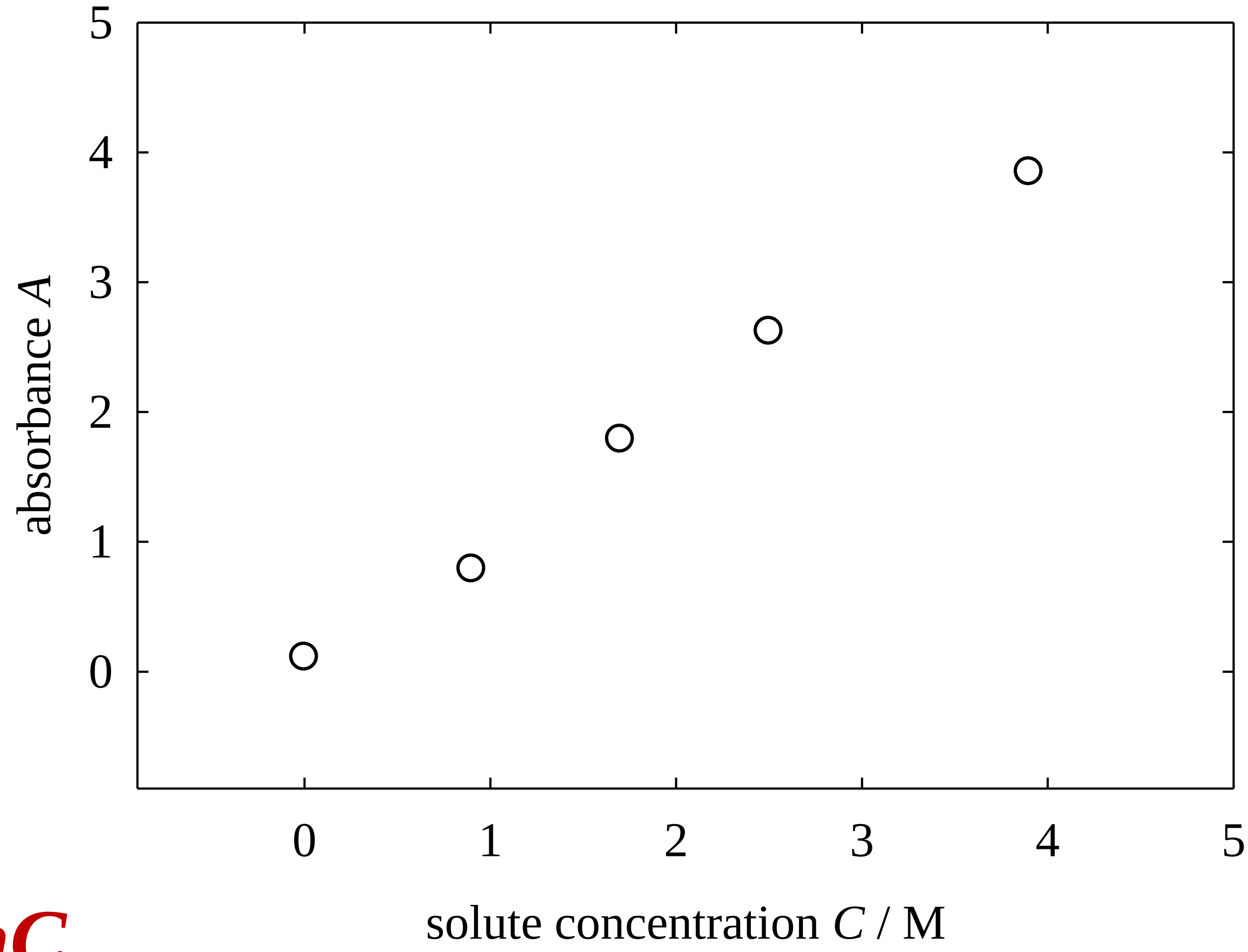
Linear Regression

a useful technique for analyzing experimental data

Suppose you measure the optical absorbance of a solution of a compound at different concentrations. Data:

concentration / M	absorbance
0.00	0.11
0.90	0.79
1.70	1.79
2.50	2.62
3.90	3.85

Plotting the data gives:



Looks like the data can be represented by a straight line:

$$A = b + mC$$

with **slope *m*** and **intercept *b*.** But there are five data points to calculate two unknowns (***m*** and ***b***)

What to Do? Statistics *to the Rescue!*

For N data points represented by the line

$$f(x) = a_0 + a_1 x$$

intercept = a_0 **slope = a_1**

the "best" values of a_0 and a_1 are evaluated from the data by minimizing S, the sum of the squared deviations between the measured and calculated values of $f(x)$.

$$S = \sum_{i=1}^{N} [f(x_i)_{\text{measured}} - f(x_i)_{\text{calculated}}]^2$$

$$= \sum_{i=1}^{N} [f(x_i)_{\text{measured}} - a_0 - a_1 x_i]^2$$

How is the Sum of Squared Deviations Minimized?

by using Partial Derivatives!

$$\left(\frac{\partial S}{\partial a_0}\right)_{a_1} = \left[\frac{\partial}{\partial a_0}\sum_{i=1}^{N}[f(x_i)_{\text{measured}} - a_0 - a_1 x_i]^2\right]_{a_1} = 0$$

$$\left(\frac{\partial S}{\partial a_1}\right)_{a_0} = \left[\frac{\partial}{\partial a_1}\sum_{i=1}^{N}[f(x_i)_{\text{measured}} - a_0 - a_1 x_i]^2\right]_{a_0} = 0$$

Minimizing *S* gives two equations (*try it!*):

$$(-2)\sum_{i=1}^{N}[\,f(x_i)_{\text{measured}} - a_0 - a_1 x_i\,] = 0$$

$$(-2)\sum_{i=1}^{N}[\,f(x_i)_{\text{measured}} - a_0 - a_1 x_i\,]\,x_i = 0$$

Two Equations. Two Unknowns.

$$N a_0 + \left(\sum_{i=1}^{N} x_i\right) a_1 = \sum_{i=1}^{N} f(x_i)_{\text{measured}}$$

$$\left(\sum_{i=1}^{N} x_i\right) a_0 + \left(\sum_{i=1}^{N} x_i^2\right) a_1 = \sum_{i=1}^{N} x_i\, f(x_i)_{\text{measured}}$$

Solve for the intercept a_0 and the slope a_1.

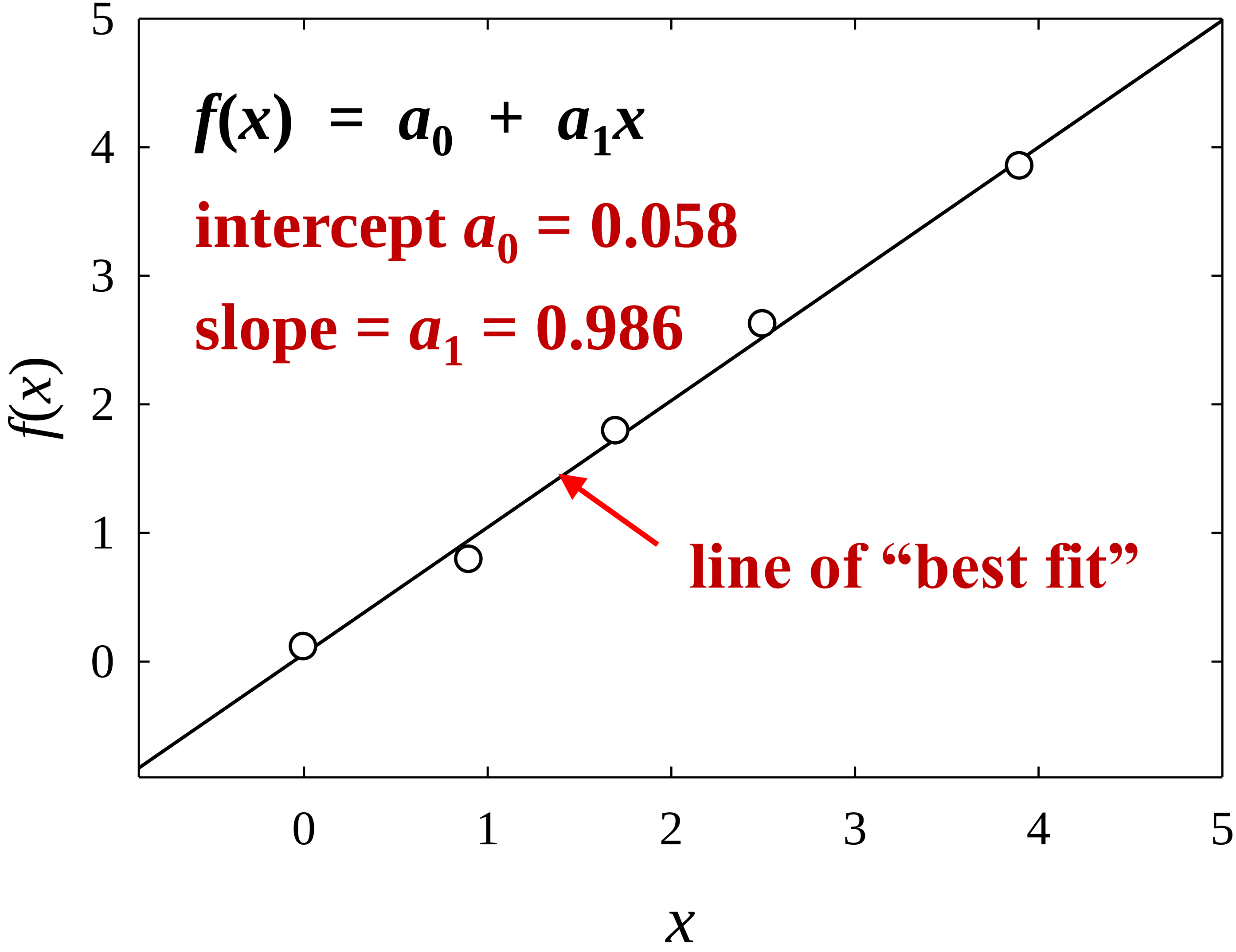
line intercept:

$$a_0 = \frac{\left(\sum_{i=1}^{N} x_i^2\right)\left(\sum_{i=1}^{N} x_i f(x_i)\right) - \left(\sum_{i=1}^{N} f(x_i)\right)\left(\sum_{i=1}^{N} x_i f(x_i)\right)}{N\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}$$

line slope:

$$a_1 = \frac{N\left(\sum_{i=1}^{N} x_i f(x_i)\right) - \left(\sum_{i=1}^{N} x_i\right)\left(\sum_{i=1}^{N} f(x_i)\right)}{N\left(\sum_{i=1}^{N} x_i^2\right) - \left(\sum_{i=1}^{N} x_i\right)^2}$$

IT WORKS!



$f(x) = a_0 + a_1x$
intercept a_0 = 0.058
slope = a_1 = 0.986
line of "best fit"
$f(x)$
x
0
1
2
3
4
5

"Line of Best Fit" (also called **linear regression** and **least-squares**) calculations are widely used in science and technology.

Can be easily extended to more complicated equations, such as

$$f(x) = a_0 + a_1x + a_2x^2 + a_3x^3 + \ldots$$

(multiple linear regression)

$$f(x) = a_0\sin(a_1x)$$

(nonlinear regression)