Data analysis with R software

Data-analyysi R-ohjelmistolla

Tommi Härkänen

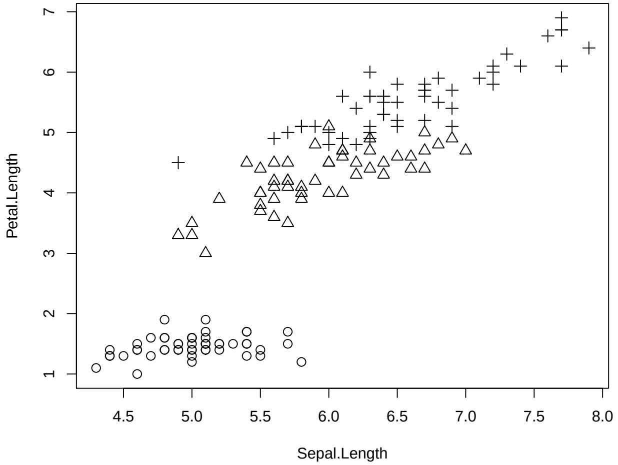
National Institute for Health and Welfare (THL), Helsinki
E-mail: `tommi.harkanen@helsinki.fi`

University of Helsinki, February 7, 2012

Contents

Association of continuous variables

Example: The iris data



Regression modeling

What is the average value of the outcome variable?

A researcher wants to know, what is the association of two (or more) continuous variables.

Simple questions:

- If the researcher measures e.g. sepal length, then what is the **average petal length**?
- How much does the petal length **change** on average, if the measured sepal length increases by 1 cm (unit of measurement)?

More complicated questions:

- Are the associations listed above different for **different species**?
- How well does the model **predict** petal length given sepal length (and possibly other variables)?

Regression modeling

Linear model for one explanatory variable (a.k.a **covariate** or independent variable) x_i for individual $i = 1, 2, \ldots, n$ is often defined as

$$Y_i = \overbrace{\beta_0 + \beta_1 x_i}^{\text{expectation}} + \epsilon_i. \tag{1}$$

The **outcome** variable is Y_i, and the **regression coefficients** are β_0 and β_1.

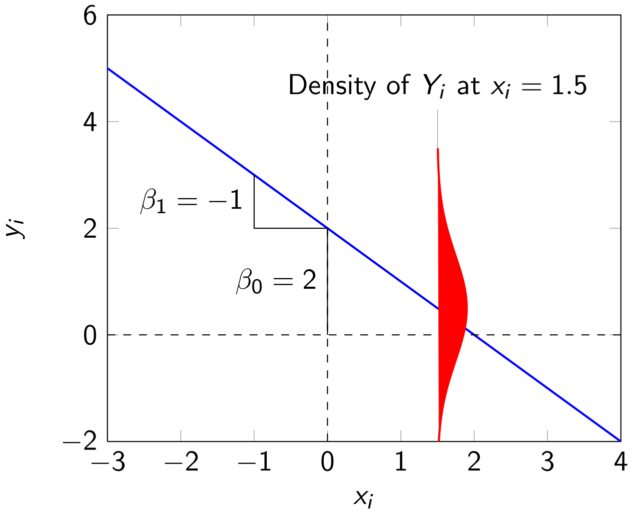
Residual error term is ϵ_i, which is often assumed to be a normally distributed random variable with mean 0 and variance σ^2.

β_0 is called the **intercept** term, which controls the average level of the outcome values. Note that if $x_i = 0$, then the expected value of the outcome is $\mathbb{E}[\, Y_i \mid x_i = 0] = \beta_0$.

β_1 controls the **association** of the outcome and the covariate. Note that if x_i increases by 1 unit, then the outcome value increases by β_1 on average.

Regression modeling

$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i = 2 + (-1) \times x_i + \epsilon_i$ and $\epsilon_i \sim \mathrm{N}(0, \sigma^2)$ where



$\sigma^2 = 1$.

Linear regression modeling in R

The basic command is `lm`:

```
lm(formula, data, subset, ...)
```

Some of the most important options are

formula The model description as a formula: `outcome` ∼ `terms` where **terms** are the covariates separated by '+' and their interactions defined using '*' or ':'.

data Optional data frame, list or environment name.

subset Optional vector specifying a subset of observations.

Example:

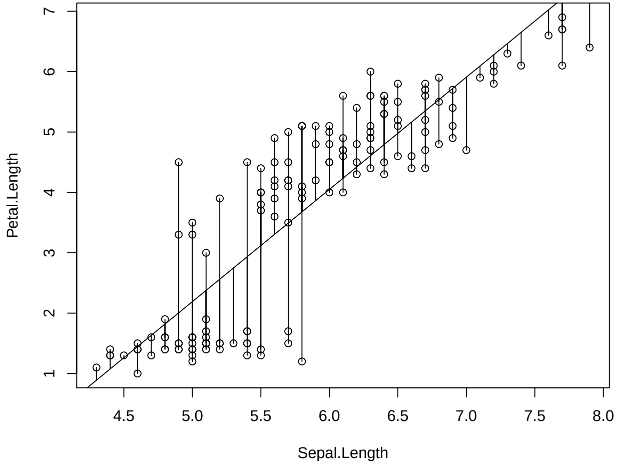
```
> lm(Petal.Length ~ Sepal.Length, data=iris, subset=Species=="setosa")

Call:
lm(formula = Petal.Length ~ Sepal.Length, data = iris, subset = Species ==
    "setosa")

Coefficients:
 (Intercept)  Sepal.Length
      0.8031        0.1316
```

Observed vs. predicted values

Vertical lines are residuals.


Petal.Length
Sepal.Length