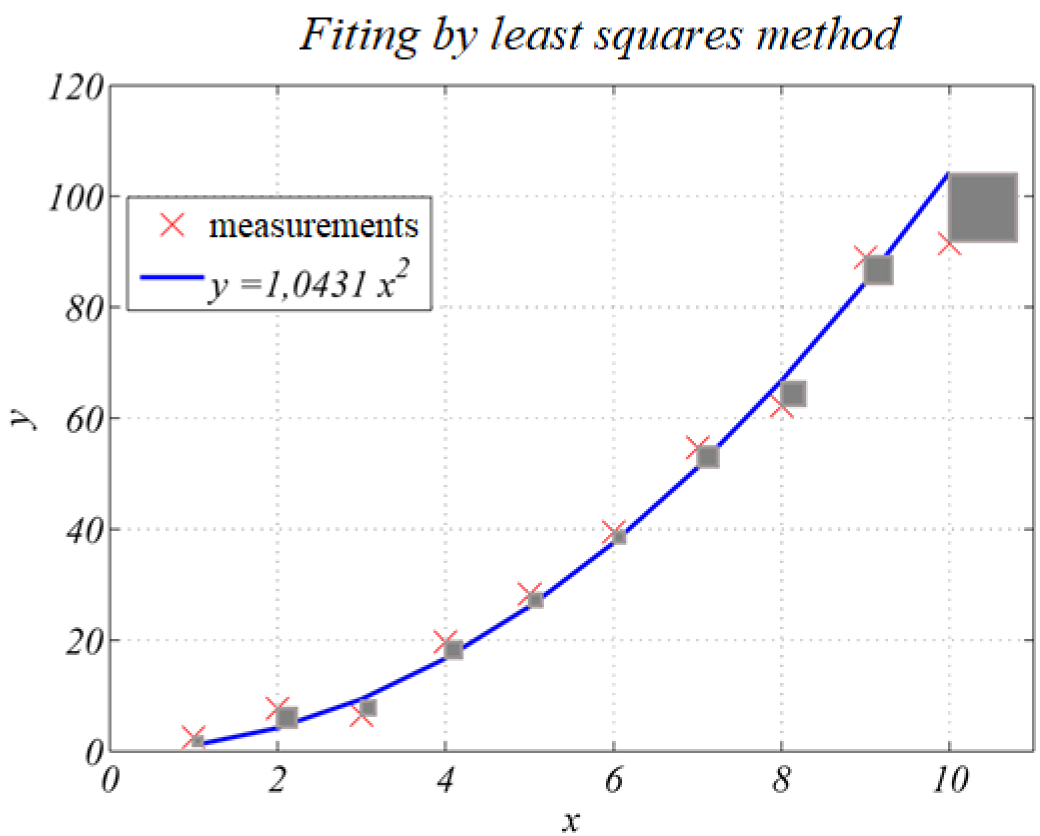
The method of least squares

Dr. Miklós BERTA
Department of Pfysics and Chemistry
Széchenyi István University

SZÉCHENYI ISTVÁN EGYETEM
SZE - GYŐR

Fizika és Kémia Tanszék

The basic principle of the method



$$\Phi(f_{opt.}) = \sum_{i=1}^{N} (y_i - f_{opt.}(x_i))^2 = \min .$$

Let's find the optimal function $y = f_{opt}(x)$ for which the sum of the squared deviations is the smallest one!

Parametrization

- Except the independent variable, the optimal function f_{opt} usually contains also a few more parameters.

$$y = f(x, p_1......p_M)$$

- Determining the optimal function in this case means determining the optimal parameters.
- The optimal parameters are determined based on the "*normal - equations*".

$$\frac{\partial \Phi}{\partial p_i} = 0, \forall i = 1...M$$

Linear regression
The case of proportionality

$$y = m.x$$

- Only the parameter *m* occurs here.
- Sum of squared deviations:

$$\Phi(m) = \sum_{i=1}^{N} (y_i - m.x_i)^2 = \sum_{i=1}^{N} y_i^2 - 2m\sum_{i=1}^{N} x_i y_i + m^2 \sum_{i=1}^{N} x_i^2$$

- The normal – equation is:

$$\frac{d\Phi}{dm} = 2m\sum_{i=1}^{N} x_i^2 - 2\sum_{i=1}^{N} x_i y_i = 0$$

- The optimal parameter is:

$$m^* = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2}$$

- The optimal parameter is calculated on the base of measu-red values with uncertainty, so it is also a *random variable*, so in addition to its *expected value*, it will also have a *standard deviation*.
- It can be proven that its *expected value itself is the optimal value derived from the normal equation*, while its *standard deviation* can be determined as follows:

$$S_0 = \sum_{i=1}^{N} (y_i - m^* x_i)^2$$

$$s = \sqrt{\frac{S_0}{N-1}}$$

$$s_{m^*} = \frac{s}{\sqrt{\sum_{i=1}^{N} x_i^2}}$$

Linear regression
The case of a general straight line

$$y = m.x + b$$

- In this case, two parameters occur, *m* - slope and *b* - offset
- Sum of squared deviations:

$$\Phi(m,b) = \sum (y_i - m.x_i - b)^2$$

- The normal – eqations:

$$\frac{\partial \Phi}{\partial m} = 0 \Rightarrow m\sum x_i^2 + b\sum x_i = \sum x_i y_i$$

$$\frac{\partial \Phi}{\partial b} = 0 \Rightarrow m\sum x_i + b.N = \sum y_i$$

- The optimal parameters:

$$m^* = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2}$$

$$b^* = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{N\sum x_i^2 - (\sum x_i)^2}$$

- Standard deviations of optimal parameters:

$$S_0 = \sum (y_i - m^*.x_i - b^*)^2, \quad s = \sqrt{\frac{S_0}{N-2}}$$

$$s_{m^*} = s\sqrt{\frac{N}{N\sum x_i^2 - (\sum x_i)^2}}, \quad s_{b^*} = s\sqrt{\frac{\sum x_i^2}{N\sum x_i^2 - (\sum x_i)^2}}$$

Correlation coefficient

- The correlation coefficient *r* measures how well the series of measured data pairs $[x_i, y_i]$ fits on a straight line:

$$r = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$\bar{x} = \frac{\sum x_i}{N}, \quad \bar{y} = \frac{\sum y_i}{N}$$

- If $|r|=1$, then the data pairs fit perfectly on a straight line. In general, if $|r|>0.8$, the fit is considered as "good".