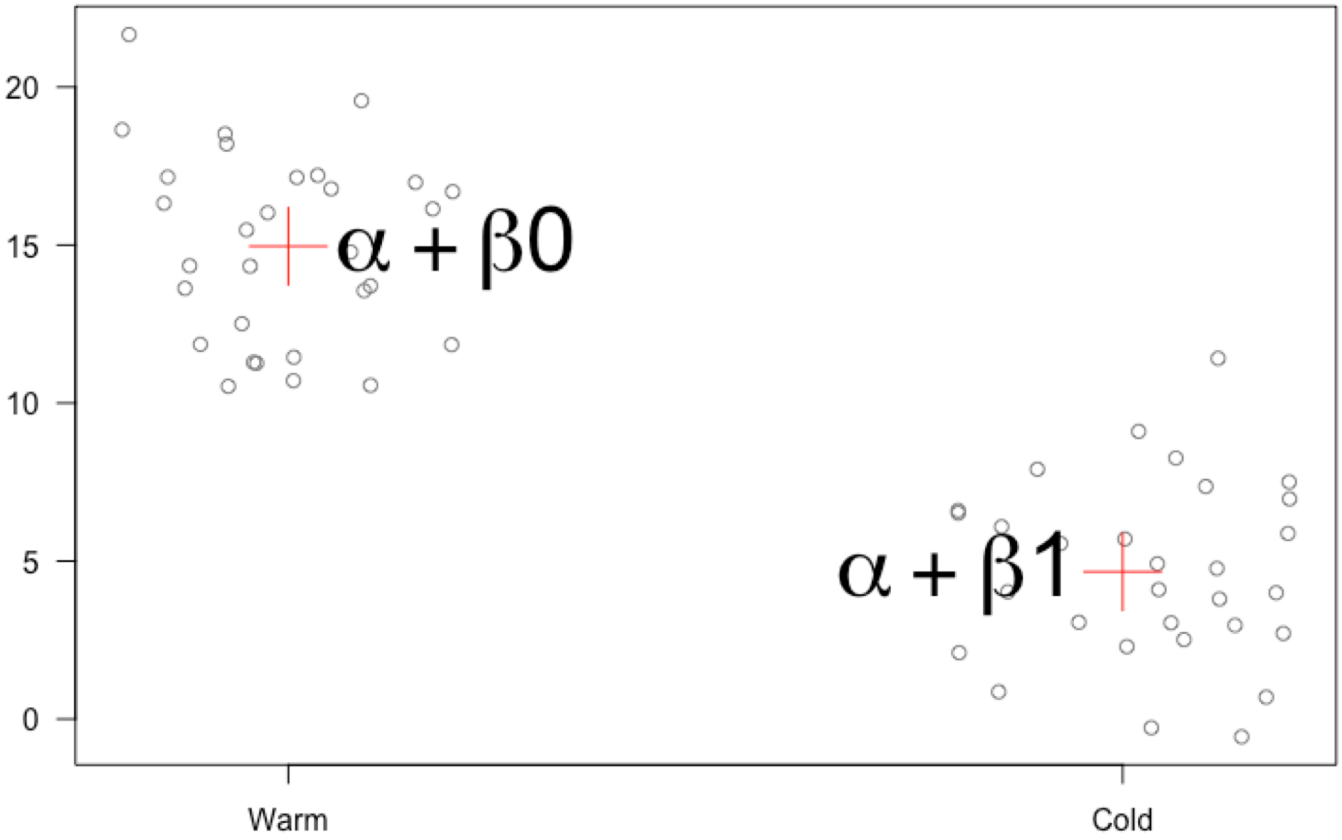
error

Examples of linear models

t-test (last week)

$$\mu_i = \alpha + \beta X_i$$



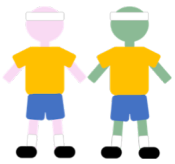


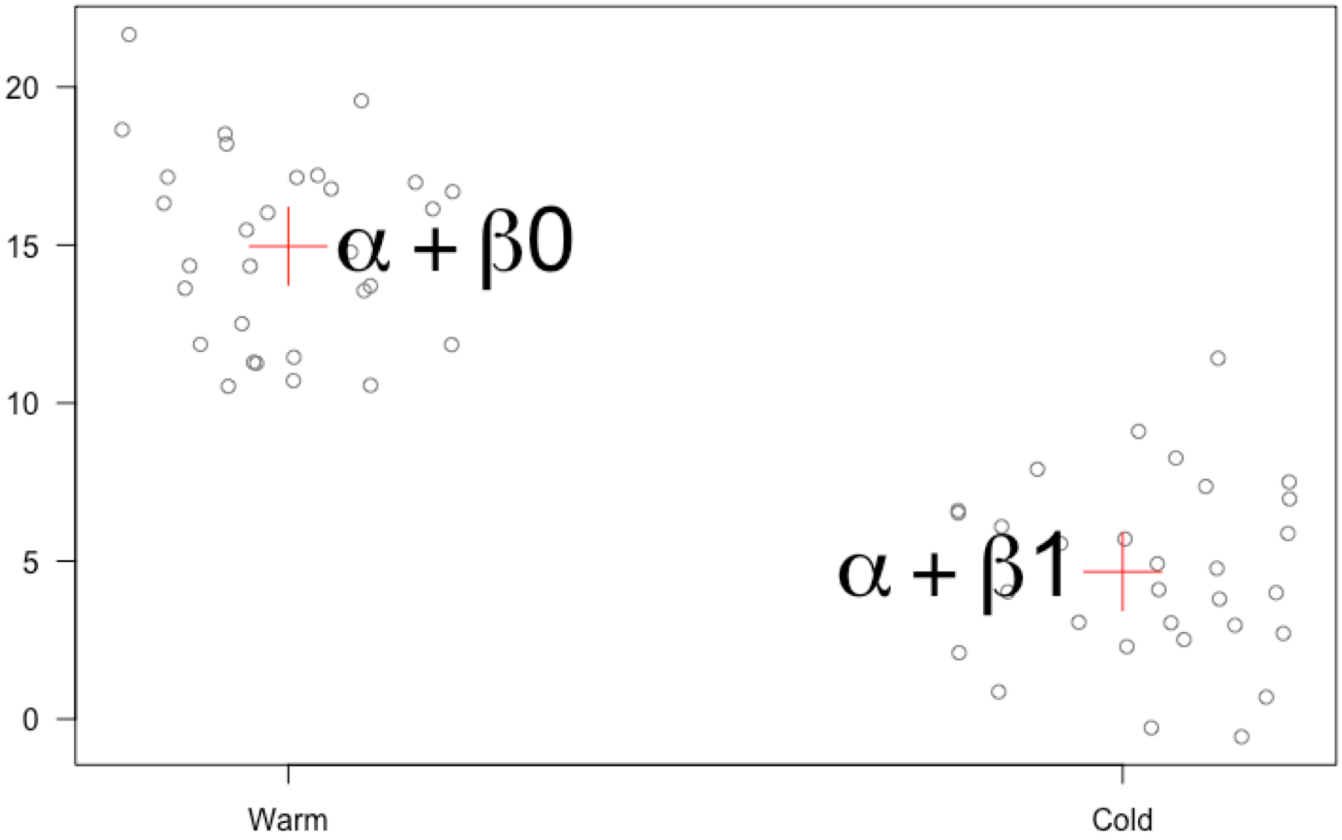
Examples of linear models

t-test (last week)

$$\mu_i = \alpha + \beta X_i$$

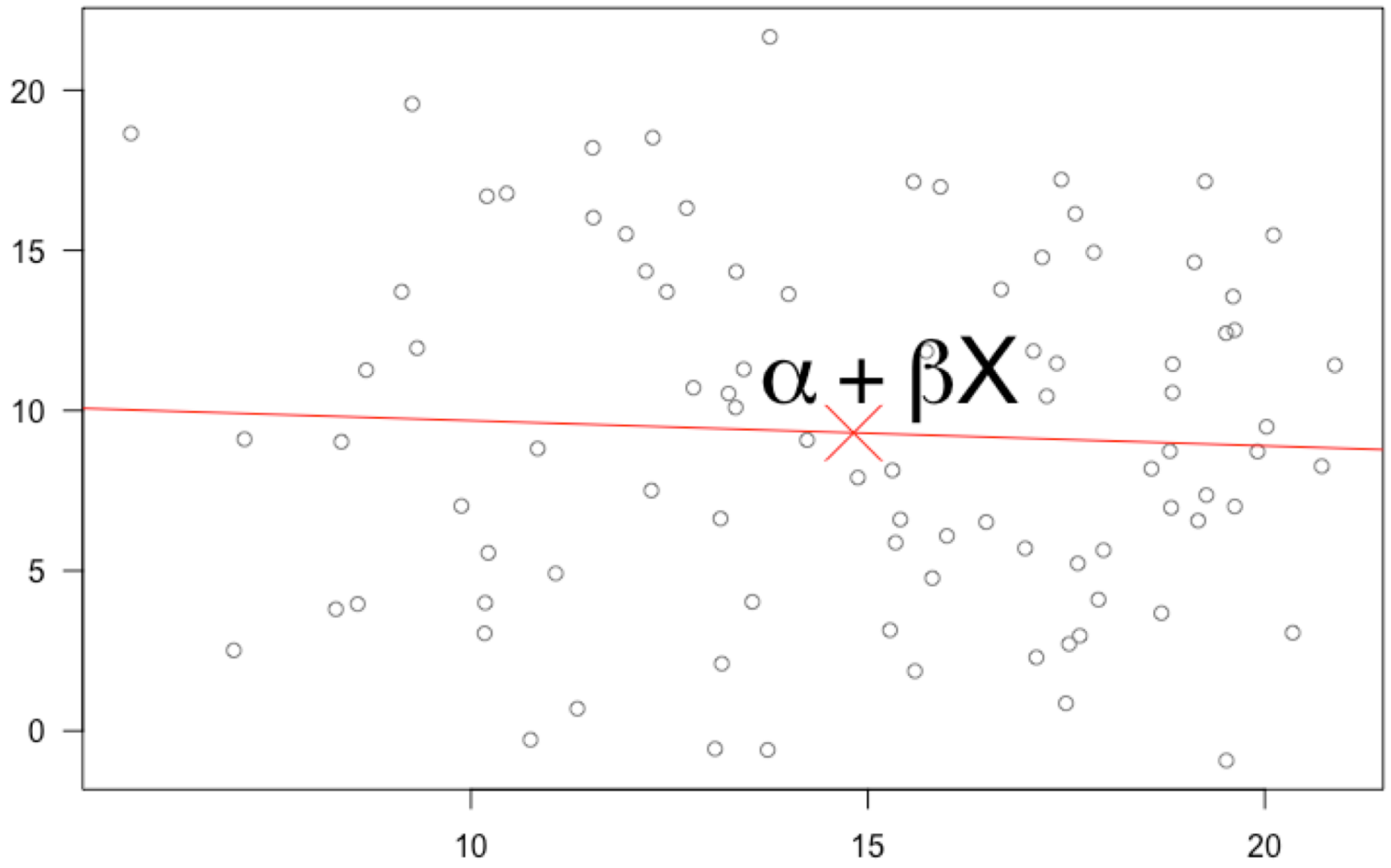
μ_i = mean of group i





Examples of linear models

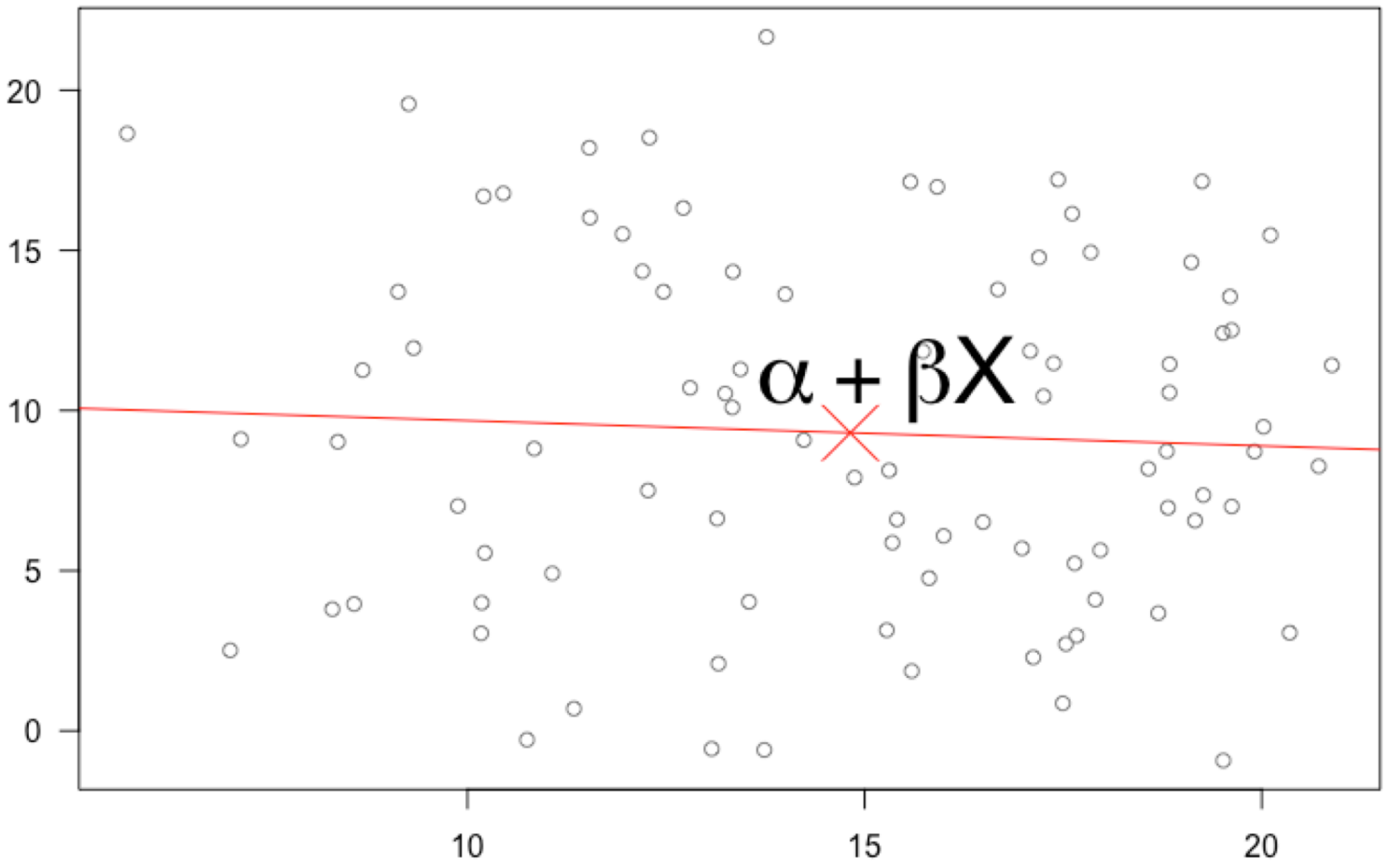
Regression (this week) $\hat{Y}_i = \alpha + \beta X_i$



Examples of linear models

Regression (this week) $\hat{Y}_i = \alpha + \beta X_i$

$\hat{Y}_i$ = estimated y value



Summary linear models

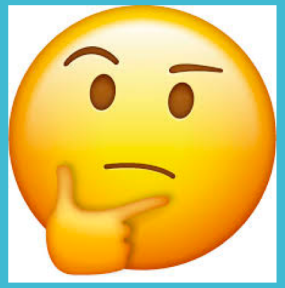
Includes **all:** t-test, anova, and regression

Use the same mathematics and equation (model)

Interpretation will change based on **type of variables**

What we will do for the rest of the course

What is linear regression?

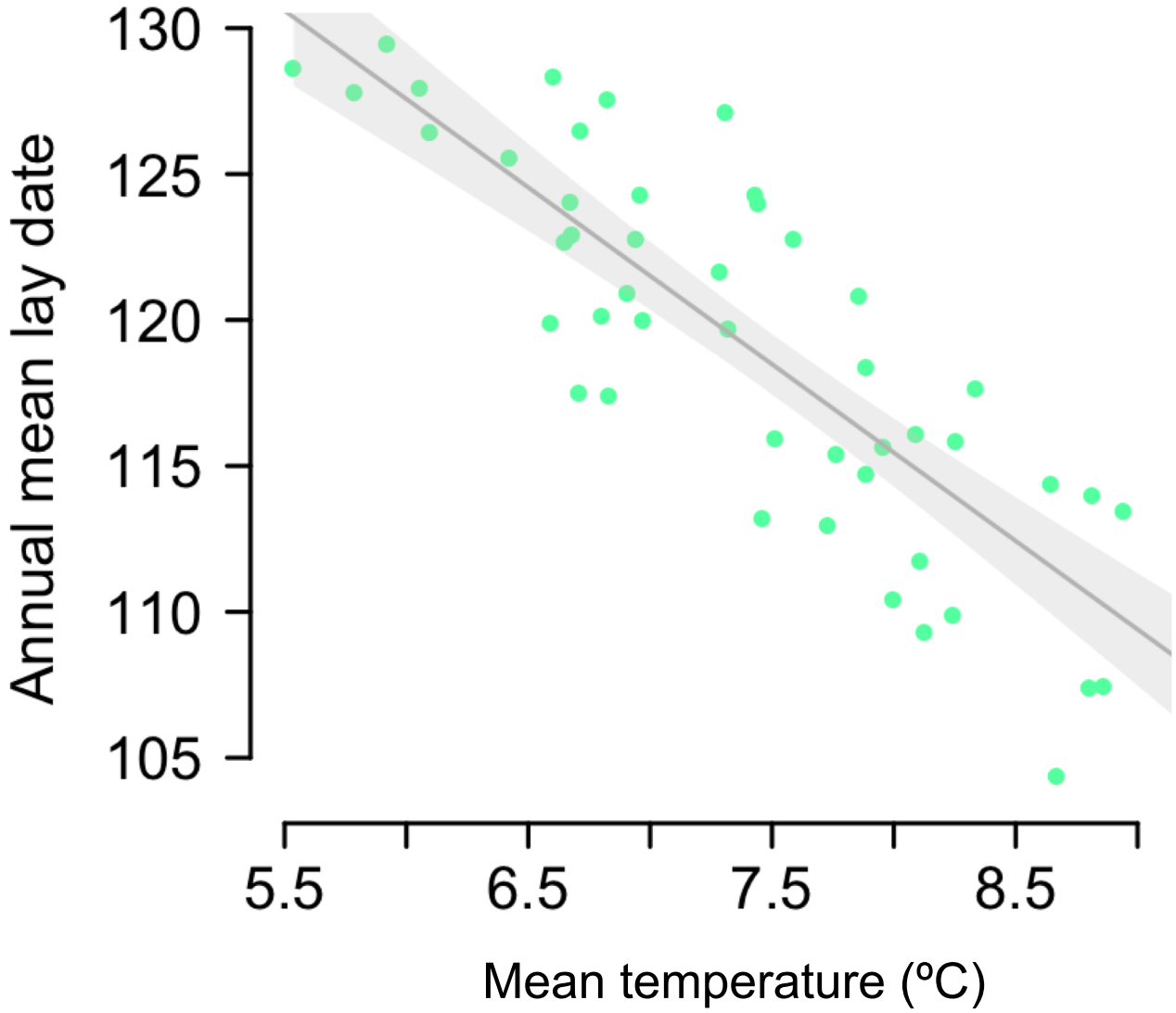


Exercise 1: Part A

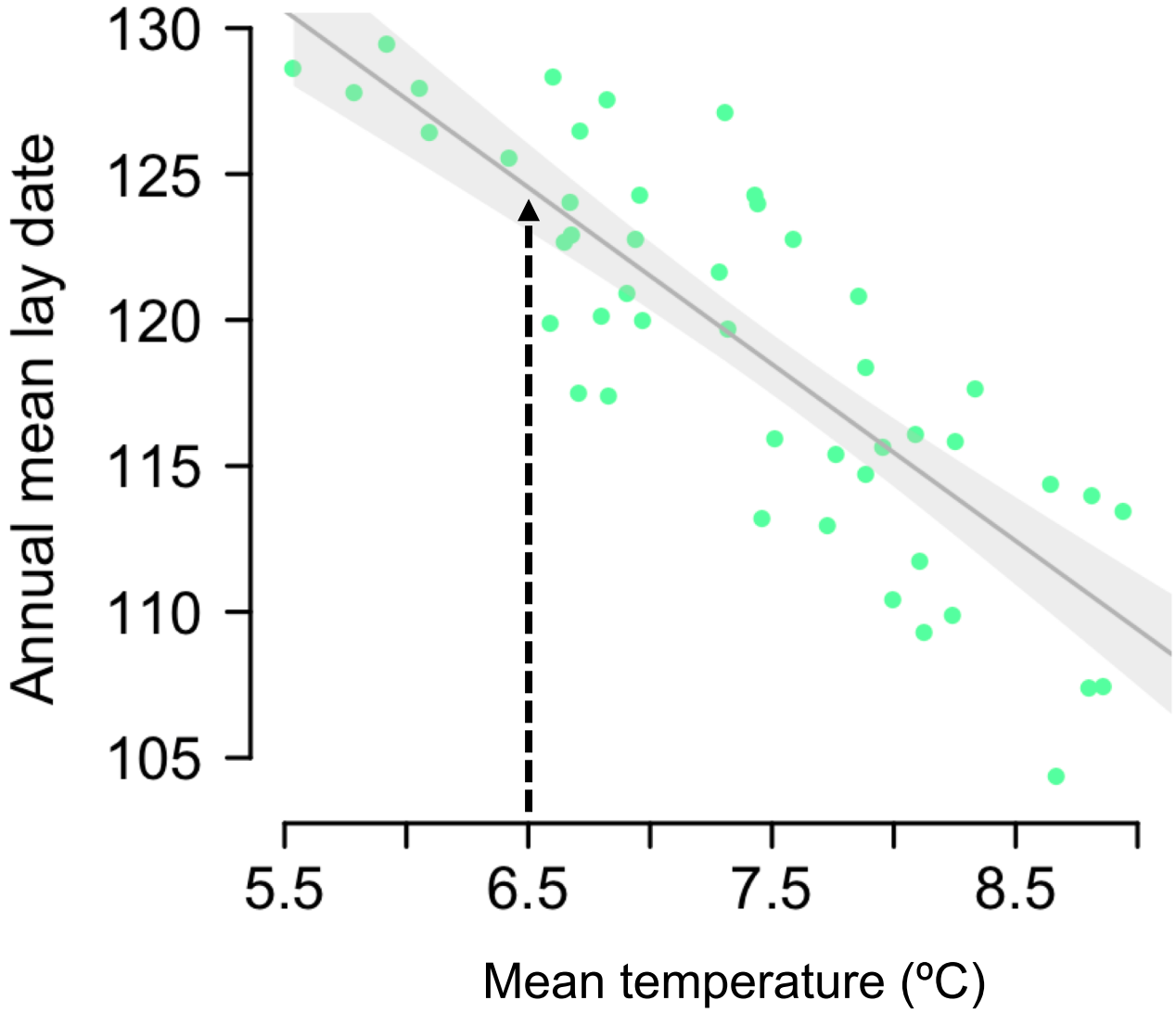
https://www.math.ntnu.no/emner/ST2304/2020v/Week05/Regression_module.html

Linear regression - Example

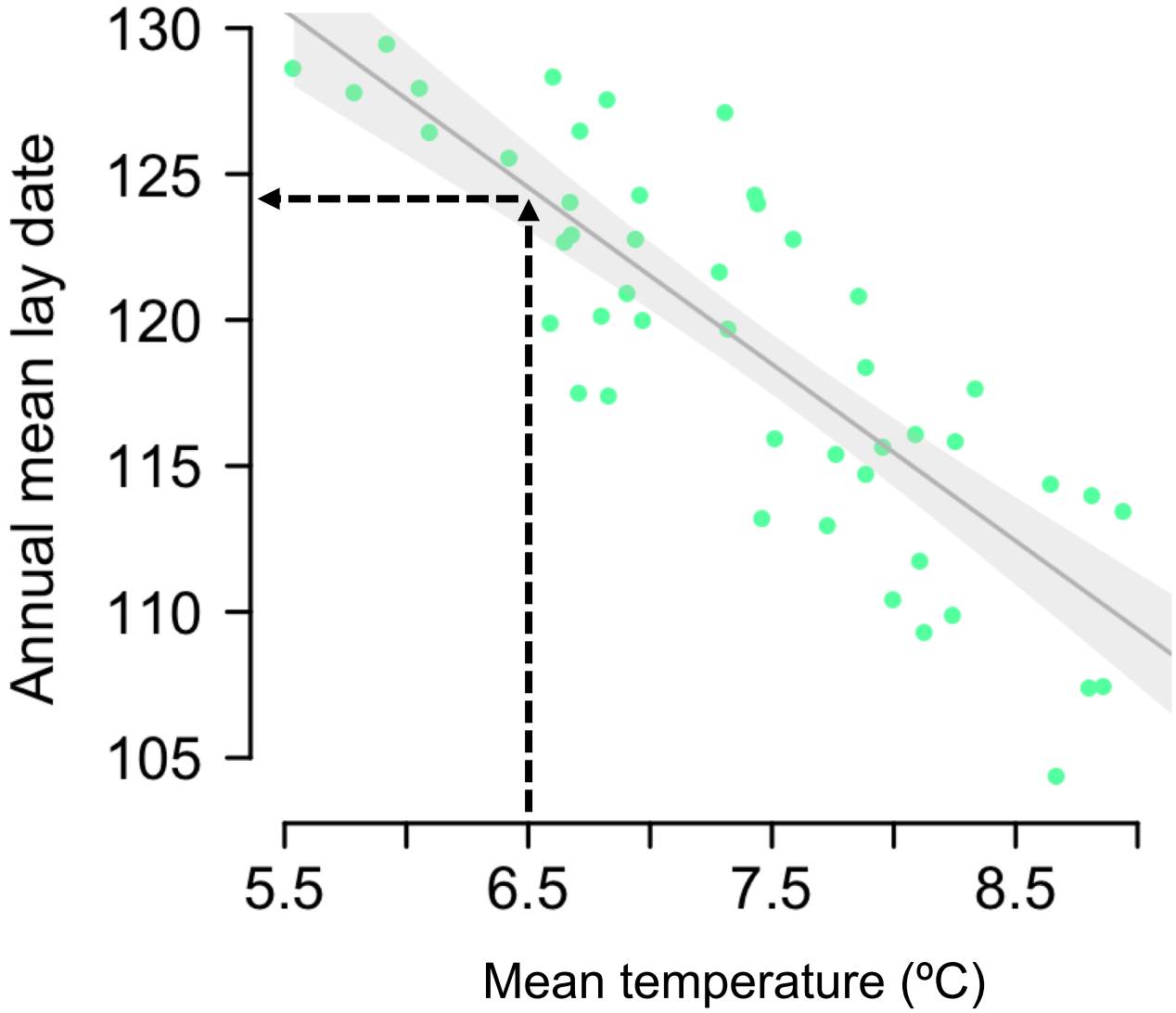
Estimate **relationship** between temperature and lay date



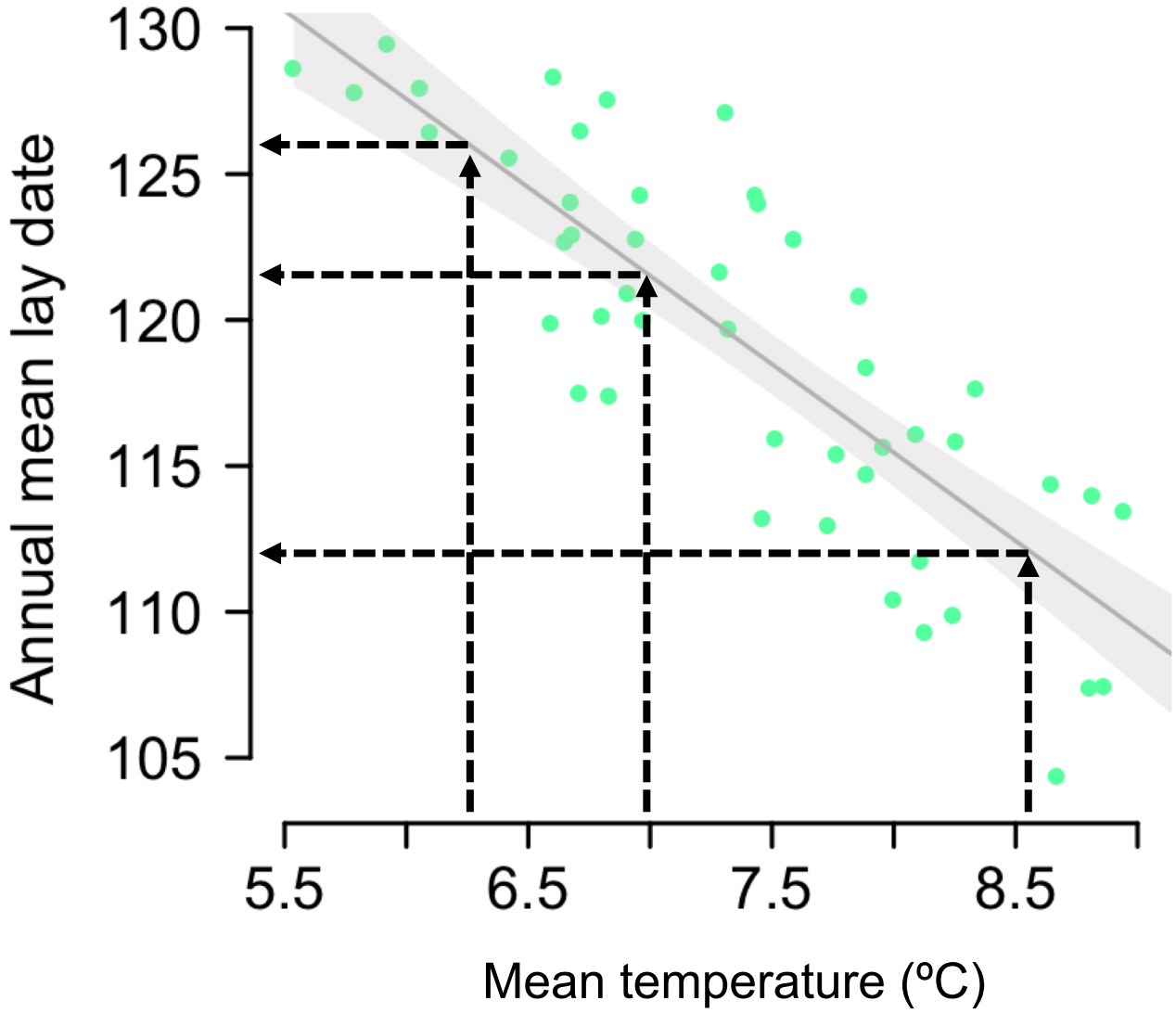
Linear regression - Example



Linear regression - Example

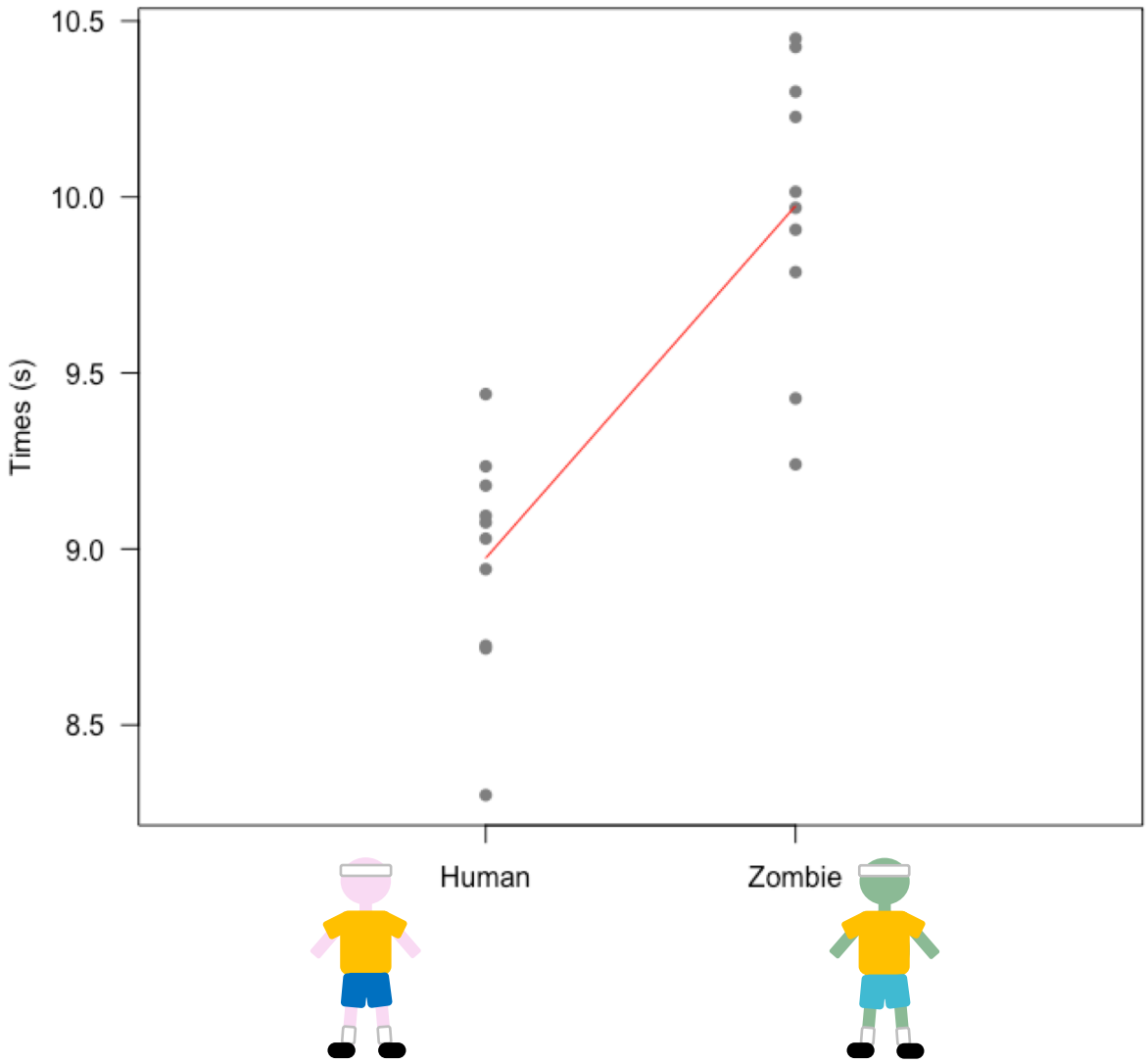


Linear regression - Example



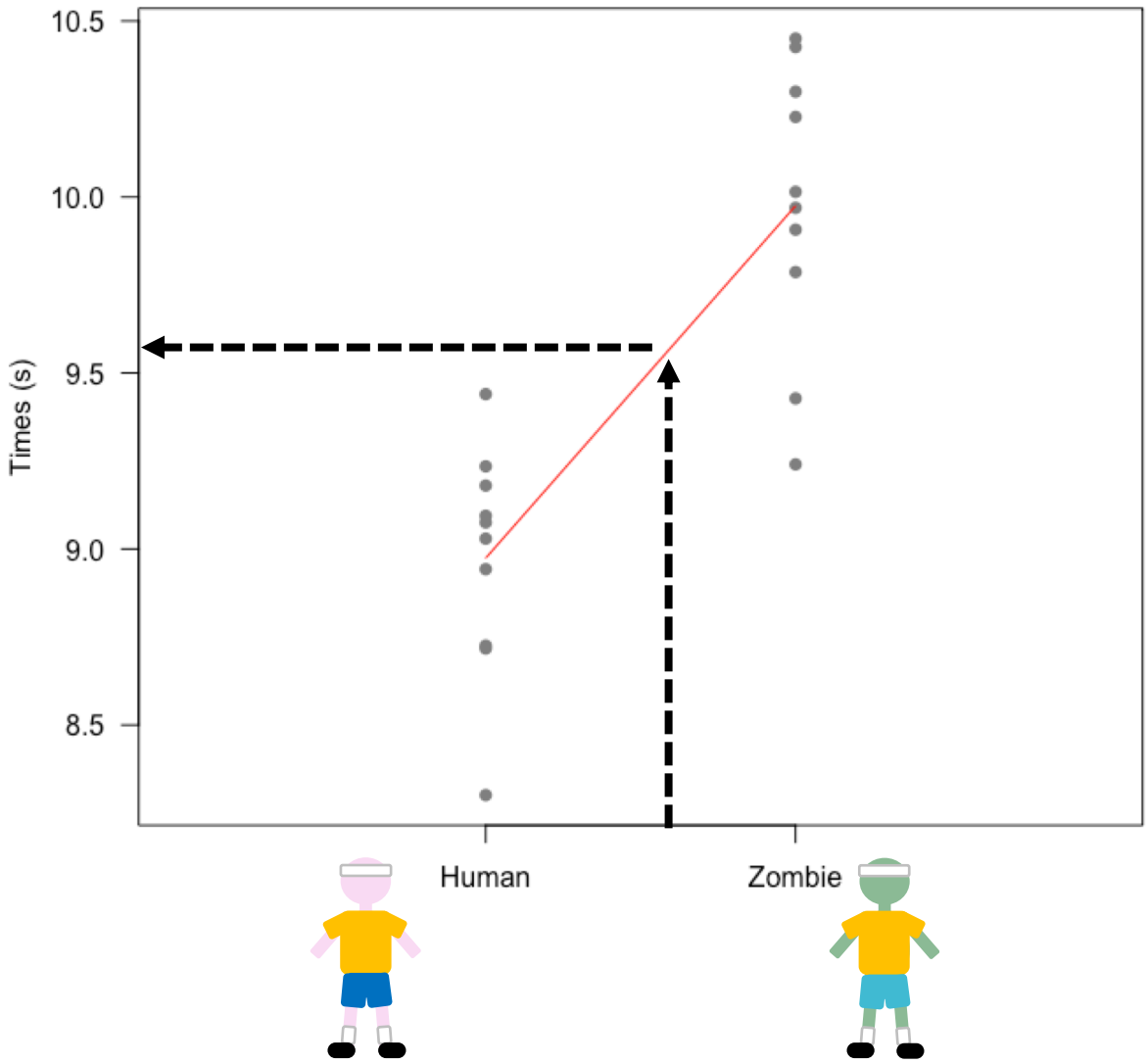
Linear regression – Example 2

Estimate **difference** in times for humans and zombies



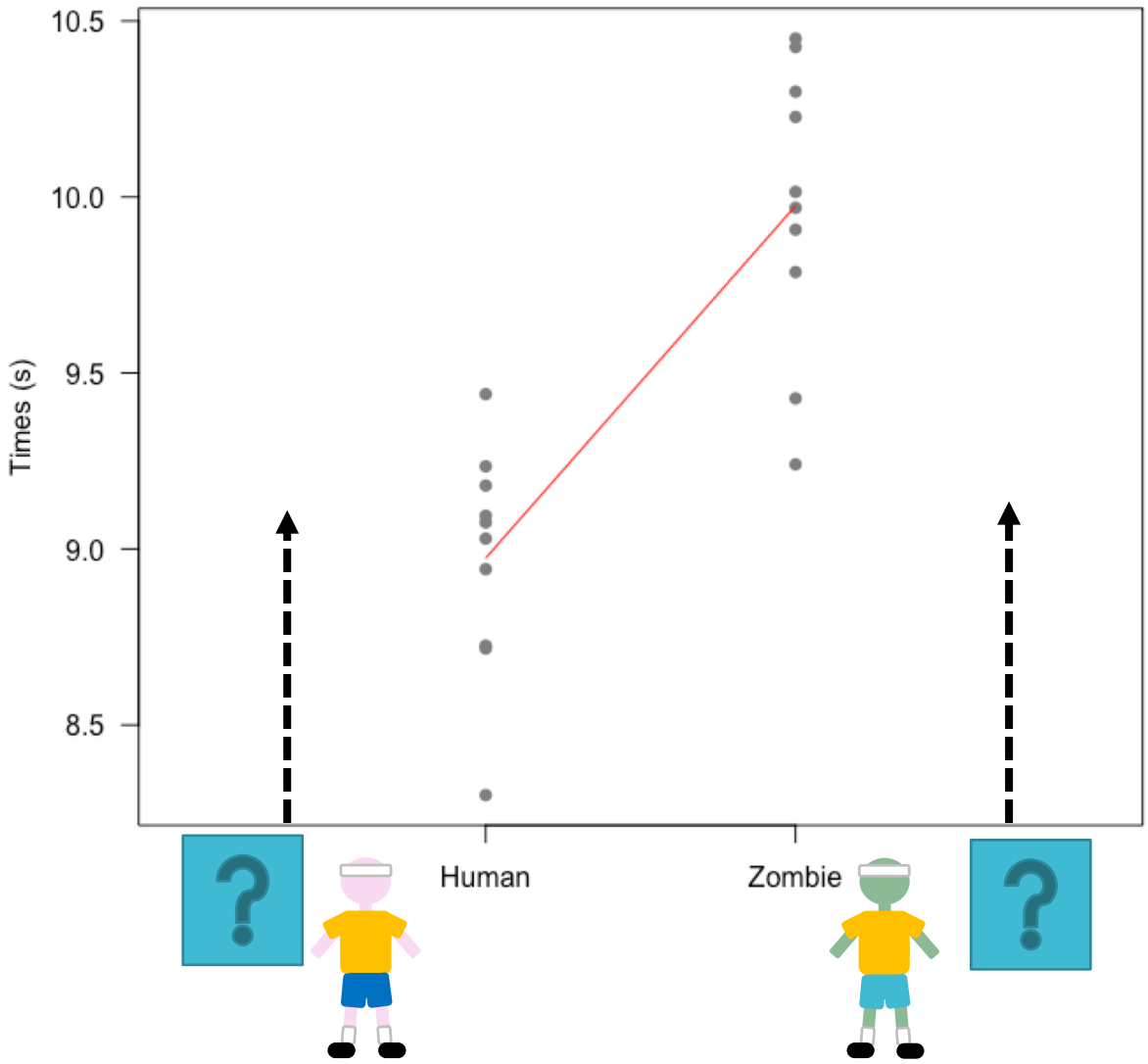
Linear regression – Example 2

Estimate **difference** in times for humans and zombies



Linear regression – Example 2

Estimate **difference** in times for humans and zombies

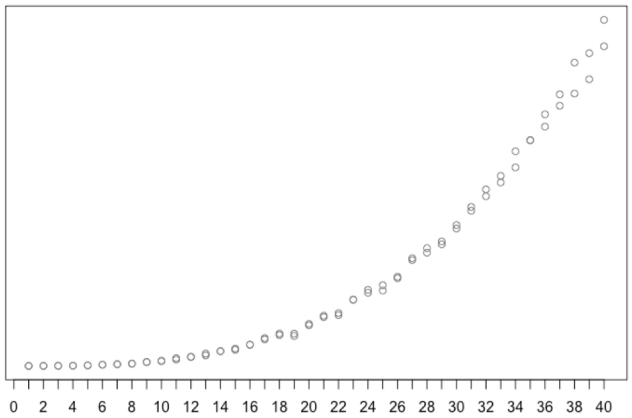


When to use regression

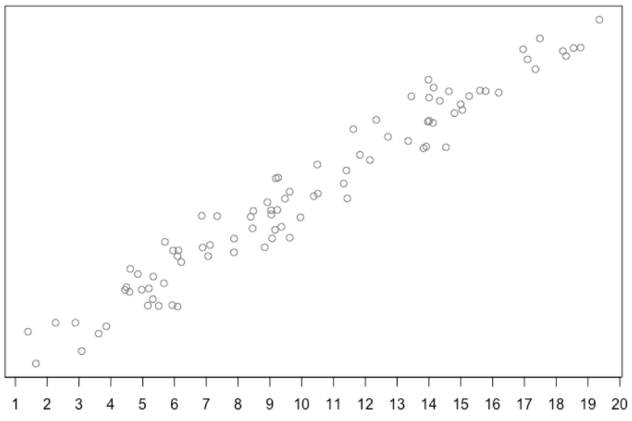
- Take a look at the four datasets below.
- For each, answer the question: **Is a continuous straight line a suitable model for this data?** (would a straight line work?)

1

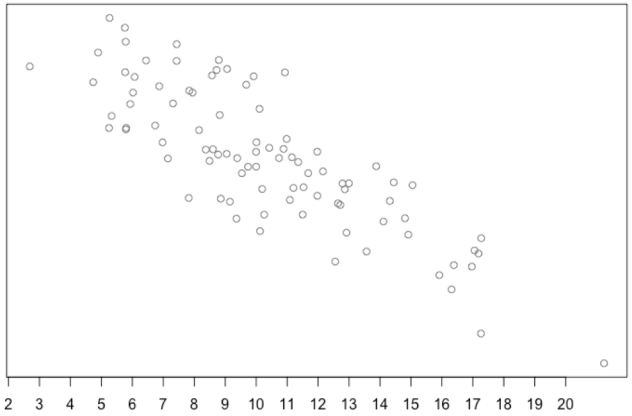
2



3



4

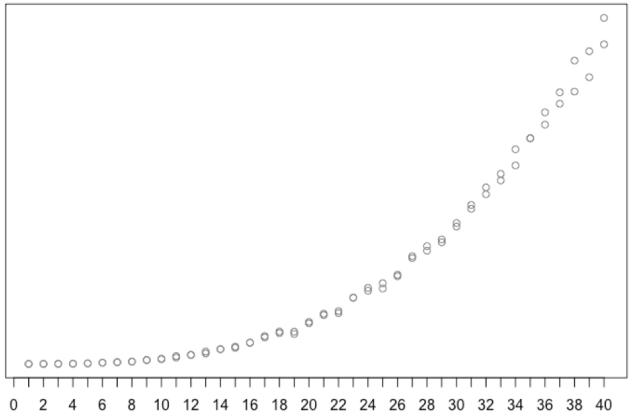


When to use regression

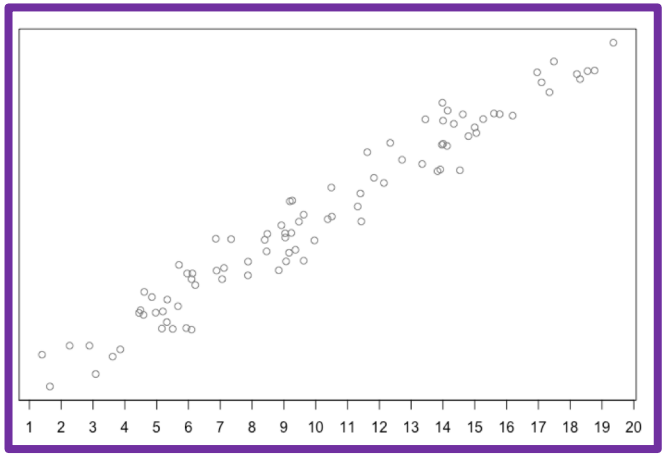
- Take a look at the four datasets below.
- For each, answer the question: **Is a continuous straight line a suitable model for this data?**

1

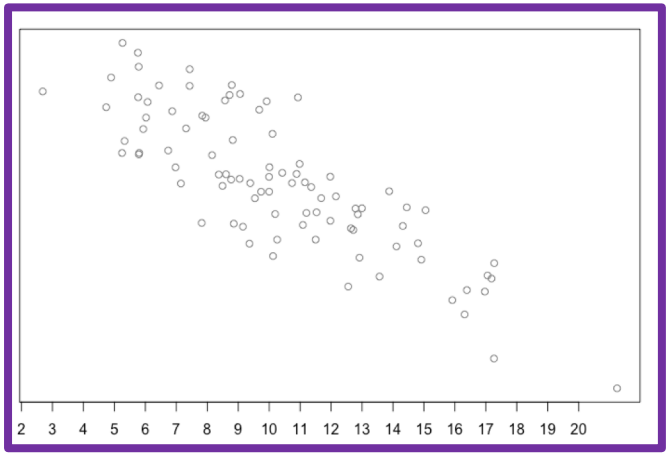
2



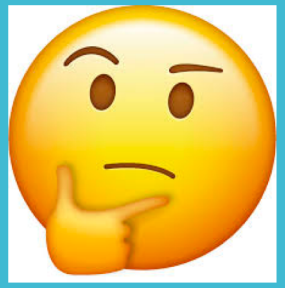
3



4

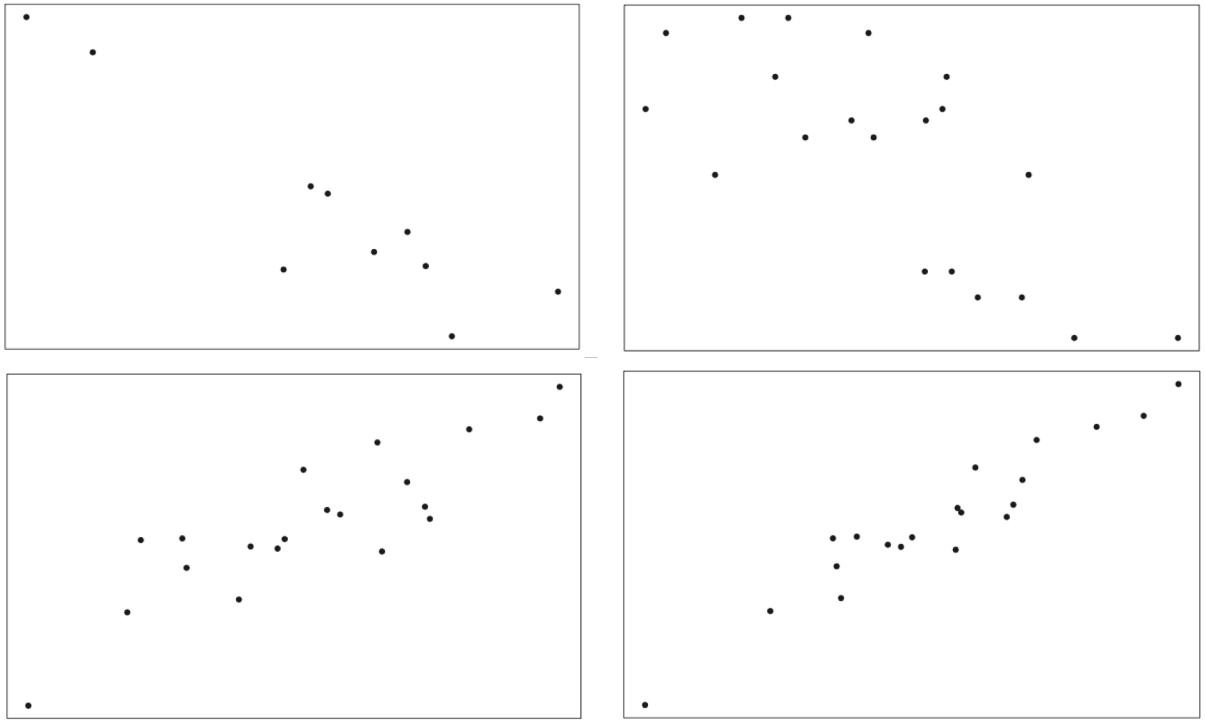


More about the model

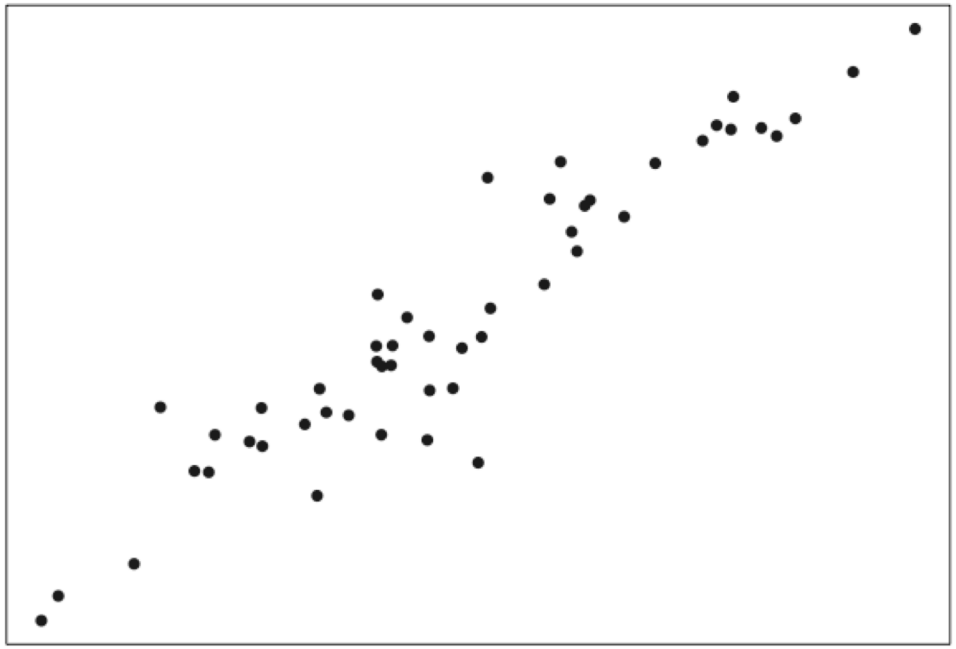


Exercise 2: What is the ‘best’ line?

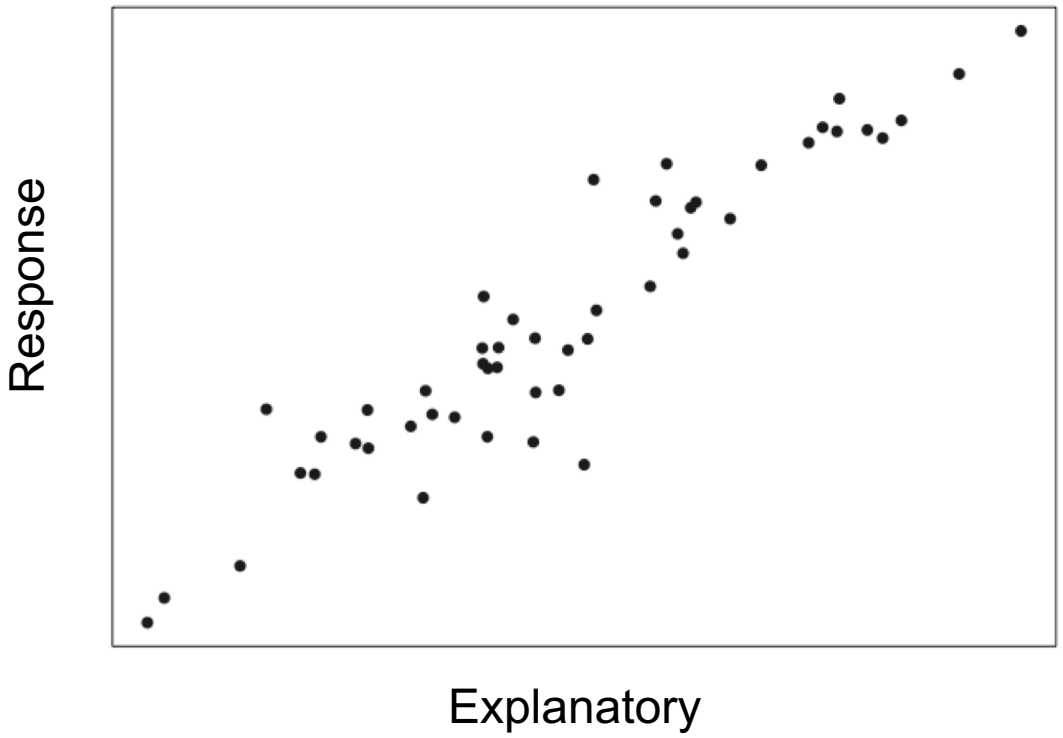
- Take a look at the four datasets below.
- For each: **draw a ‘best’ line on white boards.**



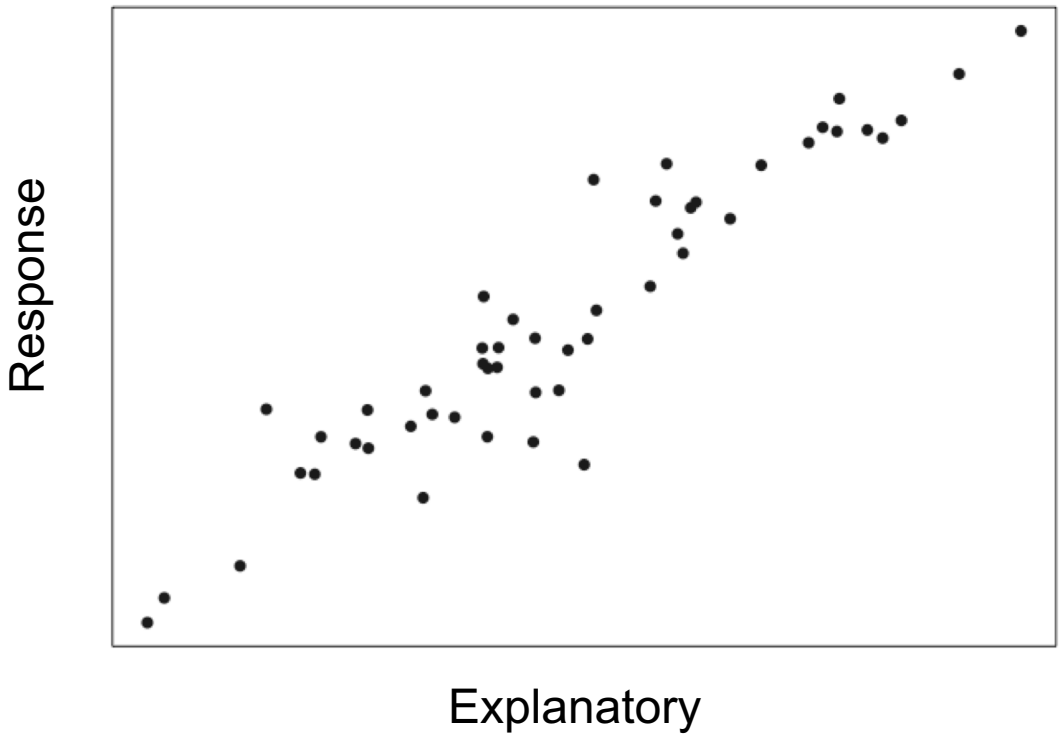
Fitting the line



Fitting the line

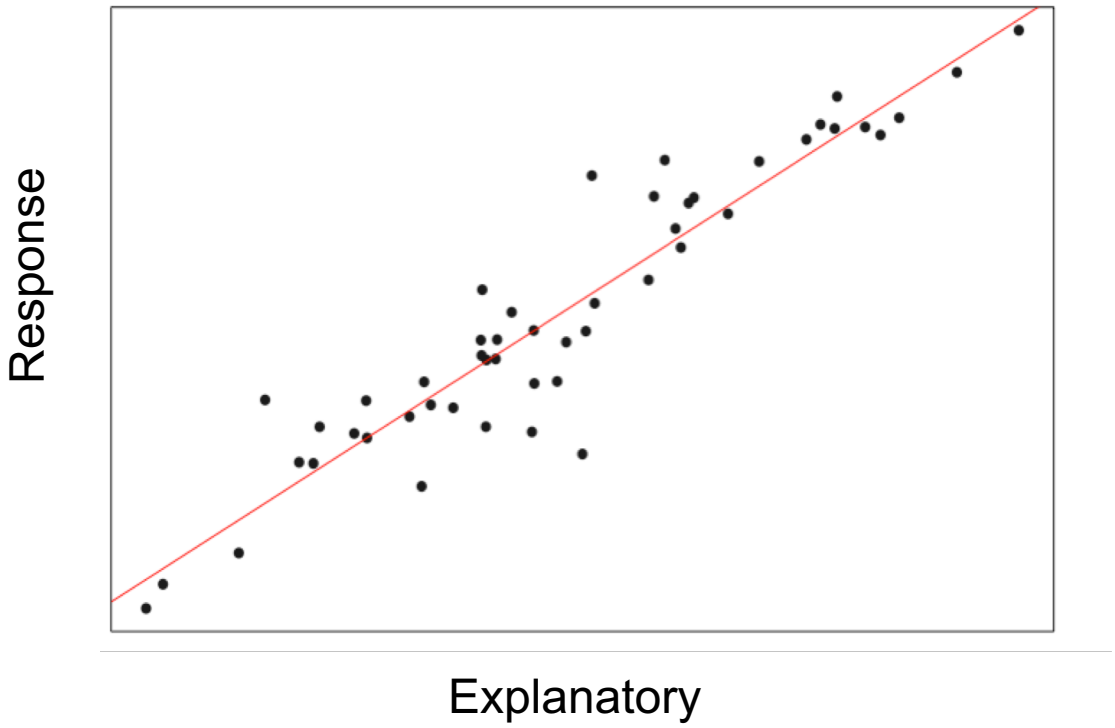


Fitting the line



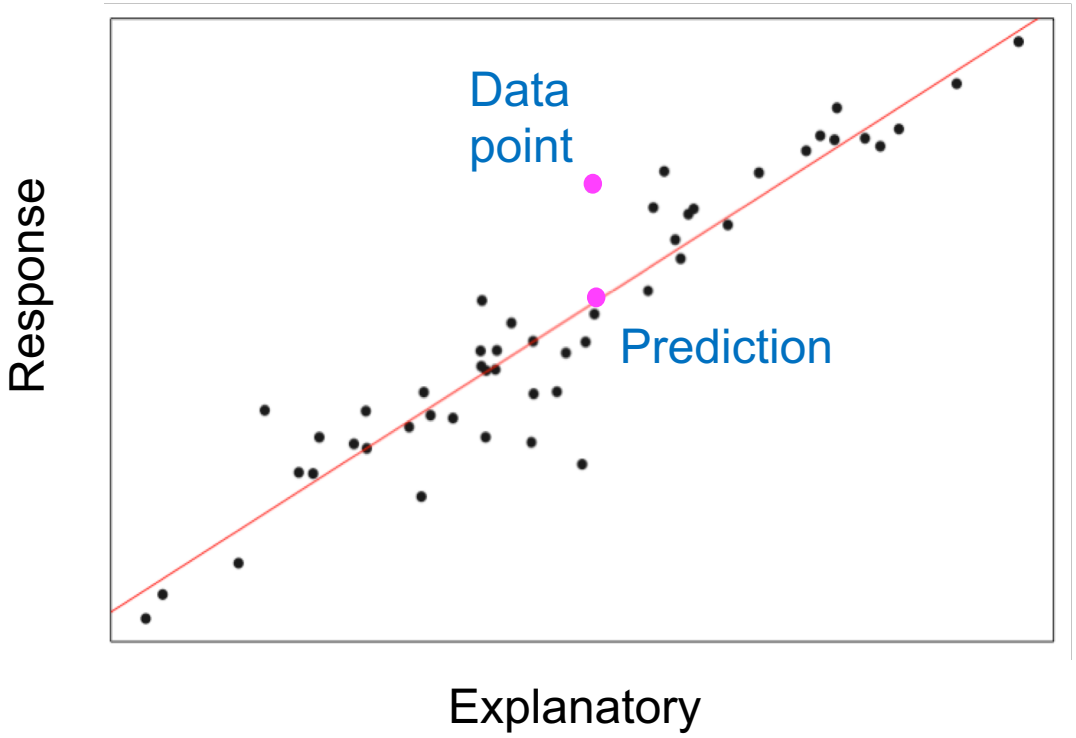
We want a model that represents how these data were generated

Fitting the line



Begin with a line – represents a relationship

Fitting the line

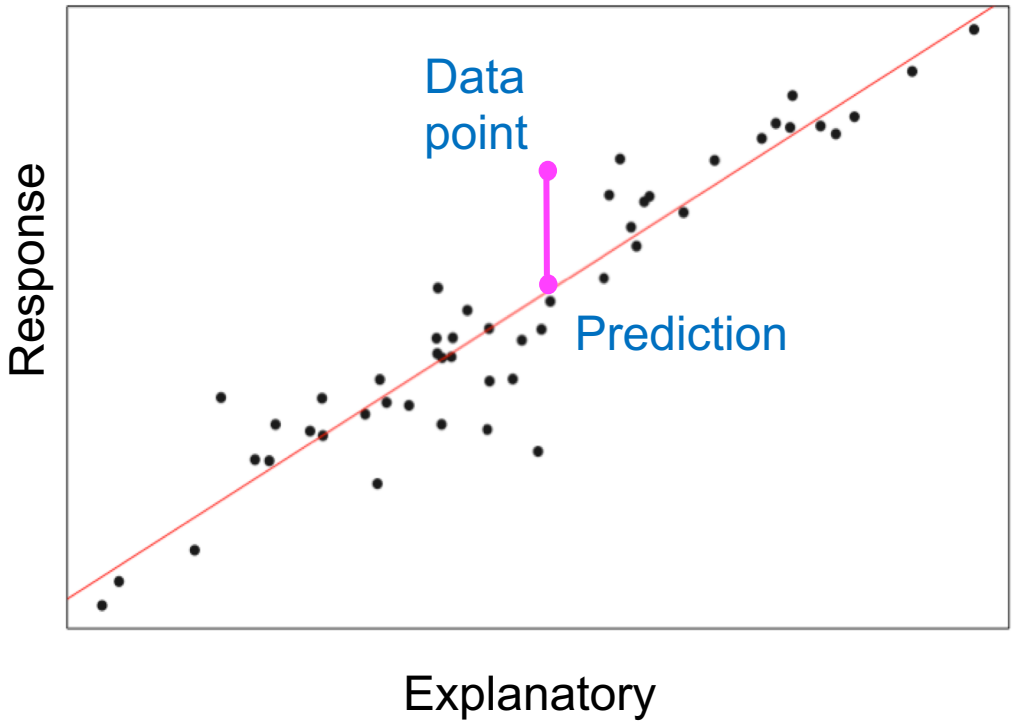


But there is also variation around this line

Fitting the line

Distance between them = error (**residual**)

$$\hat{y}_i - y_i$$



Our model

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Maximum likelihood

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Three components:

α

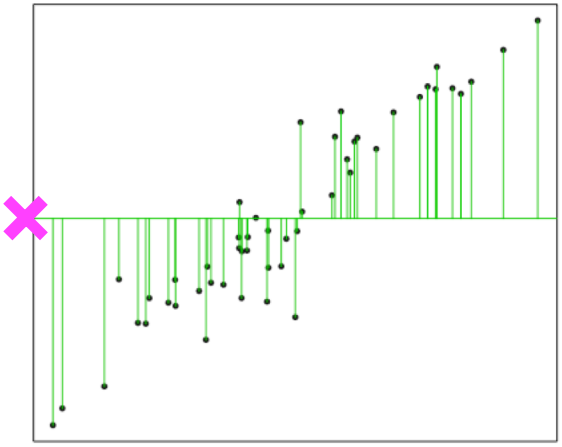
β

Control the line

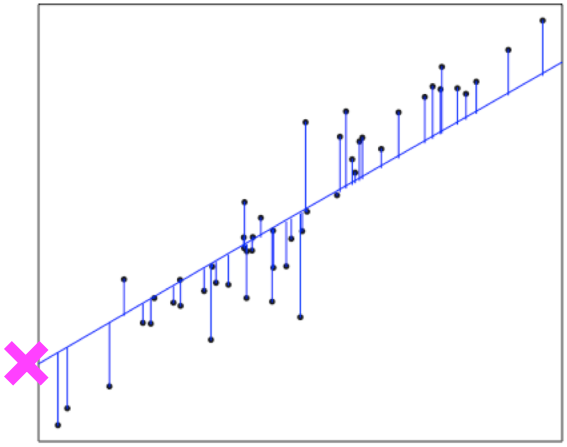
ε_i

Fitting the line

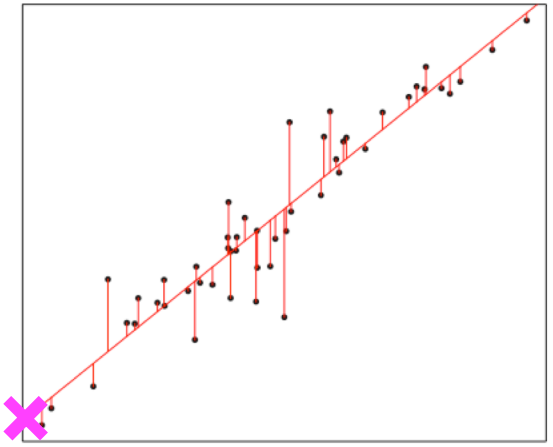
$\alpha = 10.19$
$\beta = 0$



$\alpha = 4$
$\beta = 0.6$

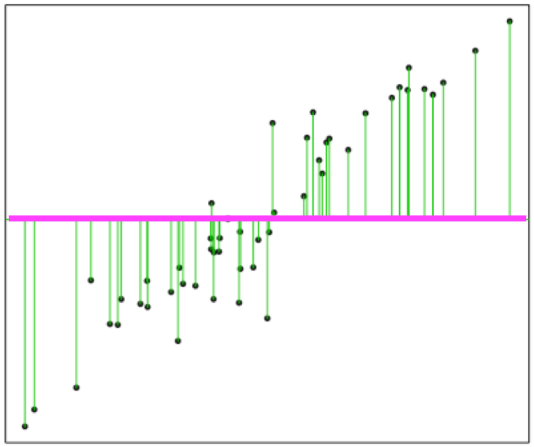


$\alpha = 1.57$
$\beta = 0.85$

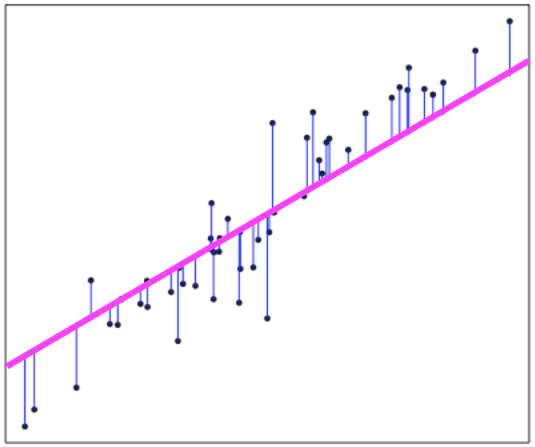


Fitting the line

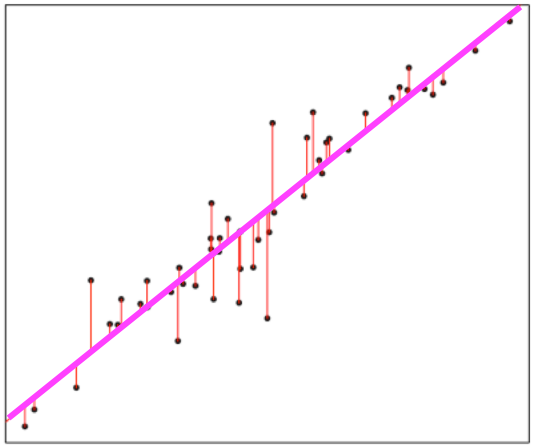
$\alpha = 10.19$
$\beta = 0$



$\alpha = 4$
$\beta = 0.6$



$\alpha = 1.57$
$\beta = 0.85$



Maximum likelihood

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

Three components:

α

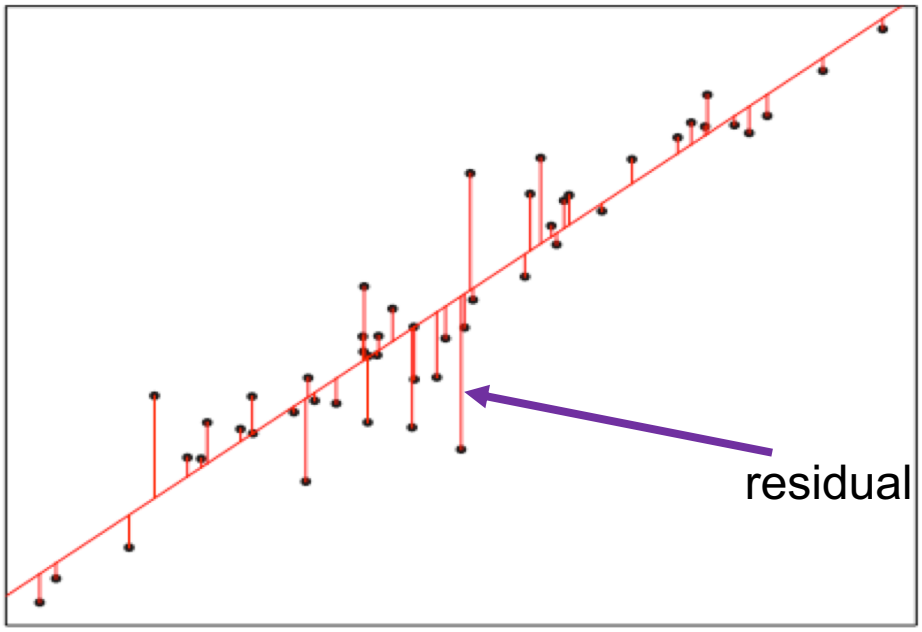
β

Control the line

ε_i → $N(0, \sigma^2)$ Variance of residuals

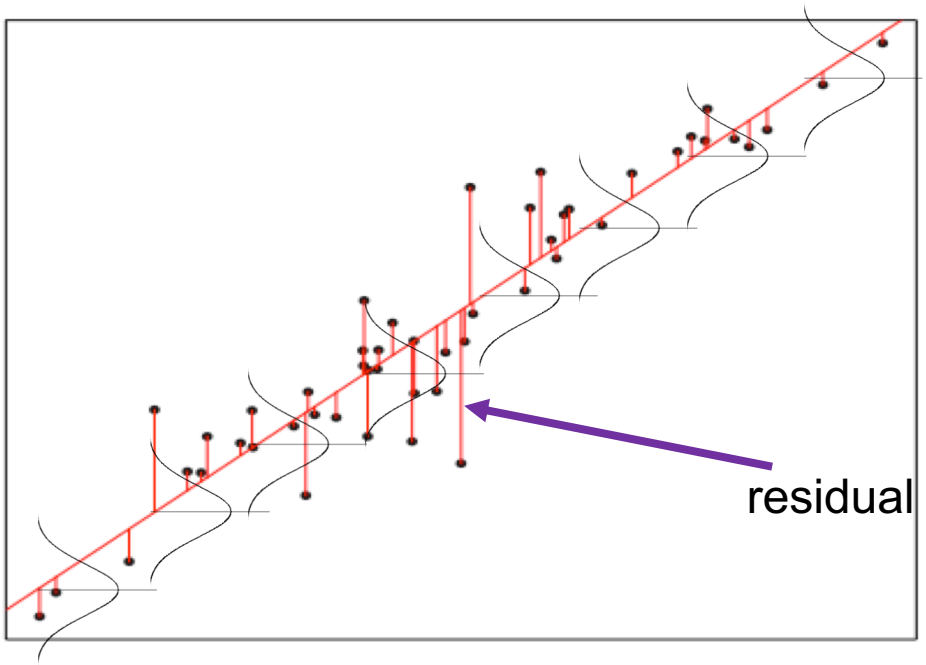
Fitting the line

We assume these residuals are normally distributed at each X value



Fitting the line

We assume these residuals are normally distributed at each X value



What is a ’best’ line?

Many different lines could be fitted to the same data

Can try to do it by eye

But also a mathematical way

Exercise 3: Try fitting a line

Go through Part B of exercise module

Summary of Part B

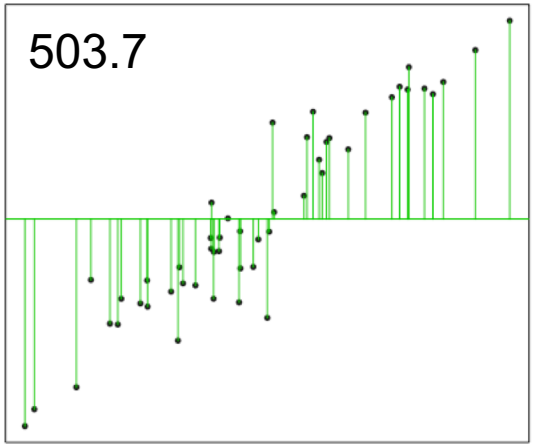
Using abline():

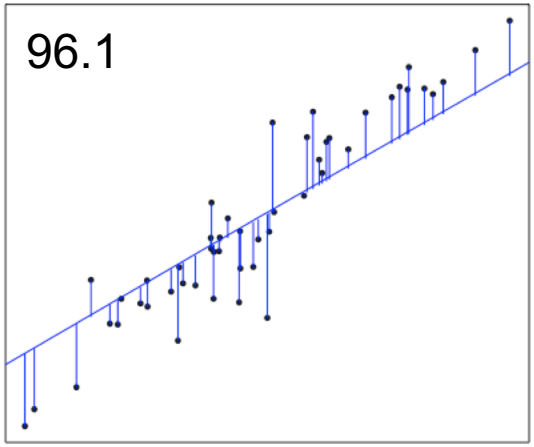
abline(a=0, b=1)

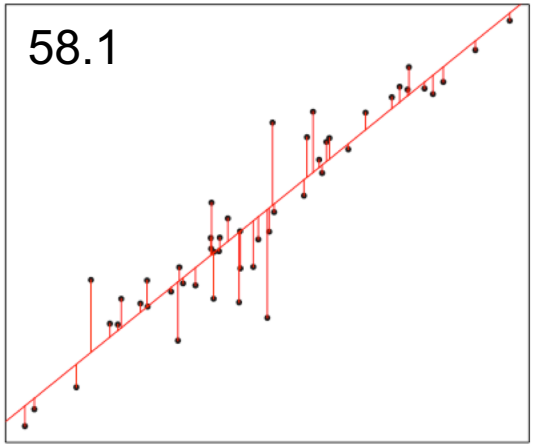
abline(a=0, b=2)

Summary of Part B

Sum of squares:

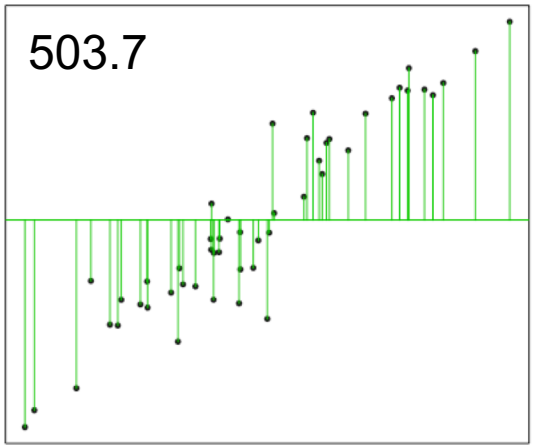


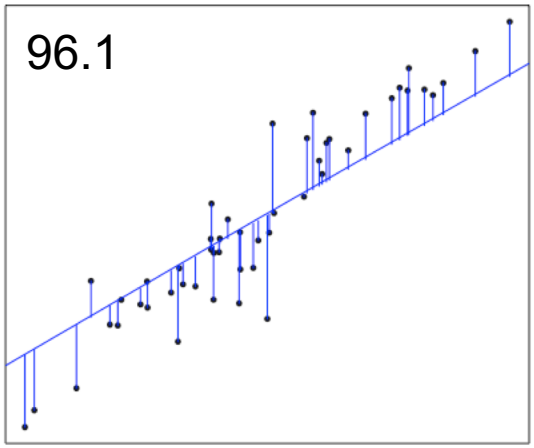


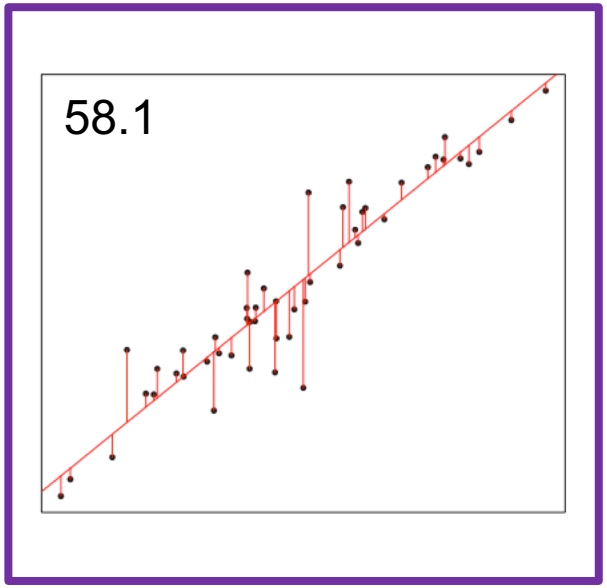


Summary of Part B

Sum of squares:

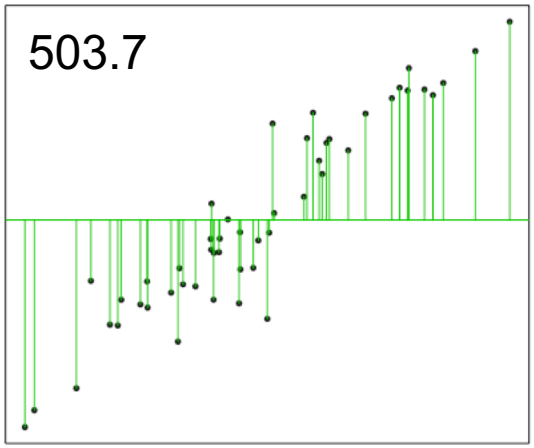


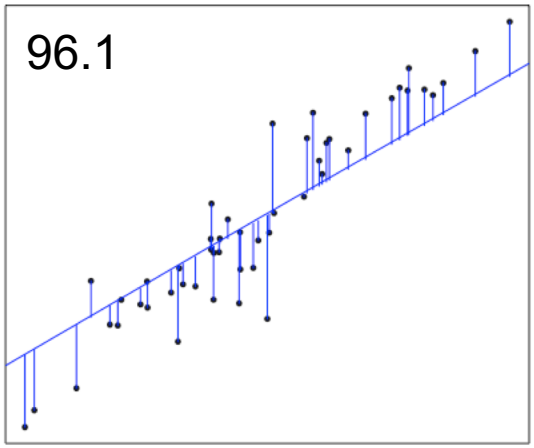


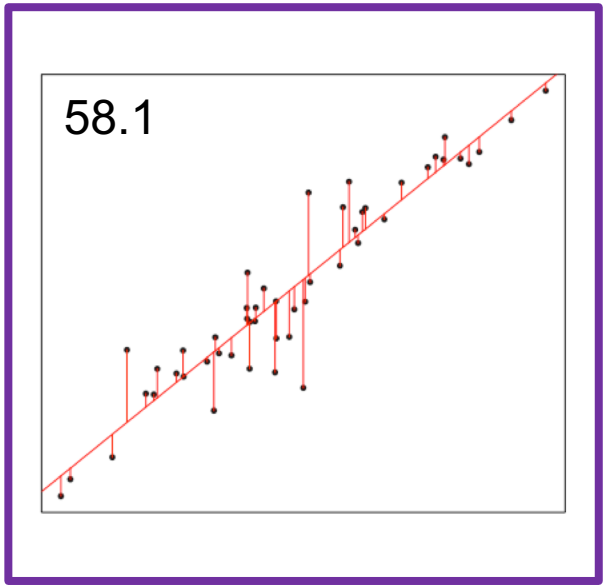


Summary of Part B

Sum of squares:

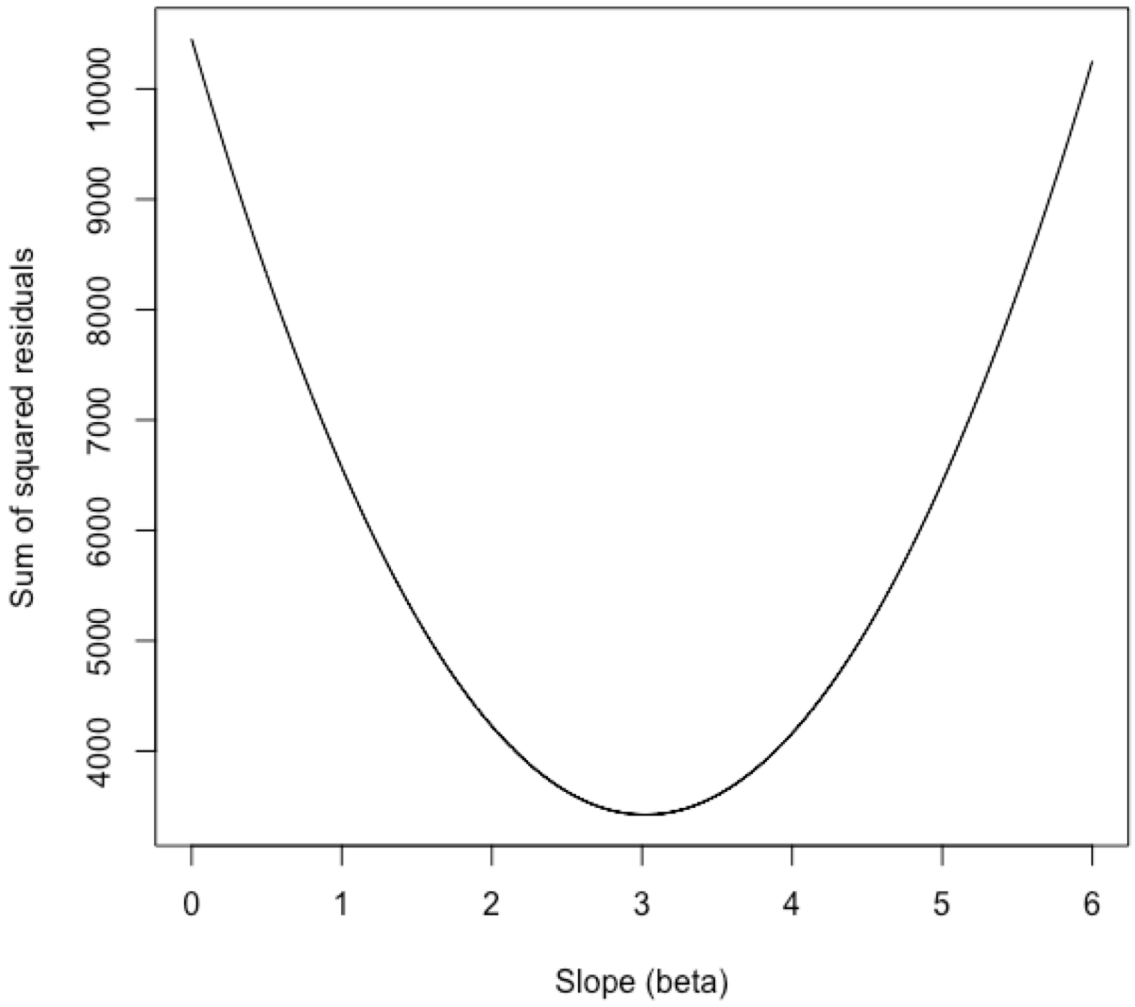




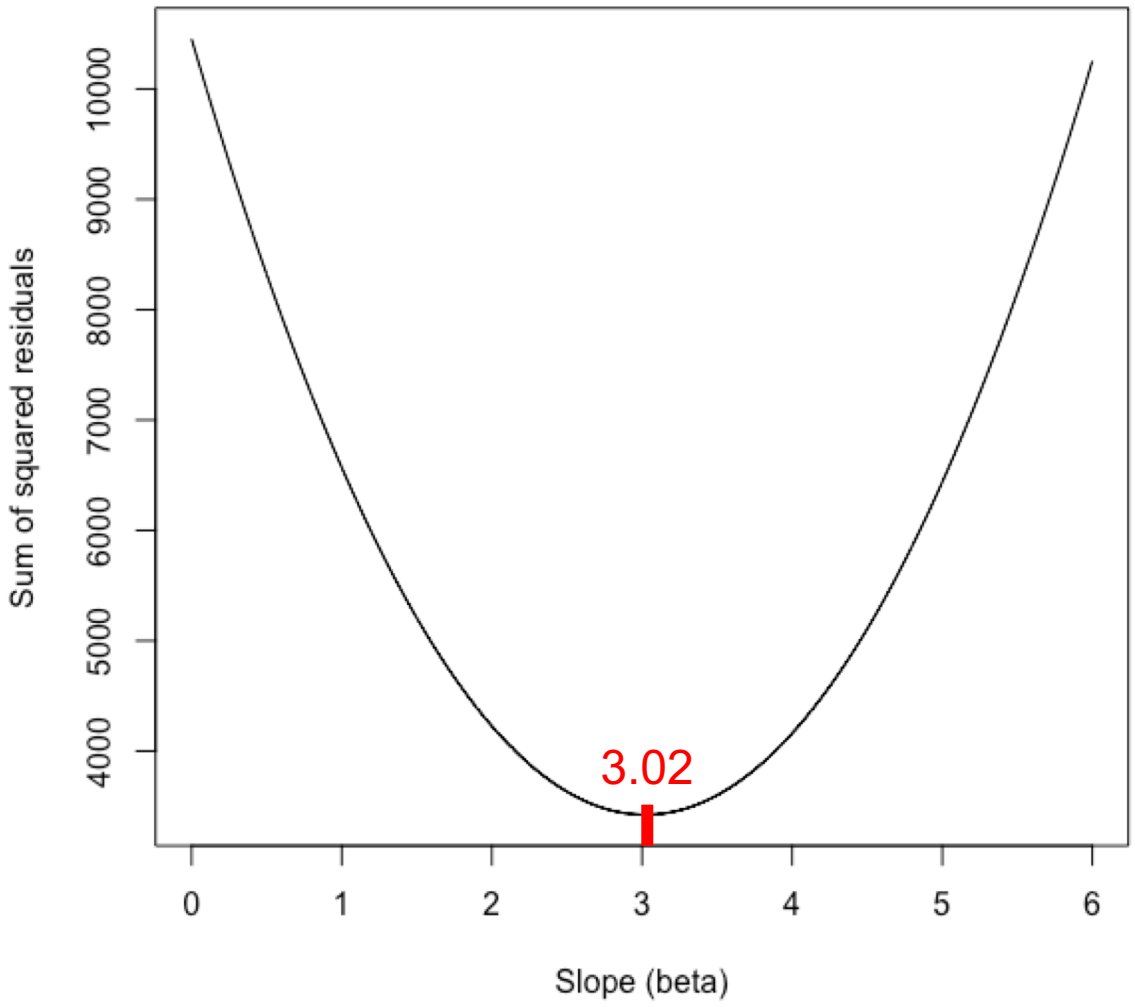


Your data best = 3423.242

Summary of Part B



Summary of Part B

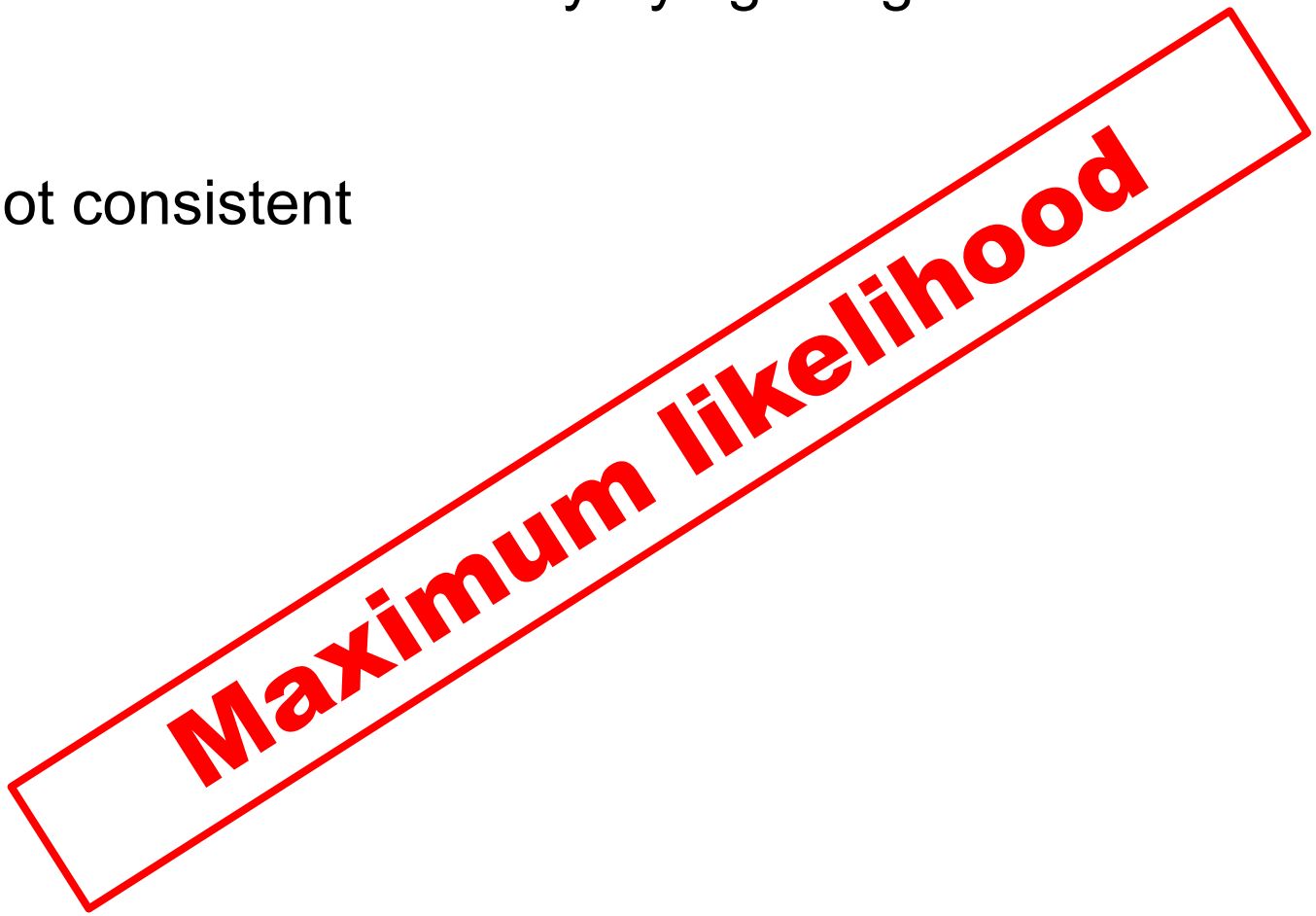


In Part B tried to find best line by trying things

Slow and not consistent

In Part B tried to find best line by trying things

Slow and not consistent



In Part B tried to find best line by trying things

Slow and not consistent

Find values of parameters that make the data most likely

Finding a best line with maximum likelihood estimation

Maximum likelihood

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

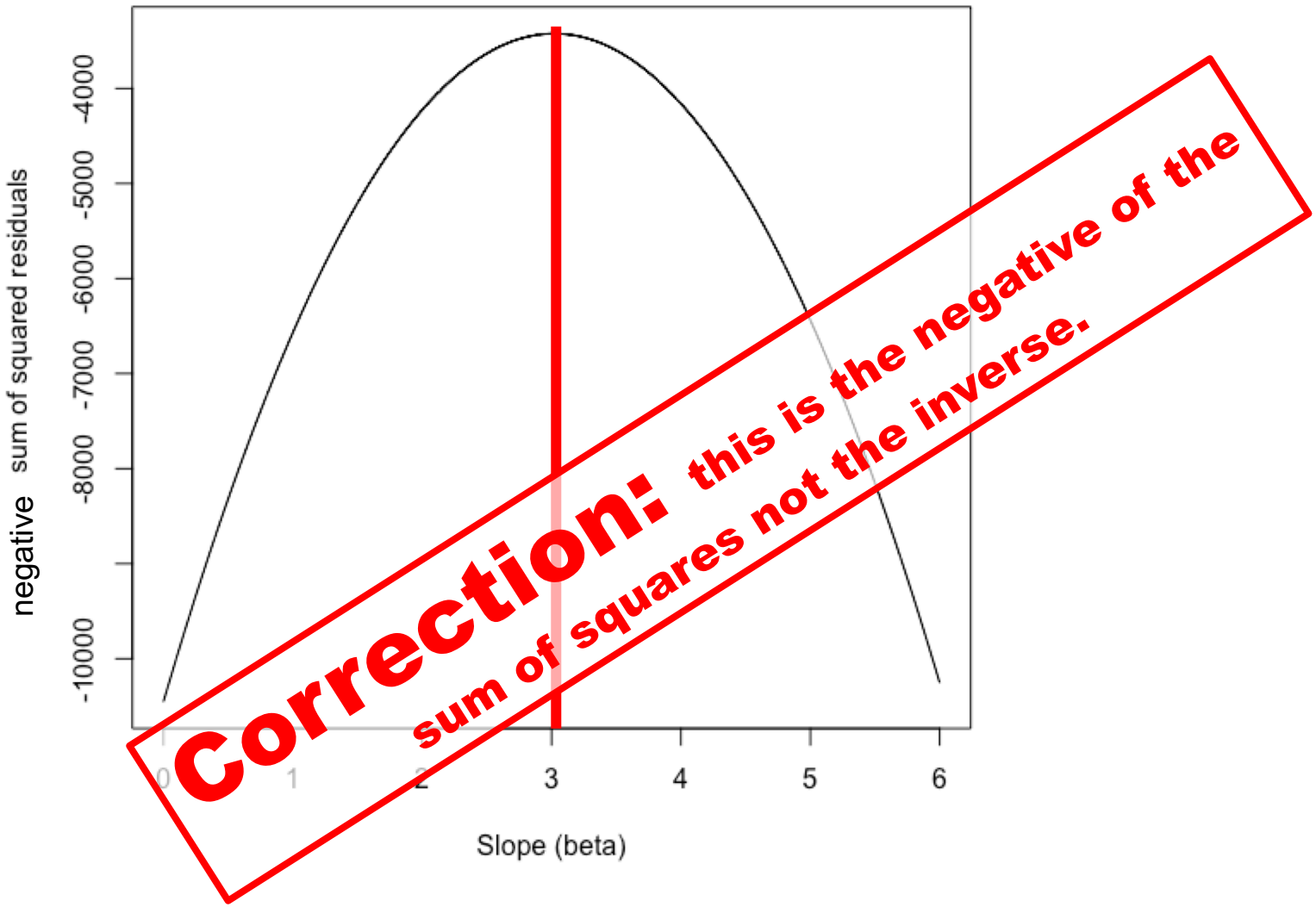
Three parameters that need to be estimated:

$$\alpha \qquad \beta \qquad \sigma^2$$

Variance is important too, even if we don't always interpret it

Summary of Part B

Maximum likelihood estimate also = 3.02



Lecture Summary

What are linear models?

What is linear regression?

How to find the best line?

Maximum likelihood and regression

Lecture Summary

What are linear models?

Broad set of models that link a response variable to an explanatory variable with linear equations.

What is linear regression?

How to find the best line?

Maximum likelihood and regression

Lecture Summary

What are linear models?

Broad set of models that link a response variable to an explanatory variable with linear equations.

What is linear regression?

A model that predicts values of a response variable from values of an explanatory variable. (lines)

How to find the best line?

Maximum likelihood and regression

Lecture Summary

What are linear models?

Broad set of models that link a response variable to an explanatory variable with linear equations.

What is linear regression?

A model that predicts values of a response variable from values of an explanatory variable. (lines)

How to find the best line?

minimize the sum of squares

Maximum likelihood and regression

we get to the best line by using maximum likelihood estimation and lm() function

Now a bit on functions