Notes on data fitting and regression

Linear Regression

Is there a linear relation between “x” and “y”

It is always a good idea to make a plot of x vs. y.

Not all relations between x and y will be linear. If they are not, you can **linearise** the relation first.

In either case:

To determine the linear dependency between “x” and “y”, we carry out a **linear regression**

Linear dependent quantities

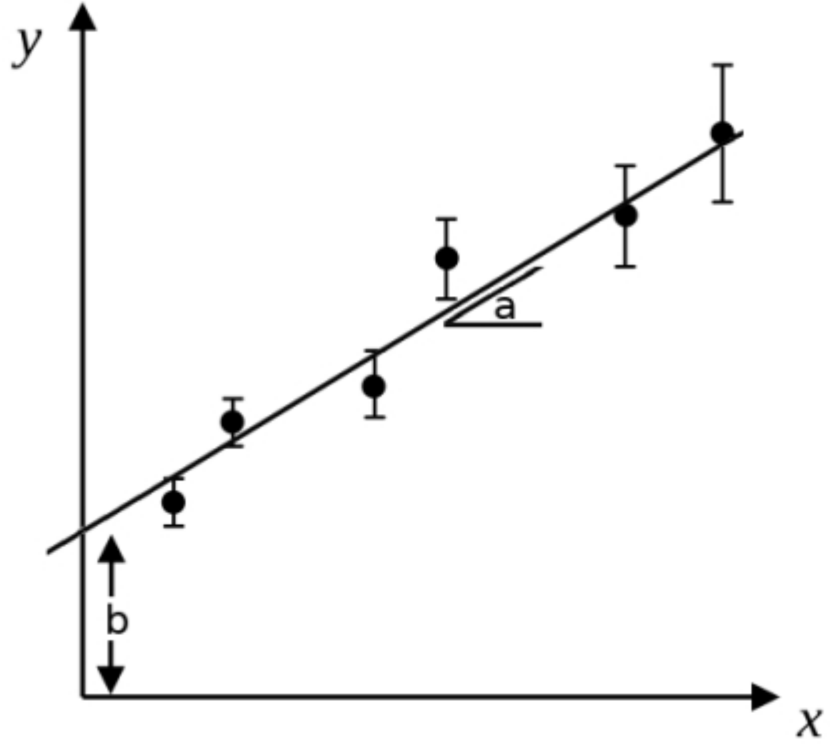
Remember that:

$$y = a \cdot x + b$$

$a: slope$

$b: axis\ intercept$

We need to find “a” and “b” for a straight line that fits best the data points.



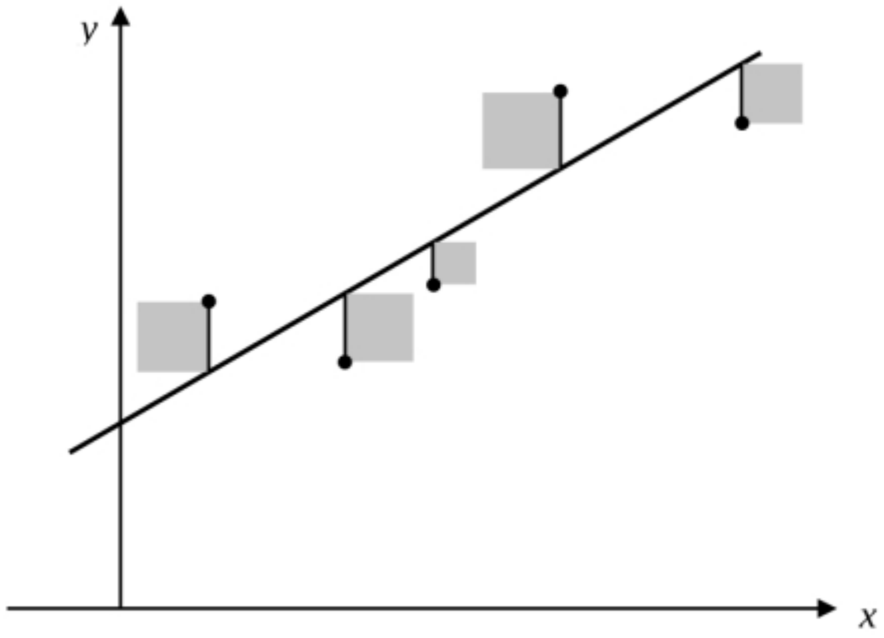
Background: least mean square

We need to minimise this expression (i.e. get the smallest sum of the grey areas)

$$\sum_{i=1}^{n}\left[y_i-\left(a\cdot x_i+b\right)\right]^2$$

by varying “a” and “b”

(partial derivatives set to 0)



Formulae I: solution

The best “a” and “b” are

$$a=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\cdot\left(y_i-\overline{y}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2} \qquad b=\overline{y}-a\cdot\overline{x}$$

with $\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$ and $\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$

Formulae II: errors

Error of the fit

$$\sigma_y = \sqrt{\frac{1}{(n-2)} \sum_{i=1}^{n} \left[y_i - (a \cdot x_i + b) \right]^2}$$

Denominator (n-2) as we are fitting 2 parameters

Errors for “a” and “b”

$$\sigma_a = \sigma_y \cdot \sqrt{\frac{1}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

$$\sigma_b = \sigma_y \cdot \sqrt{\frac{1}{n} \cdot \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}}$$

Linear regression: example

Determine the thickness of pages in a book

Measure the thickness “d” of different number “m” of pages (including the cover)

$$d_i = a \cdot m_i + b$$

The slope is the thickness of a page, and the axis intercept is the thickness of the cover.

m	d (mm)
10	3.2
20	4.2
30	5.1
40	5.8
50	6.8
60	7.7
70	8.8
80	9.7
90	10.8
100	11.7

Linear regression: example

m	d (mm)
10	3.2
20	4.2
30	5.1
40	5.8
50	6.8
60	7.7
70	8.8
80	9.7
90	10.8
100	11.7

1. Make a plot and devise a model: $d_i = a \cdot m_i + b$

2. Calculate averages of "m" and "d".

3. Calculate the slope and intercept with:

$$a = \frac{\sum_{i=1}^{n} (x_i - \overline{x}) \cdot (y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \qquad b = \overline{y} - a \cdot \overline{x}$$

a=0.0943 mm

b=2.1935 mm

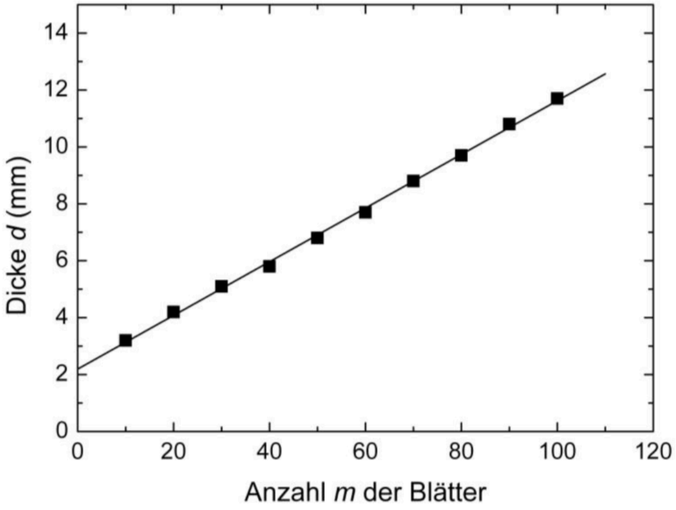
Linear regression: example

m	d (mm)
10	3.2
20	4.2
30	5.1
40	5.8
50	6.8
60	7.7
70	8.8
80	9.7
90	10.8
100	11.7

$$a=\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\cdot\left(y_i-\bar{y}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \qquad b=\bar{y}-a\cdot\bar{x}$$

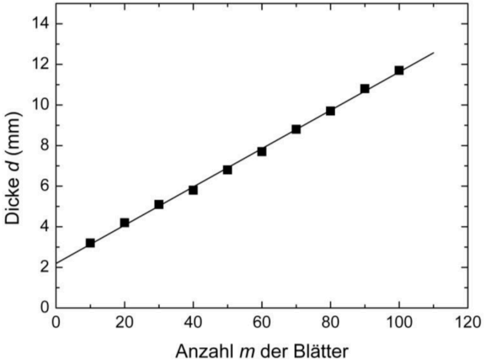
a=0.0943 mm

b=2.1935 mm



d=0.0943 m + 2,1935 [mm]

Linear regression: example



$d=0.0943\ m + 2{,}1935$ [mm]

4. Calculate the errors for the slope and intercept with our equations in slide 40:

$\sigma_d= 0.173$ mm, $\sigma_a=0.0019$ mm, and $\sigma_b=0.12$ mm.

5. We report "a" and "b" with their errors:

$a=(0.0943\pm0.0019)$ mm (average thickness of a sheet in the book)

$b=(2.19\pm0.12)$ mm (thickness of the book back cover)

Book thickness $d = a\cdot m+2b$