Analysis for CS, Winter semester 2013-2014

Course 11:

An application of local extrema of real-valued functions of several variables

Regression models

are used for

- the determination of model parameters,
- model fitting,
- assessing the importance of influencing factors,
- prediction

in all areas of human, natural and economic sciences.
$\hookrightarrow$ Computer scientists who work closely with people from these areas will definitively come across *regression models*.

The problem

Consider pairs of data

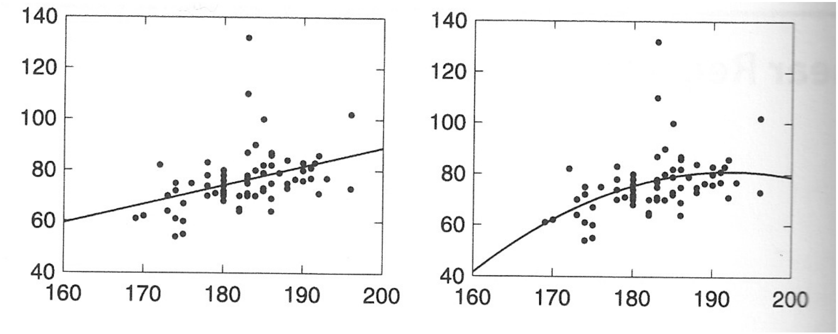
$$(x_1, y_1), \dots, (x_n, y_n)$$

obtained as observations or measurements.
Ex: x_i=height, y_i=weight of each of the 1st year CS students at the UBB
Geometrically they form a scatter plot in the plane.

$$\Downarrow$$

Find a function whose graph represents the scatter plot as closely as possible.

Regression models: line of best fit, best parabola



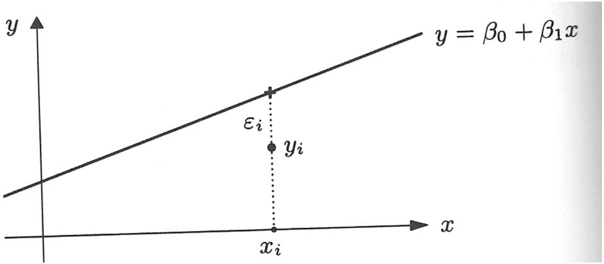
Setting up this model

The postulated relationship between x and y is linear

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

In general, the given data will not exactly lie on a straight line but deviate by ε_i

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$



Minimising the sum of squares of the errors

Define $f\colon \mathbb{R}^2 \to \mathbb{R}$ by

$$f(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

We look for a (global) minimum $(\widehat{\beta}_0, \widehat{\beta}_1)$ of f.

The stationary points of f

$$\begin{cases} \frac{\partial f}{\partial \beta_0}(\widehat{\beta}_0, \widehat{\beta}_1) = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \\ \frac{\partial f}{\partial \beta_1}(\widehat{\beta}_0, \widehat{\beta}_1) = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0. \end{cases}$$

The stationary points of f

$$\begin{cases} n\widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i\right)\widehat{\beta}_1 = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)\widehat{\beta}_0 + \left(\sum_{i=1}^{n} x_i^2\right)\widehat{\beta}_1 = \sum_{i=1}^{n} x_i y_i. \end{cases}$$

Assume that at least two x-values in the data set (x_i, y_i), $i \in \{1, \dots, n\}$ are different. (This is not a restriction.)

$$\Downarrow$$

$$\widehat{\beta}_0 = \left(\frac{1}{n}\sum y_i\right) - \left(\frac{1}{n}\sum x_i\right)\widehat{\beta}_1, \quad \widehat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2}.$$

Note that: $\left(\sum x_i\right)^2 < n\sum x_i^2$.

The Hessian matrix

$$H_f(\widehat{\beta}_0, \widehat{\beta}_1) = \begin{pmatrix} 2n & 2\sum x_i \\ 2\sum x_i & 2\sum x_i^2 \end{pmatrix}$$

is positive definite $\Rightarrow (\widehat{\beta}_0, \widehat{\beta}_1)$ is a local minimum of f.

Local minima $\Rightarrow$ global minima

Let $\emptyset \neq M \subseteq \mathbb{R}^n$ be open and let $f \in C^2(M)$. If $H_f(x)$ is positive definite for all $x \in M$, then every local minimum of f is actually a global one.

$$\Downarrow$$

Solution

The predicted regression line $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is the line of best fit through the scatter plot.

The predicted regression line

The values predicted by the model are

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i,\ \ i \in \{1, \ldots, n\}.$$

The deviations from the values y_i are called residuals

$$e_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i,\ \ i \in \{1, \ldots, n\}.$$