Using  for statistical analyses

Robert Bauer

Warnemünde, 06/05/2012

Day 8 - Agenda:

- Correlation
- Linear regression

Linear Regression

linear relationship:

$$y \sim x$$

linear model:

$$f(x) = y_i = a + bx_i + \varepsilon_i$$

- a slope
- b intercept
- y dependent variable
- x independent variable
- ε random error $\left(N\left(0, \sigma^2\right)\right)$

Linear Regression

SST (total sums of squares): total variation in the data

$$SST = \sum (y - \bar{y})^2$$

SSE (error sums of squares): residual variation, not explained by the model

$$SSE = \sum (y - \hat{y})^2$$

SSR (regression sums of squares): variation explained by the model

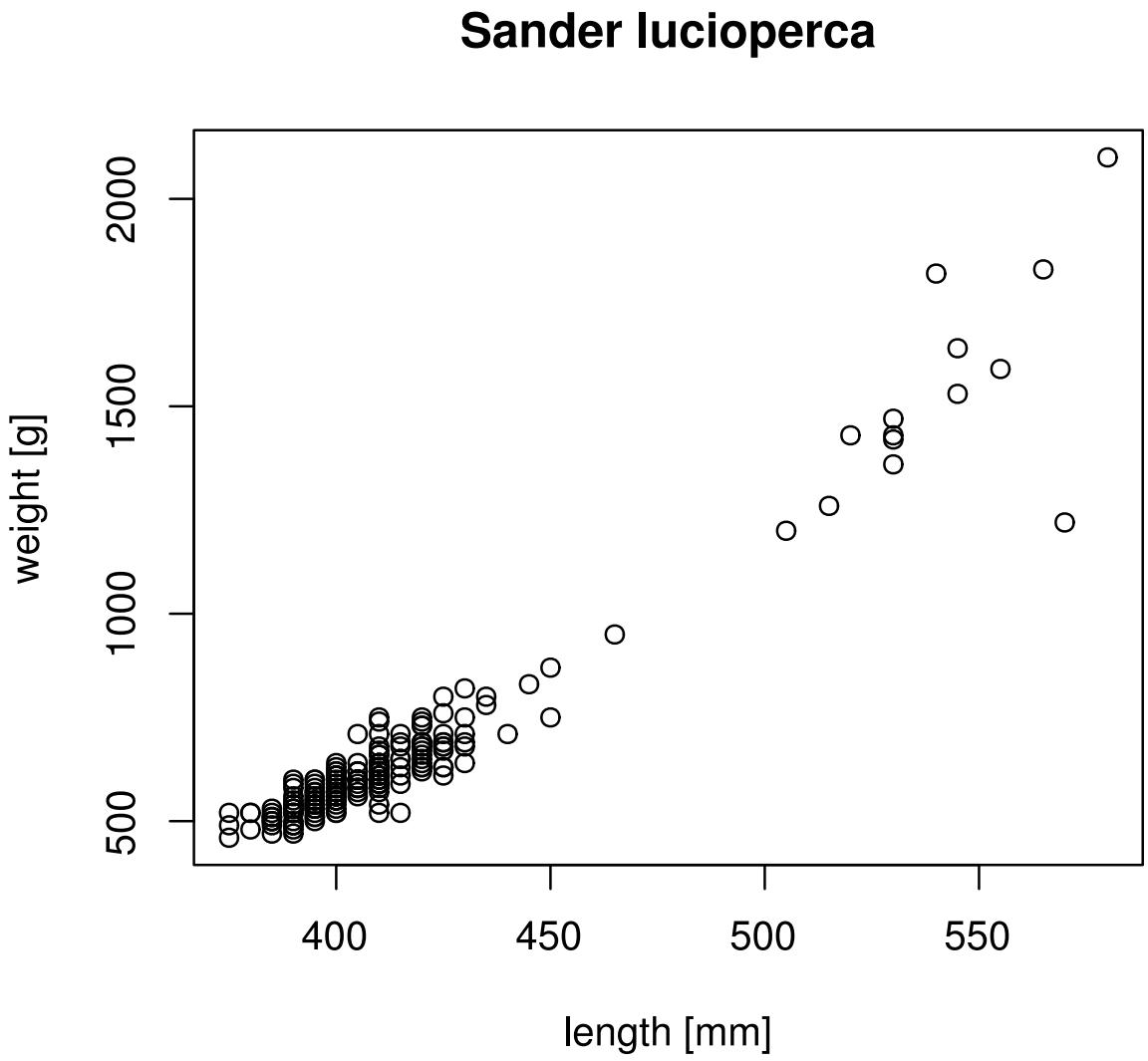
$$SSR = \sum (\bar{y} - \hat{y})^2$$

all together:

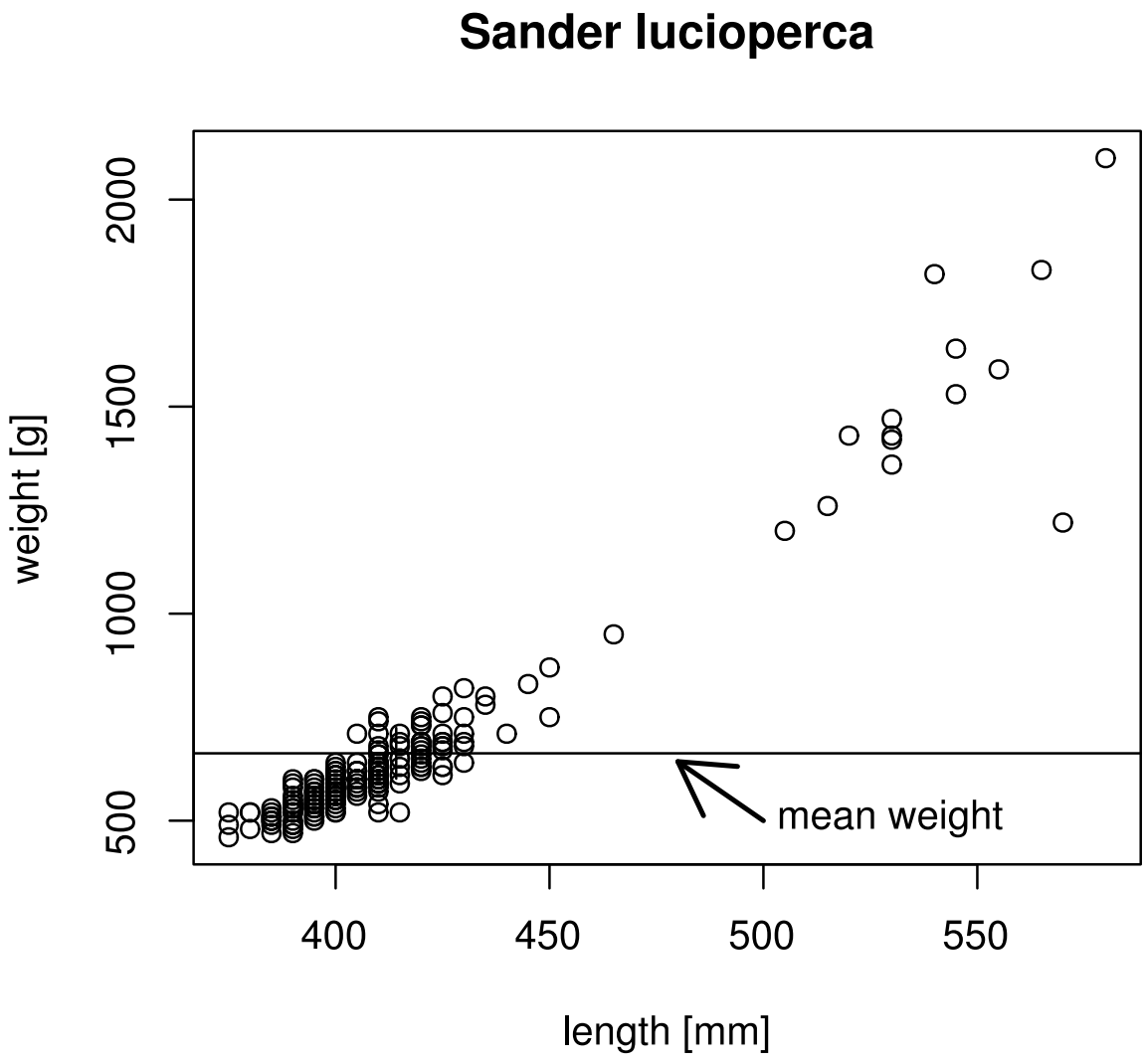
$$SST = SSE + SSR$$

Linear Regression in R

example: length vs. weight of pike-perchs (*Sander lucioperca*) (data: *pike-perch.xls*)

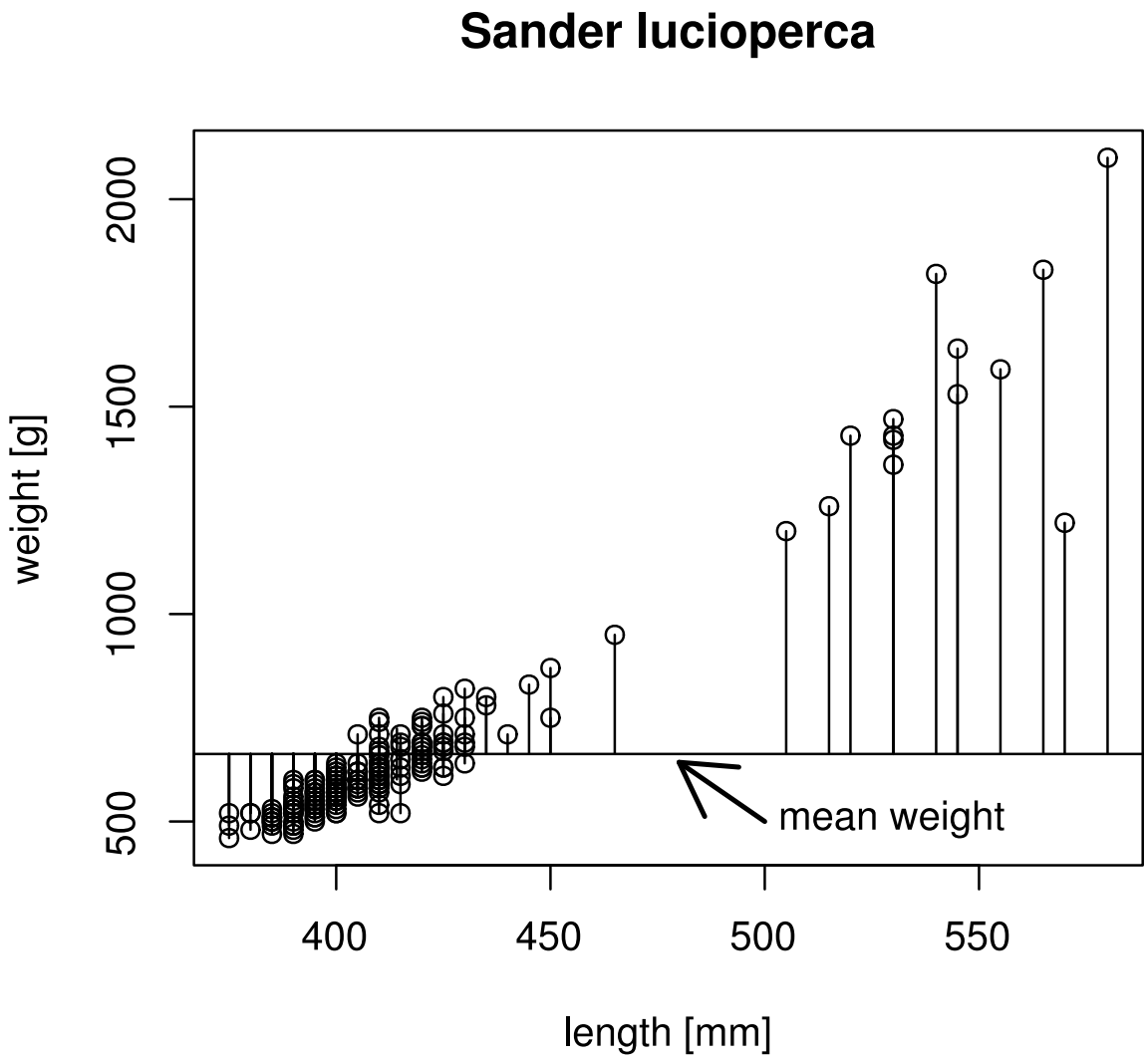


Linear Regression in R



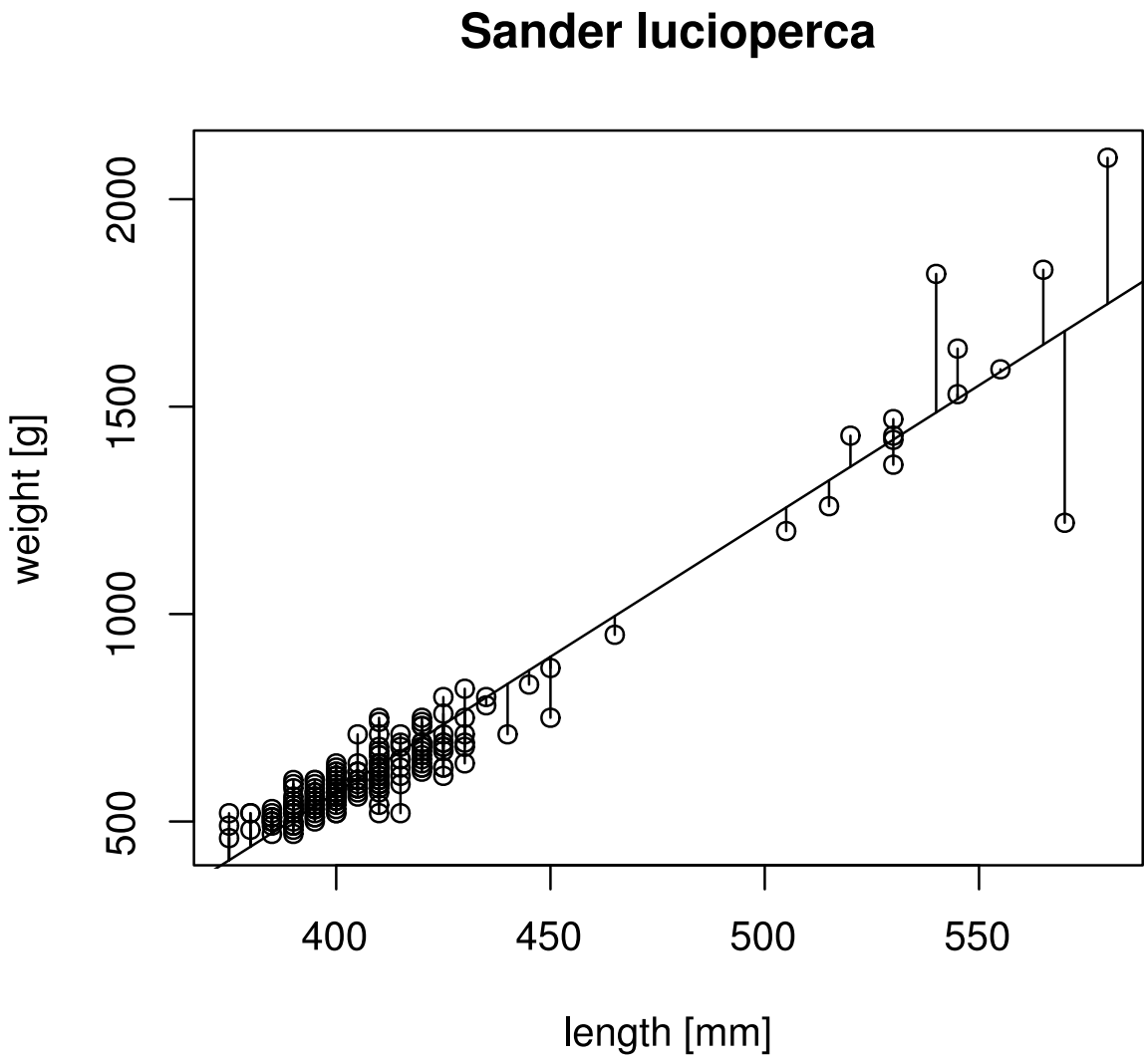
Calculating SST (total sums of squares)

$$SSE = \sum (y - \hat{y})^2$$



SSE (error sums of squares)

$$SST = \sum (y - \bar{y})^2$$

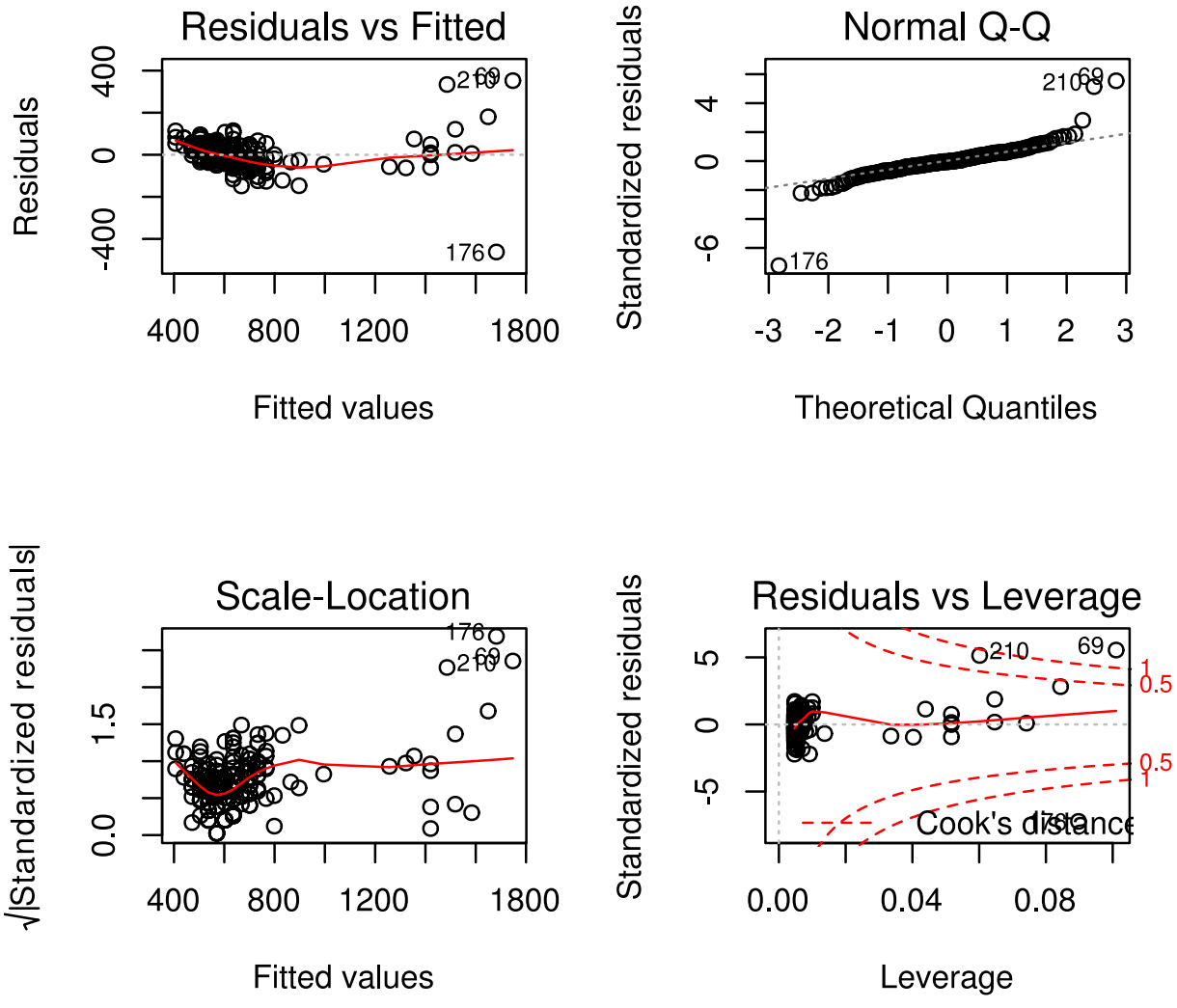


Linear Regression

Assumptions:

1. linearity of the relationship between dependent and independent variables
2. independence of the errors (no serial correlation)
3. homoscedasticity (constant variance) of the errors (residuals)
4. normality of the error distribution

Testing assumptions



Exercises

- repeat the linear regression analysis for two sub-groups (pike-perchs$<$ 500 mm and pike-perchs $>$ 500 mm) using the for-loop
- Is the data suitable for a regression analysis?
  if so,
  - fit the data to a linear model,
  - plot the data and the regression line
  - add the model equation (Hint: use the `coe()`)

Take home questions

- Which parameters are estimated in the linear regression?
- What is SST and what SSE, how do they relate to SSR?
- Why do we need an ANOVA for the linear regression model?
- What are the assumptions regarding the data to perform a linear regression analysis?

Exercises

- Import the data from the cod.xls file
- an linear relationship between length & weight as well as length and age is suggested
- Is the data suitable for a regression analysis?
  if so,
  - fit the data to a linear model,
  - plot the data, regression line and confidence interval bands,
  - add the model equation.