Session regression I: simple linear regression

Learning outcomes

- understand simple linear regression model incl. terminology and mathematical notations
- estimate model parameters and their standar error
- use model for checking the association between x and y
- use model for prediction
- assees model accuracy with RSE and R^2
- check model assumptions
- to be able to use `lm` function in R for model fitting, obtaining confidence interval and predictions

---

Introduction

Quiz: What do we already know about `simple linear regression`?

Description

- Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative, numerical) variables
  - one variable, denoted `x` is regarded as the *predictor*, *explanatory*, or *indepedent variable*, e.g. body weight (kg)
  - the other variable, denoted `y`, is regarded as the *response*, *outcome*, or *dependent variable*, e.g. plasma volume (liters)
- It is used to estimate the best-fitting straight line to describe the association

Used for to answer questions such as:

- is there a relationship between `x` exposure (e.g. body weight) and `y` outcome (e.g. plasma volume)?
- how strong is the relationship between the two variables?
- what will be a predicted value of the `y` outcome given a new set of exposure values?
- how accurately can we predict the outcome?

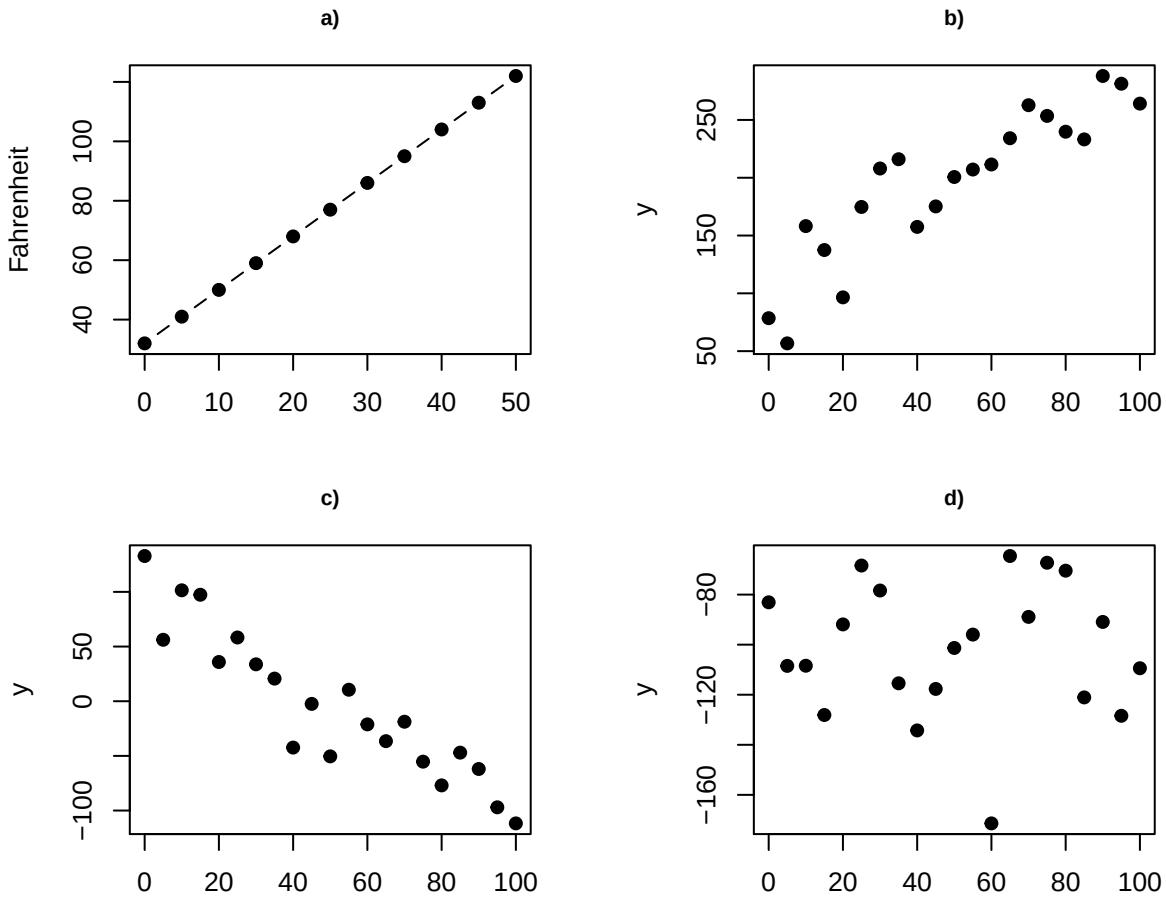


Figure 1: Determinisitc vs. statistical relationship: a) deterministic: equation exactly describes the relationship between the two variables e.g. $Fahrenheit = 9/5 * Celcius + 32$; b) statistical relationship between x and y is not perfect (increasing), c) statistical relationship between x and y is not perfect (decreasing), d) random signal

Example data

Example data contain the body weight (kg) and plasma volume (literes) for eight healthy men.

```
weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0) # body weight (kg)
plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12) # plasma volume (liters)
```

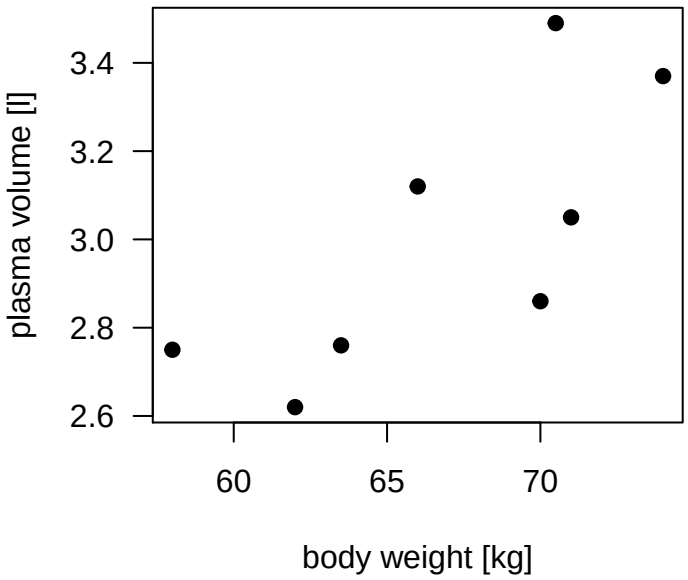


Figure 2: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice verca*.

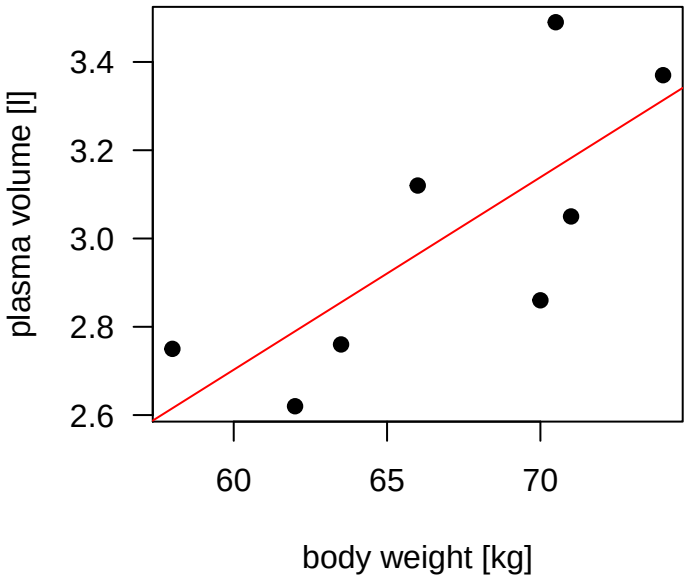


Figure 3: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice verca*. Linear regrssion gives the equation of the straight line that best describes how the outcome changes (increase or decreases) with a change of exposure variable (in red)

The equation of the regression line is:

$$y = \beta_0 + \beta_1 x$$

or mathematically using matrix notation

$$Y = \beta_0 + \beta_1 X$$

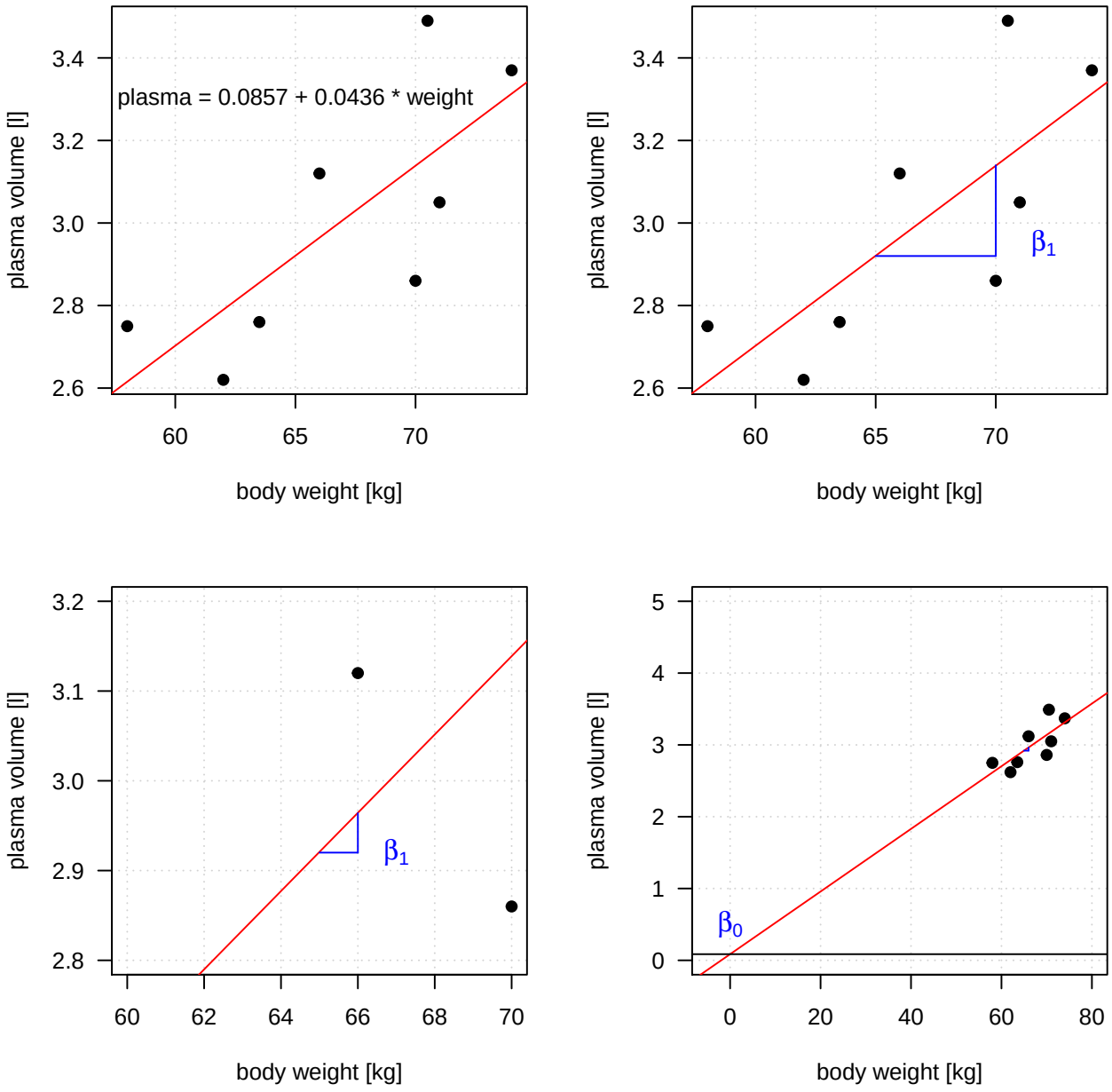
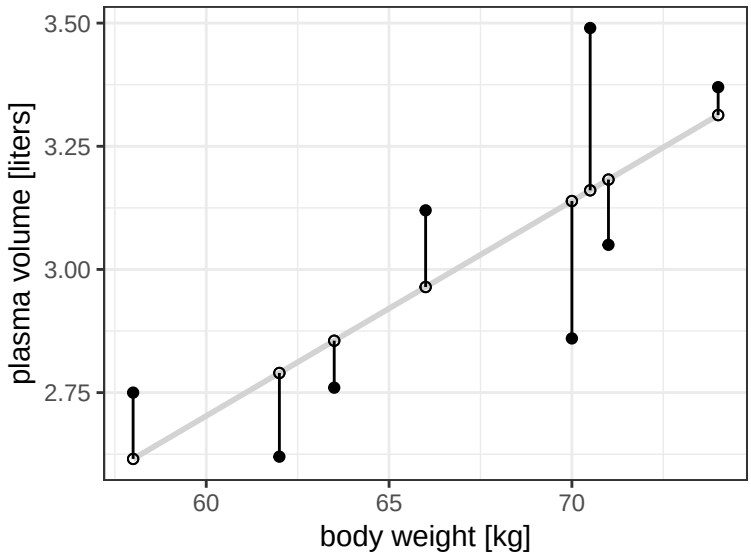


Figure 4: Scatter plot of the data shows that high plasma volume tends to be associated with high weight and *vice verca*. Linear regrssion gives the equation of the straight line that best describes how the outcome changes (increase or decreases) with a change of exposure variable (in red). Parameters explanation

Quiz: regression model parameters

Estimating the Coefficients

In practice, β_0 and β_1 are usually unknown. The best-fitting line is derived using the method of **least squares**, i.e. by finding the values of the parameters β_0 and β_1 tht minimize the sum of the squared vertical distances of the points from the line.



Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ represent n observation pairs, each of which consists of a maeasurment of X and Y, e.g. in our example we have 8 pairs of observations, e.g. (58, 2.75), (70, 2.86) etc.

```
weight <- c(58, 70, 74, 63.5, 62.0, 70.5, 71.0, 66.0) # body weight (kg)
plasma <- c(2.75, 2.86, 3.37, 2.76, 2.62, 3.49, 3.05, 3.12) # plasma volume (liters)
```

We seek to find coefficients esitmates $\hat{\beta}_0$ and $\hat{\beta}_1$ such that liner model fits the available data well, i.e. such that the resulting line is as close as possible to the 8 data points.

There are a number of ways of measuring *closeness*. By far the most common approach involves minimizing the *least squares* criterion.

Let $\hat{y}_i = \beta_0 + \hat{\beta}_1 x$ be the prediction Y based on the ith value of X. Then $\epsilon_i = y_i - \hat{y}_i$ represents the ith *residual*, i.e. the difference between the ith observed response value and the ith response value that is predicted by the linear model.

RSS, the *residual sum of squares* is defined as:

$$RSS = \epsilon_1^2 + \epsilon_2^2 + ...\epsilon_n^2$$

or equivalently as:

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + ... + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. With some calculus one gets:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_1 - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

where $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Hypothesis testing

- intro to s.e(beta0) and s.e.(beta1) incl. sampling
- intro to H0 and H1
- group work to calculate s.e.
- live demo in R to run lm

Prediction example

- by hand and live demo

Asesssing the Accuracy of the Model & Correlation

Assumptions