Linear regression with RMarkdown

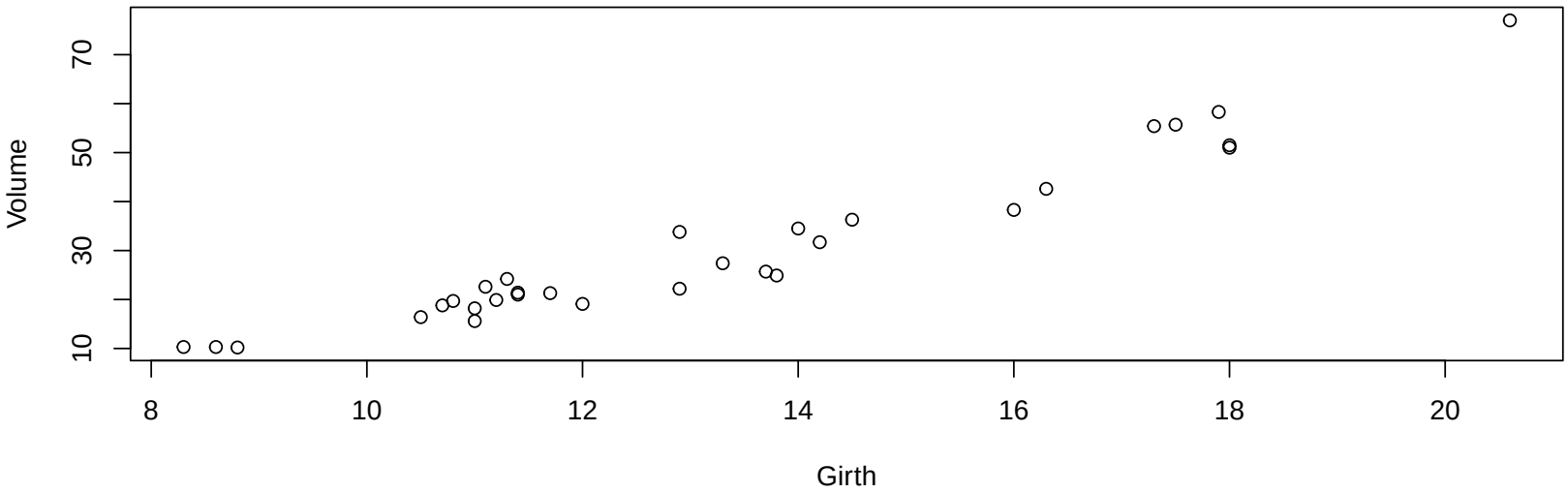
Laurent DELSOL

14 décembre 2018

Motivating example (dataset **trees**) :

How does **Volume** depend on **Girth**?

$$Volume \stackrel{?}{=} a_0 \times Girth + b_0 + \epsilon$$



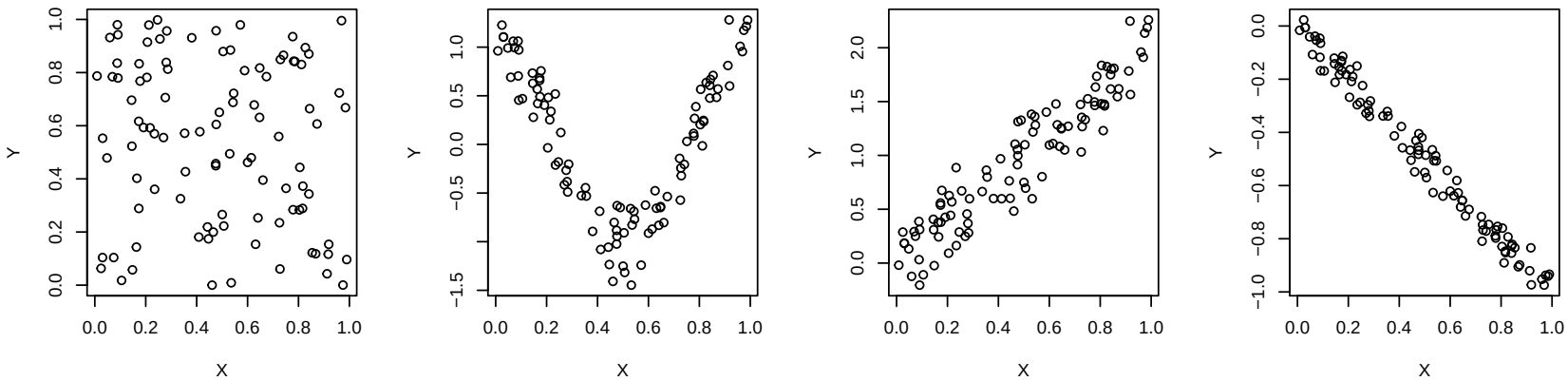
Problem formulation

- Regression problem : $Y = f(X) + \epsilon$
  - Y response variable
  - X explanatory variable
  - f unknown link function
  - ϵ error term
- Does Y depend on X? What kind of link is it? How to estimate the function f?
- Simplest case: Linear regression

$$Y = a_0 \times X + b_0 + \epsilon$$

A. Scatter plot

- Does Y depend on X?
- How intense is the link?
- Is the potential link linear?



B. Correlation coefficient

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{S_X^2 S_Y^2}} (\approx 0.9671194 \text{ in our example})$$

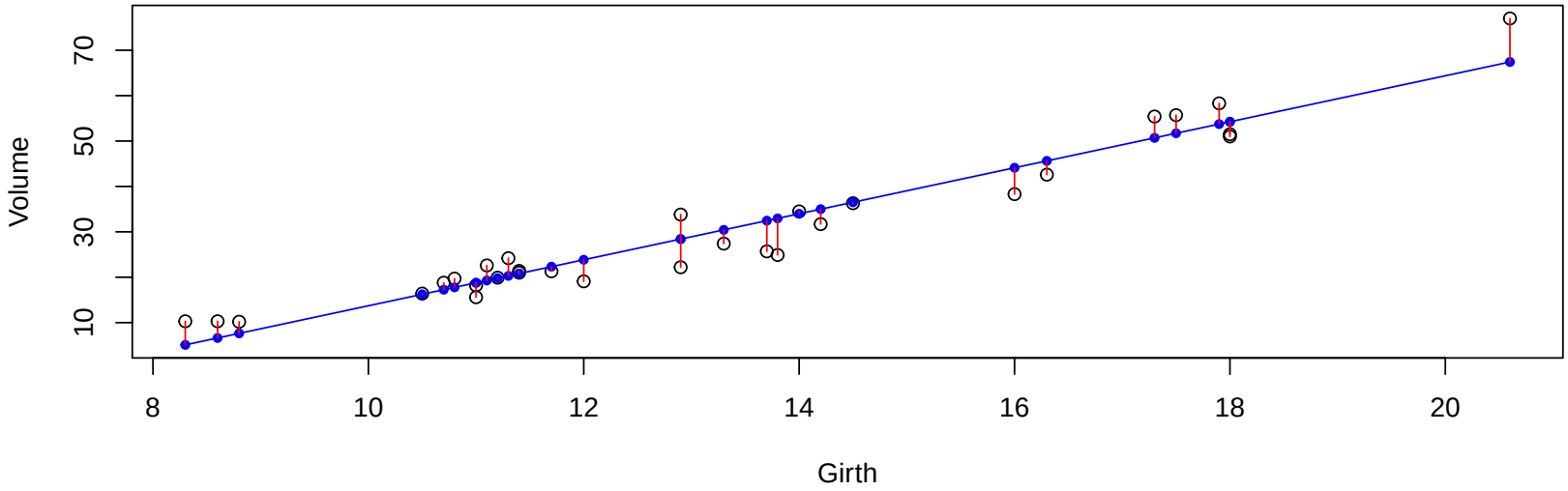
Properties

- $-1 \leq \rho(X, Y) \leq 1$
- $|\rho(X, Y)| = 1 \Leftrightarrow \exists a, b \in \mathbb{R}, \forall i \in \{1, \ldots, n\}, y_i = ax_i + b$
- Hence a strong linear link gives $|\rho(X, Y)|$ close to 1.
- The sign of $\rho(X, Y)$ is related to the way Y and X tend to evolve jointly (positively or negatively).

Regression line estimation

$$Y = a_0 X + b_0 + \epsilon_{(a_0, b_0)}$$

We look for $(\hat{a}_0, \hat{b}_0)$ minimizing the following least squares criterion:
$MC(a, b) = \sum_{i=1}^{n} (y_i - (ax_i + b))^2 = \sum_{i=1}^{n} \left(\epsilon_{i,(a,b)}\right)^2$



Regression line estimation

- $\hat{a}_0 = \frac{cov(X,Y)}{S_X^2} = \rho(X, Y)\frac{S_Y}{S_X} (\approx 5.07$ in our example)

Regression line estimation

- $\hat{a}_0 = \frac{cov(X,Y)}{S_X^2} = \rho(X, Y)\frac{S_Y}{S_X} (\approx 5.07$ in our example)
- $\hat{b}_0 = \overline{Y} - \hat{a}_0\overline{X} (\approx -36.9$ in our example)

Regression line estimation

- $\hat{a}_0 = \frac{cov(X,Y)}{S_X^2} = \rho(X, Y)\frac{S_Y}{S_X}$ (≈ 5.07 in our example)
- $\hat{b}_0 = \overline{Y} - \hat{a}_0\overline{X}$ (≈ -36.9 in our example)
- Modelling quality : coefficient of determination

$$R^2(X, Y) = \frac{\sum_{i=1}^{n}(\hat{a}_0 X_i + \hat{b}_0 - \overline{\hat{a}_0 X + \hat{b}_0})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = (\rho(X, Y))^2$$

In our example: $R^2(X, Y) \approx 0.9353199$

Regression line estimation

- $\hat{a}_0 = \frac{cov(X,Y)}{S_X^2} = \rho(X, Y)\frac{S_Y}{S_X}$ (≈ 5.07 in our example)
- $\hat{b}_0 = \overline{Y} - \hat{a}_0\overline{X}$ (≈ -36.9 in our example)
- Modelling quality : coefficient of determination

$$R^2(X, Y) = \frac{\sum_{i=1}^{n}(\hat{a}_0 X_i + \hat{b}_0 - \overline{\hat{a}_0 X + \hat{b}_0})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = (\rho(X, Y))^2$$

In our example: $R^2(X, Y) \approx 0.9353199$

- Estimation of the mean value of Volume when Girth=12.9

$$\hat{Volume} = 5.0658564 \times 12.9 + -36.9434591 \approx 28.4060887$$

Adding a third variable?

