Linear regression

- In order to test a linear relationship between two observables need to plot one against the other and fit a line.
- This is called *linear regression* or *least-squares fit*.
- Can also be used to validate non-linear relationships after casting them into linear form (e.g. exponentials appear linear on semilog plots, see lecture 2)
- The method of least squares can also be applied for non-linear regression (fitting polynomials).

Linear regression

- Suppose we have N data points (x_i, y_i) with an uncertainty σ_y for each y_i .

<u>Assumptions</u>

- We expect y_i and x_i to be connected via a linear relationship $y_i = A + Bx_i$
- Assume that error bar σ_y is the same for all y_i.
- Assume no uncertainty in x, i.e. $\sigma_x = 0$. (No horizontal error bars)
- Assume that uncertainty in the y_i's is governed by the normal distribution.

Linear regression

Two main questions:

- What are the values of A and B that give the best line fit ?
- How good is the line fit (“goodness of fit”) ? Does the data support a linear relationship between x and y ?

We will use principle of maximum likelihood to find the linear fit parameters A and B.

Furthermore, will define a *figure of merit,* χ^2 (chi squared), that reflects the goodness of the fit.

Maximum likelihood

Each of the measurements y_i is governed by a Gaussian distribution around the true value $A + Bx_i$, where A and B are the fit parameters we want to determine.

The probability of obtaining the observed value y_i is just

$$P(y_i) \propto \frac{1}{\sigma_y} e^{-(y_i - A - Bx_i)^2/2\sigma_y^2}$$

The probability of obtaining a complete set of N measurements y_i is just

$$P(y_1, y_2, .., y_N) = P(y_1)P(y_2)..P(y_N) \propto \frac{1}{\sigma_y} e^{-\chi^2/2}$$

χ^2 - Chi squared

$$P(y_1, y_2, .., y_N) = P(y_1)P(y_2)..P(y_N) \propto \frac{1}{\sigma_y^N} e^{-\chi^2/2}$$

The exponent is the so-called Chi-squared χ^2:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

χ^2 is the sum of the squared differences between observed and expected values (here it is $A + Bx_i$) normalized by the variance of the observables.

If there is perfect agreement then $\chi^2 = 0$. If the observed $y_i's$ are on average one σ_y away from the expected values, then $\chi^2 \approx N$

Least-squares fitting

According to the principle of maximum likelihood, want to maximise the probability $P(y_1, y_2, .., y_N) \propto \frac{1}{\sigma_y^N} e^{-\frac{\chi^2}{2}}$ with respect to the free parameters, which are A and B in our case.

This probability is maximum when the exponent, $\propto \chi^2$ is a minimum:

Therefore, need to solve

$$\frac{\partial \chi^2}{\partial A} = 0 \quad \text{and} \quad \frac{\partial \chi^2}{\partial B} = 0$$

Least-squares fitting

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

$$\frac{\partial \chi^2}{\partial A} = \frac{-2}{\sigma_y^2} \sum_{i=1}^{N} (y_i - A - Bx_i) = 0$$

$$\frac{\partial \chi^2}{\partial B} = \frac{-2}{\sigma_y^2} \sum_{i=1}^{N} x_i(y_i - A - Bx_i) = 0$$

Least-squares fitting

This can be written as simultaneous equations for A and B

$$AN + B\sum x_i = \sum y_i$$

$$A\sum x_i + B\sum x_i^2 = \sum x_i y_i$$

Here we use the shorthand $\sum \quad \equiv \quad \sum_{i=1}^{N}$

Solving for A and B is a straightforward problem of linear algebra. Two linear equations with two unknowns.

Least square fitting

Finally we obtain

$$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{\Delta}$$

$$B = \frac{N \sum xy - \sum x \sum y}{\Delta}$$

where

$$\Delta = N \sum x^2 - \left(\sum x\right)^2$$

Note that these estimates for A and B **do not depend** on the error bars of the $y_i's$, namely σ_y.

Example – measuring Hooke's law

Objective: Obtain spring constant from five measurements with different masses:

$$l = A + Bm$$

Here, $x_i \leftrightarrow m_i$ and $y_i \leftrightarrow l_i$

Table 8.1. Masses m_i (in kg) and lengths l_i (in cm) for a spring balance. The "x" and "y" in quotes indicate which variables play the roles of x and y in this example.

Trial number i	"x" Load, m_i	"y" Length, l_i	m_i^2	$m_i l_i$
1	2	42.0	4	84
2	4	48.4	16	194
3	6	51.3	36	308
4	8	56.3	64	450
5	10	58.6	100	586
$N = 5$	$\Sigma m_i = 30$	$\Sigma l_i = 256.6$	$\Sigma m_i^2 = 220$	$\Sigma m_i l_i = 1{,}622$

Example – measuring Hooke's law

Compute the fit parameters:

$$\Delta = N\sum m^2 - \left(\sum m\right)^2 = 5 \cdot 220 - 30^2 = 200$$

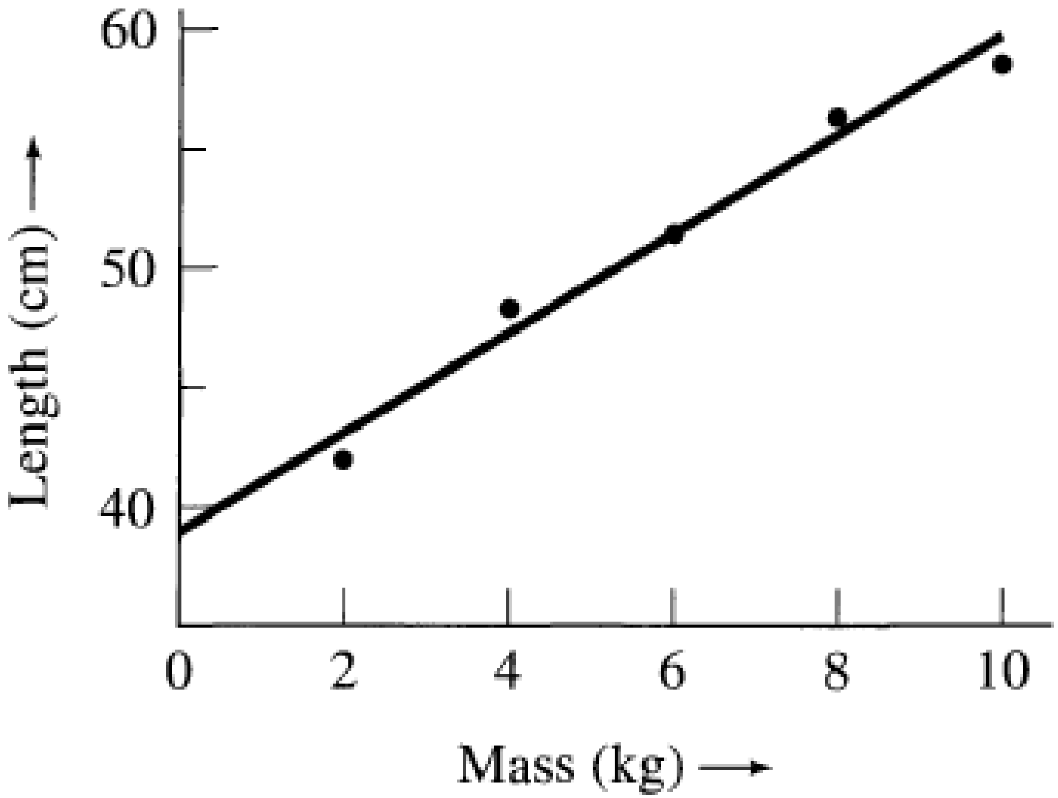
$$A = \frac{\sum m^2 \sum l - \sum m \sum ml}{\Delta} = \frac{220 \cdot 256.6 - 30 \cdot 1622}{200} = 39.0cm$$

$$B = \frac{N\sum ml - \sum m \sum l}{\Delta} = \frac{5 \cdot 1622 - 30 \cdot 256.6}{200} = 2.06\ cm/kg$$

Here, A is the unloaded length of the spring and $B = g/k$, where k is the spring constant ($F = kx \rightarrow x = \frac{F}{k} = \frac{mg}{k} = \left(\frac{g}{k}\right)m$).

Example – measuring Hooke's law

Least squares fit. Note, that you should always plot with error bars. This example just illustrates that error bars are not required to find A and B.



Uncertainties in A and B

- So far the error bars σ_y didn't play any role. They are not needed to find fit parameters A and B.

- However, to estimate the uncertainties in A and B will need to know the uncertainties in the observed y_i's.

- Finding the uncertainties σ_A and σ_B can be achieved through error propagation.

Uncertainties in A and B

$$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{\Delta}$$

$$\Delta = N\sum x^2 - \left(\sum x\right)^2$$

What is σ_A ?

Since $\sigma_x = 0$, there is no uncertainty in Δ which depends on x only.

Likewise, $\sum x^2$ and $\sum x$ have no uncertainties associated with it and can be considered constants.

Uncertainties in A and B

$$A = \left(\frac{\sum x^2}{\Delta}\right)\sum y - \left(\frac{\sum x}{\Delta}\right)\sum xy$$

The terms in the brackets are constants and have no uncertainty as they depend on x only.

From the general error propagation formula, we can calculate σ_A as follows:

$$\sigma_A^2 = \sum_{i=1}^{N}\left(\frac{\partial A}{\partial y_i}\sigma_y\right)^2$$

Uncertainties in A and B

$$\frac{\partial A}{\partial y_i} = \left(\frac{\sum x^2}{\Delta}\right)\frac{\partial(\sum y)}{\partial y_i} - \left(\frac{\sum x}{\Delta}\right)\frac{\partial(\sum xy)}{\partial y_i}$$

$$\frac{\partial(\sum y)}{\partial y_i} = \frac{\partial}{\partial y_i}(y_1 + y_2 + .. + y_N) = 1 \quad for\ all \quad i$$

$$\frac{\partial(\sum xy)}{\partial y_i} = \frac{\partial}{\partial y_i}(x_1 y_1 + x_2 y_2 + .. + x_N y_N) = x_i$$

Therefore,

$$\frac{\partial A}{\partial y_i} = \left(\frac{\sum x^2}{\Delta}\right)\cdot 1 - \left(\frac{\sum x}{\Delta}\right)\cdot x_i$$

Uncertainties in A and B

Finally,

$$\sigma_A^2 = \sum_{i=1}^{N}\left(\frac{\partial A}{\partial y_i}\sigma_y\right)^2 = \sigma_y^2\sum_{i=1}^{N}\left(\left(\frac{\sum x^2}{\Delta}\right) - \left(\frac{\sum x}{\Delta}\right)x_i\right)^2$$

$$\sigma_A^2 = \frac{\sigma_y^2}{\Delta^2}\sum_{i=1}^{N}\left(\sum x^2 - x_i\sum x\right)^2$$

$$\sigma_A^2 = \frac{\sigma_y^2}{\Delta^2}\sum_{i=1}^{N}\left(\left(\sum x^2\right)^2 - 2x_i\sum x\sum x^2 + \left(x_i\sum x\right)^2\right)$$

$$\sigma_A^2 = \frac{\sigma_y^2}{\Delta^2}\left(N\left(\sum x^2\right)^2 - 2\sum x\sum x\sum x^2 + \sum x^2\left(\sum x\right)^2\right)$$

Uncertainties in A and B

$$\sigma_A^2 = \frac{\sigma_y^2}{\Delta^2}\left(N\left(\sum x^2\right)^2 - 2\left(\sum x\right)^2\sum x^2 + \sum x^2\left(\sum x\right)^2\right)$$

$$\sigma_A^2 = \frac{\sigma_y^2}{\Delta^2}\left(N\left(\sum x^2\right)^2 - \left(\sum x\right)^2\sum x^2\right)$$

$$\sigma_A^2 = \frac{\sigma_y^2}{\Delta^2}\sum x^2\left(N\sum x^2 - \left(\sum x\right)^2\right) = \frac{\sigma_y^2}{\Delta^2}\sum x^2\Delta$$

$$\boxed{\sigma_A = \sigma_y\sqrt{\frac{\sum x^2}{\Delta}}}$$

Uncertainties in A and B

Now calculate σ_B

$$B = \frac{N\sum xy - \sum x \sum y}{\Delta}$$

$$B = \left(\frac{N}{\Delta}\right)\sum xy - \left(\frac{\sum x}{\Delta}\right)\sum y$$

Terms in the brackets are constants. Error in B given by

$$\sigma_B^2 = \sum_{i=1}^{N} \left(\frac{\partial B}{\partial y_i}\sigma_y\right)^2$$

Uncertainties in A and B

$$\frac{\partial B}{\partial y_i} = \left(\frac{N}{\Delta}\right)\frac{\partial(\sum xy)}{\partial y_i} - \left(\frac{\sum x}{\Delta}\right)\frac{\partial(\sum y)}{\partial y_i}$$

Already computed

$$\frac{\partial(\sum xy)}{\partial y_i} = \frac{\partial}{\partial y_i}(x_1 y_1 + x_2 y_2 + .. + x_N y_N) = x_i$$

$$\frac{\partial(\sum y)}{\partial y_i} = \frac{\partial}{\partial y_i}(y_1 + y_2 + .. + y_N) = 1 \quad for\ all \quad i$$

Therefore,

$$\frac{\partial B}{\partial y_i} = \left(\frac{N}{\Delta}\right) \cdot x_i - \left(\frac{\sum x}{\Delta}\right) \cdot 1$$

Uncertainties in A and B

Finally,

$$\sigma_B^2 = \sum_{i=1}^{N} \left(\frac{\partial B}{\partial y_i} \sigma_y \right)^2 = \sigma_y^2 \sum_{i=1}^{N} \left(\left(\frac{N}{\Delta} \right) x_i - \left(\frac{\sum x}{\Delta} \right) \right)^2$$

$$\sigma_B^2 = \frac{\sigma_y^2}{\Delta^2} \sum_{i=1}^{N} (N x_i - \sum x)^2$$

$$\sigma_B^2 = \frac{\sigma_y^2}{\Delta^2} \sum_{i=1}^{N} \left((N x_i)^2 - 2N x_i \sum x + (\sum x)^2 \right)$$

$$\sigma_B^2 = \frac{\sigma_y^2}{\Delta^2} \left(N^2 \sum x^2 - 2N \sum x \sum x + N (\sum x)^2 \right)$$

Uncertainties in A and B

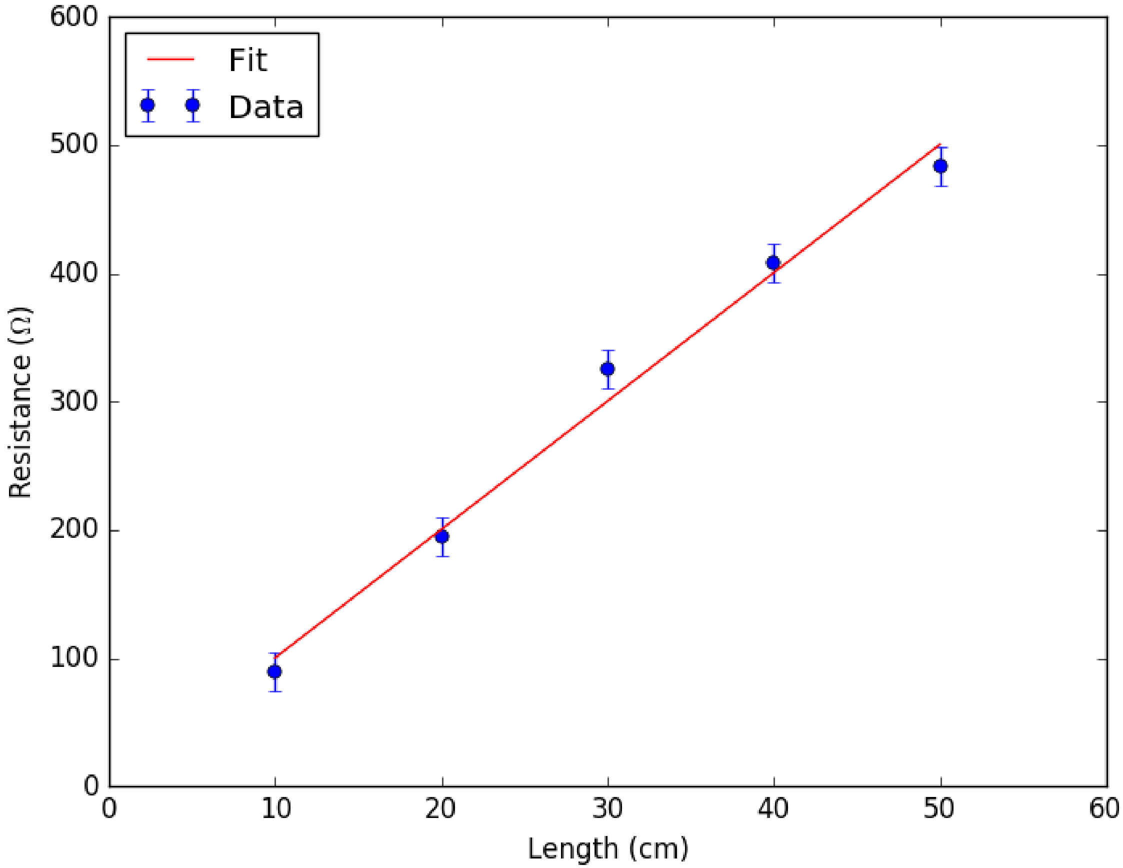
$$\sigma_B^2 = \frac{\sigma_y^2}{\Delta^2}\left(N^2 \sum x^2 - 2N\left(\sum x\right)^2 + N\left(\sum x\right)^2\right)$$

$$\sigma_A^2 = \frac{\sigma_y^2}{\Delta^2}\left(N^2 \sum x^2 - N\left(\sum x\right)^2\right)$$

$$\sigma_A^2 = \frac{\sigma_y^2}{\Delta^2} N\left(N \sum x^2 - \left(\sum x\right)^2\right) = \frac{\sigma_y^2}{\Delta^2} N\Delta$$

$$\boxed{\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}}}$$

Example: Resistance vs length



$$A = 0 \pm 16\ \Omega$$

$$B = 10.0 \pm 0.5\,\frac{\Omega}{cm}$$

χ^2 (Chi squared) – Goodness of fit

- The uncertainties in A and B alone are not sufficient to tell us whether the data follows a linear relationship.

- Need a figure of merit to tell us if the data is consistent with the proposed model. In our case the model is a linear relationship between y and x.

- An often used figure of merit is Chi-squared - the squared difference between the data points and the proposed model:

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

χ^2 (Chi squared) – Goodness of fit

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$

- $\chi^2 \approx N$ if data points are within σ_y of the line fit.

- Can define the so-called reduced Chi squared $\tilde{\chi}^2$:

$$\tilde{\chi}^2 = \frac{\chi^2}{\nu}$$

- Degrees of freedom:

ν = number of data points – independent constraints

χ^2 (Chi squared) – Goodness of fit

- For line fit, $\nu = N - 2$, so

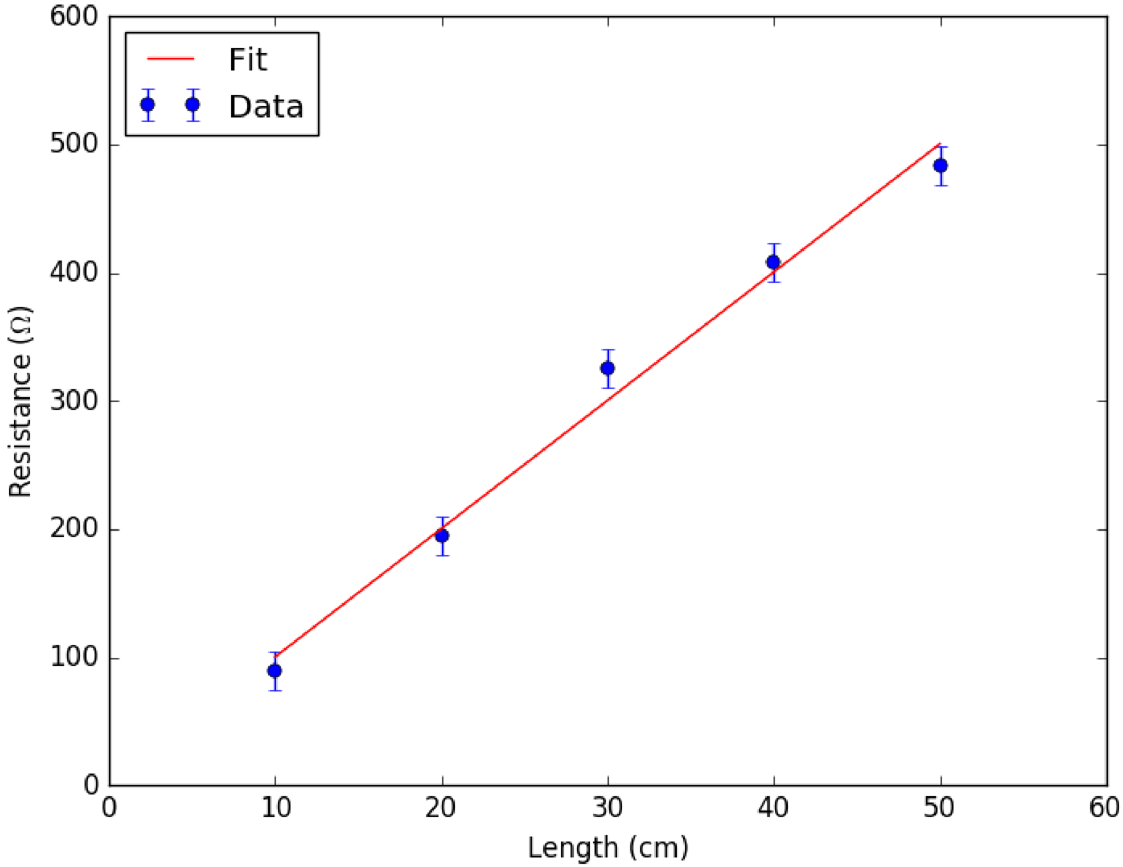
$$\tilde{\chi}^2 = \frac{\chi^2}{N-2}$$

- E.g. line fit for two points (N=2) is meaningless. Can always find a line that goes exactly through the two data points. Therefore $\tilde{\chi}^2$ ill defined for $N \leq 2$

<u>Possible outcomes:</u>

- $\tilde{\chi}^2 \gg 1$: Data cannot be described by the model (in our case: linear fit)
- $\tilde{\chi}^2 > 1$: poor fit, either data not fully captured by model or error bars have been underestimated
- $\tilde{\chi}^2 \approx 1$: good fit consistent with the error bars
- $\tilde{\chi}^2 < 1$: fit is “too good”. Error bars are probably overestimated.

Example: Resistance vs length



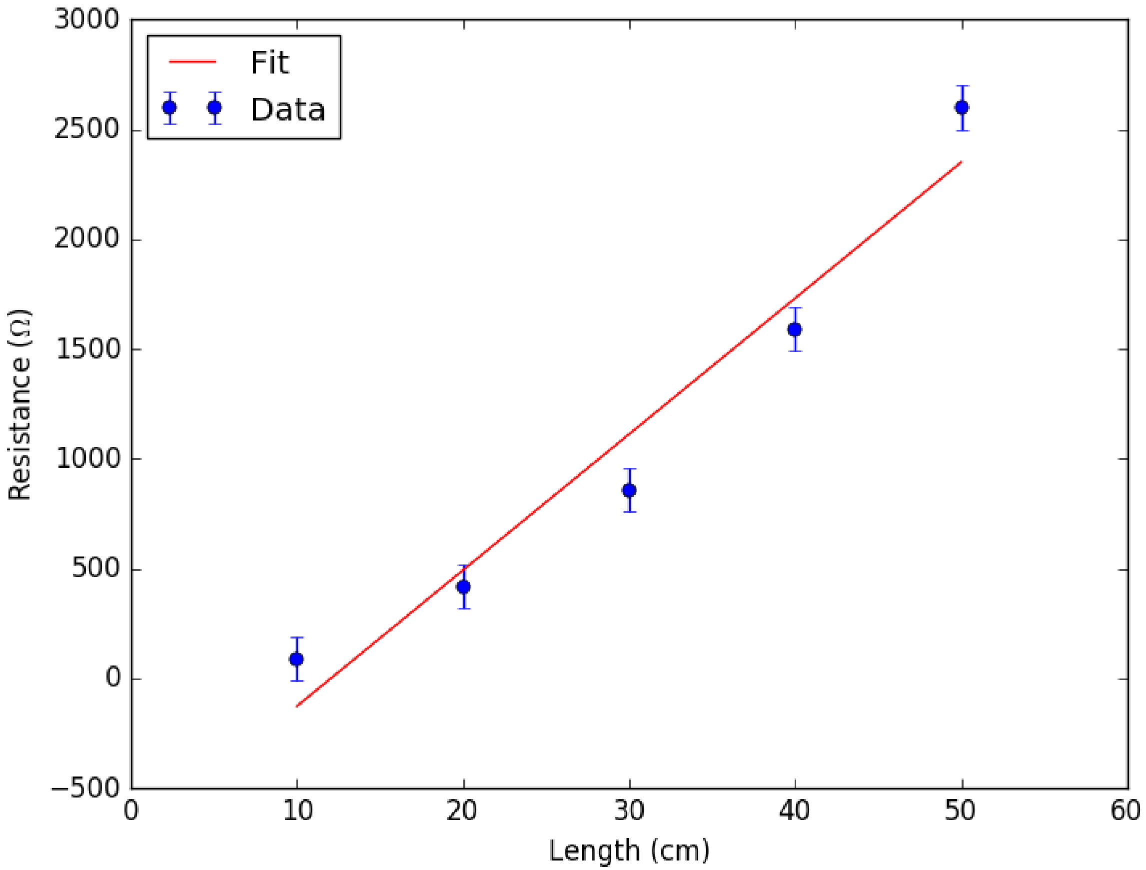
$$A = 0 \pm 16\ \Omega$$

$$B = 10.0 \pm 0.5\,\frac{\Omega}{cm}$$

$$\tilde{\chi}^2 = 1.6$$

Not the best fit, but reasonable. Error bars are slightly underestimated.

Example: bad fit



$$A = -745 \pm 105\ \Omega$$

$$B = 62 \pm 3\,\frac{\Omega}{cm}$$

$$\tilde{\chi}^2 = 6.6$$

Bad fit, does not look like a linear relationship.

Important: Always use common sense! Do not just blindly fit data.
Plot it and see if the fit makes sense on physical grounds.
E.g. resistance should be zero at zero length.
Not the case here ($A = -745\Omega$!!)

Extensions to linear regression

- The error bars on the y_i's are not necessarily the same. In this case each data point is weighted by their respective (error bar)^2. Data points with large error bars have less weight than points with smaller uncertainties.

- In most cases, one of the variables has a much larger fractional error. Can ignore error bars on the other variables. e.g. typical Hookes's law experiment: Mass $\approx 100 \pm 1$ gram, extension 10 ± 1 mm. Fractional error in extension much larger. No need to plot error bars for mass.

- Least square fitting can easily be extended to non-linear fits. e.g. quadratic fit leads to

$$\chi^2 = \sum_{i=1}^{N} \frac{\left(y_i - A - Bx_i - Cx_i^2\right)^2}{\sigma_y^2}$$

- Principle of maximum likelihood will lead to 3 equations with 3 unknowns (A, B, C)

Summary linear regression

Fit $y = A + bx$:

$$A = \frac{\sum x^2 \sum y - \sum x \sum xy}{\Delta}; \qquad B = \frac{N\sum xy - \sum x \sum y}{\Delta}$$

where $\quad \Delta = N\sum x^2 - \left(\sum x\right)^2$

$$\sigma_A = \sigma_y \sqrt{\frac{\sum x^2}{\Delta}}; \qquad \sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}}$$

$$\tilde{\chi}^2 = \frac{1}{N-2} \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}$$