Introductory examples

Jean-Michel Marin

University of Montpellier
Faculty of Sciences

HAX912X - 2023/2024

Simple linear regression

Let's start with a simple illustrative example. During an experiment in 1849, botanist Joseph Dalton Hooker measured the atmospheric pressure p_i and the boiling temperature of the water y_i in various parts of the Himalayas.

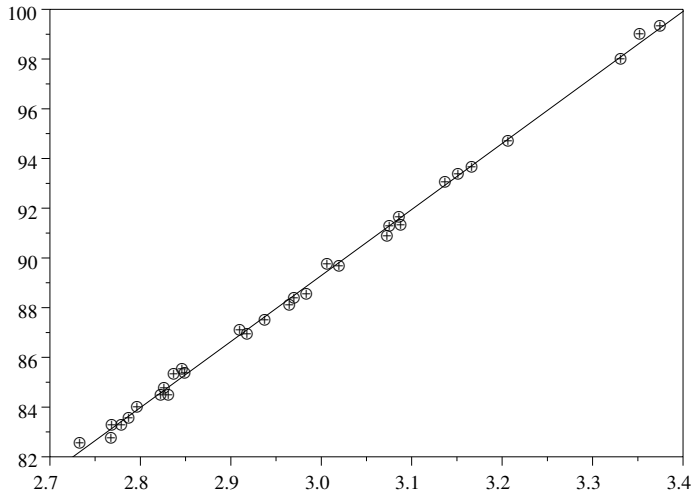
According to the laws of physics, y_i should be (at first approximation) proportional to the logarithm of p_i. We therefore posit the model

$$y_i = \beta_1 + \beta_2 x_i + u_i, \qquad x_i = \log(p_i)$$

u_i represents measurement error

Simple linear regression

Water bulb temperature measured at various locations in the Himalayas as a function of the logarithm of pressure



Simple linear regression

This figure also shows the straight line estimated by least squares

We see a very good fit

If u_i is assumed to be Gaussian, we have the parametric model $y_i \sim \mathcal{N}(\beta_1 + \beta_2 x_i, \sigma^2)$

The parameter σ^2 represents the variance of the points on the right (measured vertically) and the estimate of σ here gives 0.2

Simple linear regression

This example illustrates how the regression model attempts to explain a quantity y (the response) as a function of other quantities x (vector of explanatory variables, or regressors, or factors, one in the example) by separating the deterministic from the random and quantifying these two aspects by the β_i on the one hand and σ^2 on the other hand

Multiple linear regression

Consider the variables, each relating to the totality of the United States (i being the year index)

- P_i : production
- K_i : capital (value of factories, etc.)
- T_i : work done (based on a calculation of the total number of workers)

We seek to explain P_i using the variables (K_i, T_i)

Multiple linear regression

The Cobb and Douglas model is

$$P = \alpha_1 K^{\alpha_2} T^{\alpha_3}$$

which suggests the statistical model

$$\log(P_i) = \log(\alpha_1) + \alpha_2 \log(K_i) + \alpha_3 \log(T_i) + u_i,$$

$$E[u_i] = 0, \quad E[u_i^2] = \sigma^2$$

The regressors are $x_i = (1, \log(K_i), \log(T_i))$, the response is $y_i = \log(P_i)$ and the parameters $\beta = (\log(\alpha_1), \alpha_2, \alpha_3)$

Multiple linear regression

The logarithm and changes in variables have made it possible to render the model linear (with respect to β)

$$y_i = \beta_1 + \beta_2 \log(K_i) + \beta_3 \log(T_i) + u_i$$

Cobb and Douglas had observations over $n = 24$ years (from 1899 to 1922) and determined estimates of β_2 and β_3 : approximately 1/4 and 3/4 respectively

Non linear parametric regression

We observe pairs $(x_i, y_i)_{1 \leqslant i \leqslant n}$ where y_i is the concentration of active ingredient in a drug at time x_i after manufacture

The linear model $y_i = \beta_1 + \beta_2 x_i + u_i$ is certainly inadequate

We start from a specific model considered to be realistic

$$y_i = \beta_1 e^{-\beta_2 x_i} + u_i$$

It's the analogue of the previous one in a non linear situation

Polynomial regression

Same example as above, we observe pairs $(x_i, y_i)_{1 \leqslant i \leqslant n}$. where y_i is the concentration of active ingredient in a drug at time x_i after manufacture

This time, we start with an abstract parametric model

$$y_i = \sum_{j=0}^{J} \beta_j x_i^j + u_i$$

where J is assumed to be known. The linearity in β of this equation means that that β_j is easily estimated by least squares : this is exactly the same as finding the polynomial of degree J that passes as close as possible to the points (x_i, y_i)

Non parametric regression

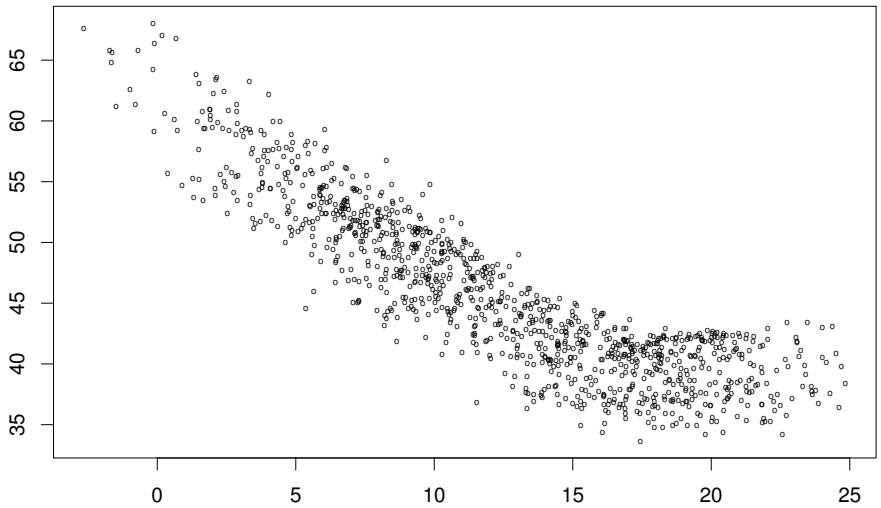
Same example as above, we observe pairs $(x_i, y_i)_{1 \leqslant i \leqslant n}$ where y_i is the concentration of active ingredient in a drug at time x_i after manufacture

$$y_i = f(x_i) + u_i, \quad u_i \sim \mathcal{N}(0, \sigma^2)$$

This involves estimating the function f and $sigma^2$.

Piecewise linear regression

Average electricity consumption in France, at 2am, based on outdoor temperature (average over previous 24 hours), data for 3 years (1095 points)



We might be tempted to consider a piecewise linear model here

Logistic regression

For a bank, this involves measuring the risk it takes on in assigning a credit to a customer.

The bank has data on its former customers. Each customer who has requested a credit in the past is an individual, and the answer $y \in \{0, 1\}$ is a variable indicating whether there has been a repayment problem

The regressor x is a vector containing

- quantitative variables : income, age...
- qualitative variables : gender,

Logistic regression

The logistic model : y is a Bernoulli random variable $\mathcal{B}(1, p_x)$ with

$$p_x = \frac{1}{1 + e^{-x\beta}}$$

where β is a vector column of parameters characterizing the influence of each regressor on the response (so that $x\beta$ is a scalar product)

p_x represents the risk taken by the bank to authorize a credit to a customer with x regressors

Goals

Regression can be seen as the simplest framework for the parametric modelling of sequences of non-stationary random variables

In practice, the main applications are as follows

- **Prediction/Simulation** of responses knowing the regressors
- Determination of **significant factors**

As we've just seen, the method involves setting up a more or less realistic model, on which it's a good idea to stand back