Noise floor example?

Curve fitting

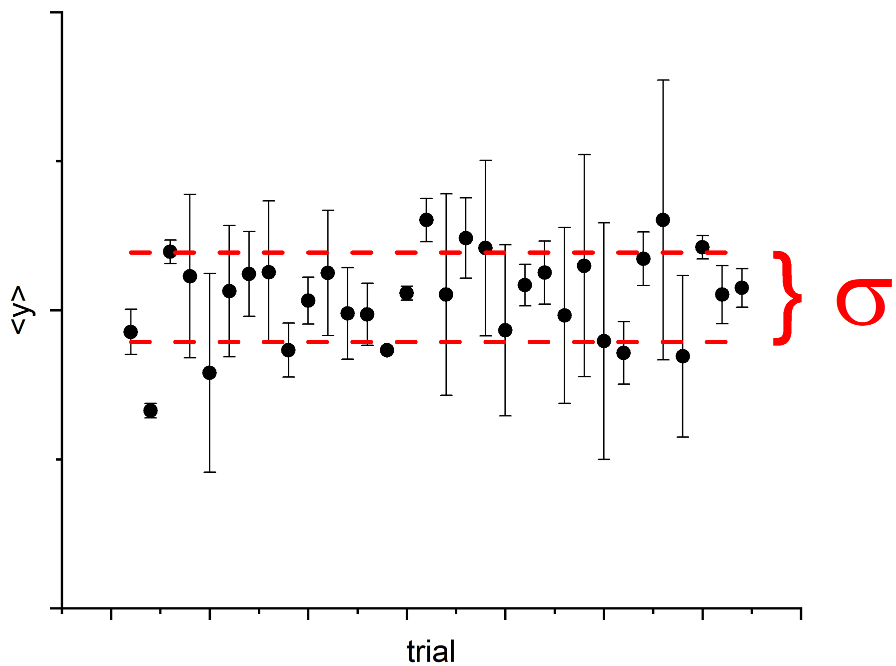
P. LeClair

PH 4/591 Fall 2022

based on material from A. Piepke

Summary statistics

We have studied how to summarize a data set y_i of $i = 1, \dots, N$ independent measurements with a mean and standard deviation:



Mean:

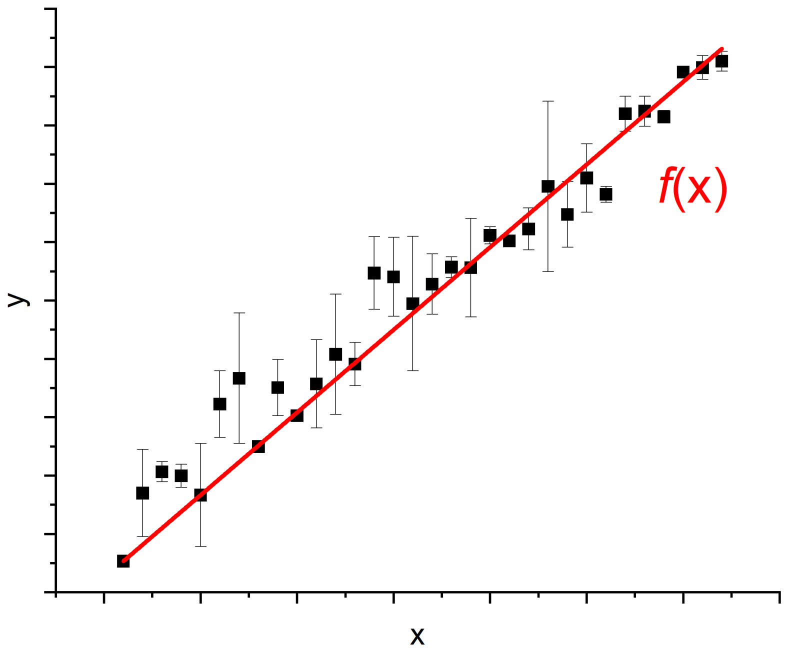
$$\langle y \rangle = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i$$

Standard deviation:

$$s^2 = \frac{1}{N-1} \cdot \sum_{i=1}^{N} (y_i - \langle y \rangle)^2$$

What do we do if the data exhibits a linear correlation? Assume the variation in i goes along with changing some observable x not plotted here.

Linear relationships



This data better described by line

$$y = f(x) = p_2 \cdot x + p_1$$

How do we find the *optimal* p_1 and p_2?

- Assume fluctuations are random (Gaussian), errors symmetric.
- Assume the data represents the *most likely* outcome of the measurement.
- Then: principle of maximum likelihood allows parameter estimation.

Model functions

- N data pairs (x_i, y_i).
- Want $f(x_i)$ that describes y_i so $f(x_i) \approx y_i$.
- The function *f* is our fit model. What is reasonable form for *f* ?
- Justified by how well f fits the data *and physical plausibility*
- Often starts by eyeballing it!
- *f* needs *M* tunable parameters $p_1, \ldots, p_M$, therefore:
  - $y_i \approx f(x_i; p_1, \ldots, p_M)$.
- Once a functional form has been chosen, game is to determine the numerical values of p_i (and their errors) that best fit the data.
- Assume the statistical fluctuations of the data are Gaussian

Likelihood

- Probability P for observing y_i for an independent variable x_i is given by:

Our description of the "mean"

$$P(x_i; p_1, \dots, p_M)\,dx = \frac{1}{\sqrt{2\cdot\pi}}\cdot\frac{1}{s_i}\cdot exp\left(-\frac{[y_i - f(x_i, p_1, \dots, p_M)]^2}{2\cdot s_i^2}\right)dx$$

- Assume each y_i subject to fluctuations of known standard dev s_i (but x_i free of fluctuations)
- f plays the role of the underlying true value of y
- Want $p_1, \dots, p_M$ that maximize the likelihood of observing the union of all N pairs (x_i, y_i).

Likelihood

- Maximize product of individual data-pair wise probs:

$$P_s = \prod_{i=1}^{N} P(x_i; y_i, p_1, \dots, p_M)$$

Easier to find the max of ln P_s … ln(x) is monotonic

$$ln(P_s) = \sum_{i=1}^{N} ln[P(x_i; y_i, p_1, \dots, p_M)]$$

$$= \sum_{i=1}^{N} ln\left[\frac{1}{\sqrt{2 \cdot \pi}} \cdot \frac{1}{s_i} \cdot exp\left(-\frac{[y_i - f(x_i, p_1, \dots, p_M)]^2}{2 \cdot s_i^2}\right)\right]$$

Likelihood

$$ln(P_s) = \sum_{j=1}^{N} ln\left[\frac{1}{\sqrt{2\cdot\pi}}\cdot\frac{1}{s_i}\cdot exp\left(-\frac{[y_i - f(x_i, p_1, \ldots, p_M)]^2}{2\cdot s_i^2}\right)\right]$$

$$ln(P_s) = \sum_{i=1}^{N} ln\left(\frac{1}{\sqrt{2\cdot\pi}}\cdot\frac{1}{s_i}\right) - \frac{1}{2}\cdot\sum_{i=1}^{N}\frac{[y_i - f(x_i, p_1, \ldots, p_M)]^2}{s_i^2}$$

- Set of p_j that maximizes the likelihood are the *best fit parameters*.

$$\frac{\partial P_s(x_i; y_i, p_1, \ldots, p_M)}{\partial p_j} = 0$$

Likelihood

$$\frac{\partial}{\partial p_j}\left[\sum_{i=1}^{N} \underbrace{ln\left(\frac{1}{\sqrt{2\cdot\pi}}\cdot\frac{1}{s_i}\right)}_{\text{does not depend on } p_j\text{, all derivatives are 0}} - \frac{1}{2}\cdot\underbrace{\sum_{i=1}^{N}\frac{[y_i - f(x_i; p_1, \ldots, p_M)]^2}{s_i^2}}_{\text{minus sign: likelihood is maximal when this expression is minimal}}\right] = 0$$

for $j = 1, \ldots, M$

$$\chi^2 = \sum_{i=1}^{N}\frac{[y_i - f(x_i; p_1, \ldots, p_M)]^2}{s_i^2}$$

- This is the *chi-square statistic*.
- Measures the deviation of our data from the fit function f.
- Choose p_j to minimize χ^2 and thus maximize likelihood

Chi square

$$\chi^2(N, x_i; y_i, s_i, p_1, \ldots, p_M) = \frac{1}{N-M} \cdot \sum_{i=1}^{N} \frac{[y_i - f(x_i,; p_1, \ldots, p_M)]^2}{s_i^2}$$

This method of finding the best-fit f is called *chi-square minimization*. It works for just about any function.

Obtain the unknown parameters p_i by simultaneously solving the M equations:

$$\frac{\partial}{\partial p_j}\chi^2 = \frac{\partial}{\partial p_j} \sum_{i=1}^{N} \frac{[y_i - f(x_i; p_1, \ldots, p_M)]^2}{s_i^2} = 0 \quad for\ j = 1, \ldots, M$$

Usually leads to a system of M non-linear equations that can't be solved analytically … numerical methods required

Numerical methods

- The linear case is analytically solvable
  - $y_i = f(x_i) = p_2 \cdot x_i + p_1$
- Every decent analysis program does this
- Excel does the absolute bare minimum
- Some will do much more – matlab, mathematica, originlab, python … many options, learn one of these.

Linear case

Describe data with linear function:
(each measurement has own s_i)

$$\chi^2 = \frac{1}{N-M} \cdot \sum_{i=1}^{N} \frac{(y_i - p_2 \cdot x_i - p_1)^2}{s_i^2}$$

Find the values of p_1 and p_2 which minimize χ^2

$$\frac{\partial \chi^2}{\partial p_2} = 0 \quad \text{and} \quad \frac{\partial \chi^2}{\partial p_1} = 0$$

$$\frac{\partial \chi^2}{\partial p_2} = -2 \cdot \sum_{i=1}^{N} \frac{(y_i - p_2 \cdot x_i - p_1) \cdot x_i}{s_i^2} = 0$$

$$\frac{\partial \chi^2}{\partial p_1} = -2 \cdot \sum_{i=1}^{N} \frac{(y_i - p_2 \cdot x_i - p_1)}{s_i^2} = 0$$

2 eqns 2 unknowns
Need at least two data pairs to solve this problem; more improve precision.

Simultaneously solve two inhomogeneous linear equations.

Linear case

$$-\sum_{i=1}^{N}\frac{y_i \cdot x_i}{s_i^2} + p_2 \cdot \sum_{i=1}^{N}\frac{x_i^2}{s_i^2} + p_1 \cdot \sum_{i=1}^{N}\frac{x_i}{s_i^2} = 0$$

$$-\sum_{i=1}^{N}\frac{y_i}{s_i^2} + p_2 \cdot \sum_{i=1}^{N}\frac{x_i}{s_i^2} + p_1 \cdot \sum_{i=1}^{N}\frac{1}{s_i^2} = 0$$

Solve for p_1 and p_2.

Tedious; details in appendix

Algebra ensues …

$$p_1 = \frac{1}{\Delta} \cdot \left(\sum_{i=1}^{N}\frac{x_i^2}{s_i^2} \cdot \sum_{i=1}^{N}\frac{y_i}{s_i^2} - \sum_{i=1}^{N}\frac{x_i}{s_i^2} \cdot \sum_{i=1}^{N}\frac{x_i \cdot y_i}{s_i^2} \right)$$

$$p_2 = \frac{1}{\Delta} \cdot \left(\sum_{i=1}^{N}\frac{1}{s_i^2} \cdot \sum_{i=1}^{N}\frac{x_i \cdot y_i}{s_i^2} - \sum_{i=1}^{N}\frac{x_i}{s_i^2} \cdot \sum_{i=1}^{N}\frac{y_i}{s_i^2} \right)$$

where

$$\Delta = \left[\sum_{i}\frac{1}{s_i^2} \cdot \sum_{i}\frac{x_i^2}{s_i^2} - \left(\sum_{i}\frac{x_i}{s_i^2} \right)^2 \right]$$

Linear result

Slope p_2 of the fit line:

$$\boxed{p_2 = \frac{1}{\Delta} \cdot \left(\sum_{i=1}^{N} \frac{1}{s_i^2} \cdot \sum_{i=1}^{N} \frac{x_i \cdot y_i}{s_i^2} - \sum_{i=1}^{N} \frac{x_i}{s_i^2} \cdot \sum_{i=1}^{N} \frac{y_i}{s_i^2} \right)}$$

Intercept p_1 of the fit line:

$$\boxed{p_1 = \frac{1}{\Delta} \cdot \left(\sum_{i=1}^{N} \frac{x_i^2}{s_i^2} \cdot \sum_{i=1}^{N} \frac{y_i}{s_i^2} - \sum_{i=1}^{N} \frac{x_i}{s_i^2} \cdot \sum_{i=1}^{N} \frac{x_i \cdot y_i}{s_i^2} \right)}$$

$$\Delta = \sum_{i=1}^{N} \frac{1}{s_i^2} \cdot \sum_{i=1}^{N} \frac{x_i^2}{s_i^2} - \left(\sum_{i=1}^{N} \frac{x_i}{s_i^2} \right)^2$$

i: number of measurements, $i=1, 2, \ldots, N$

x_i: independent variable

y_i: dependent variable

s_i: standard deviation of y_i

s_{p2}: standard deviation of p_2

s_{p1}: standard deviation of p_1

Linear result (equal s_i)

If all uncertainties are equal ($s = s_i$), simple enough Excel can do it:

Slope p_2 of the fit line:

$$\boxed{p_2 = \frac{1}{\Delta'} \cdot \left(N \cdot \sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i \right)}$$

$$\Delta' = N \cdot \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2$$

Intercept p_1 of the fit line:

$$\boxed{p_1 = \frac{1}{\Delta'} \cdot \left(\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} x_i \cdot y_i \right)}$$

Linear result - uncertainty

- If all uncertainties can be assumed to be equal ($s = s_i$), you can determine this common uncertainty from the data.
- e.g. uncertainties are instrumental and you've used the same instrument for all

$$s^2 = \frac{1}{N-2} \cdot \sum_{i=1}^{N} (y_i - p_2 \cdot x_i - p_1)^2$$

Linear result - uncertainty

What about the error on the fitted slope p_2 and intercept p_1?
(general case of unequal s_i again)

$$\frac{\partial \chi^2}{\partial p_2} = p_1 \cdot \sum_i \frac{x_i}{s_i^2} + p_2 \cdot \sum_i \frac{x_i^2}{s_i^2} - \sum_i \frac{x_i \cdot y_i}{s_i^2} = 0$$

Solve for slope p_2…

$$p_2 = \frac{\sum \frac{x_i \cdot y_i}{s_i^2} - p_1 \cdot \sum \frac{x_i}{s_i^2}}{\sum \frac{x_i^2}{s_i^2}}$$

The slope p_2 and intercept p_1 are dependent on each other (correlated).

Must be taken into account when calculating the error of interpolated or extrapolated values $y = p_2 \cdot x + p_1$

Linear result - uncertainty

We use error propagation to find the errors on p_1 and p_2, assuming the errors s_j of the individual measurements y_j are uncorrelated.

$$s_{p_2}^2 = \sum_{i=1}^{N} \left(\frac{\partial p_2}{\partial y_i}\right)^2 \cdot s_i^2 \qquad s_{p_1}^2 = \sum_{i=1}^{N} \left(\frac{\partial p_1}{\partial y_i}\right)^2 \cdot s_i^2$$

After some manipulation:

$$\boxed{s_{p_2}^2 = \frac{1}{\Delta} \cdot \sum_{i=1}^{N} \frac{1}{s_i^2}} \qquad \boxed{s_{p_1}^2 = \frac{1}{\Delta} \cdot \sum_{i=1}^{N} \frac{x_i^2}{s_i^2}}$$

Where, as before: $$\Delta = \sum_{i=1}^{N} \frac{1}{s_i^2} \cdot \sum_{i=1}^{N} \frac{x_i^2}{s_i^2} - \left(\sum_{i=1}^{N} \frac{x_i}{s_i^2}\right)^2$$

Linear result - uncertainty

For the special case of equal uncertainties $s = s_i$ for all y_i-values the uncertainties s_{p_1} and s_{p_2} of the fitted intercept p_1 and slope p_2 are:

$$\boxed{s_{p_1}^2 = \frac{s^2}{\Delta'} \cdot \sum_{i=1}^{N} x_i^2} \qquad \boxed{s_{p_2}^2 = N \cdot \frac{s^2}{\Delta'}}$$

$$\Delta' = N \cdot \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2$$

$$s^2 = \frac{1}{N-2} \cdot \sum_{i=1}^{N} (y_i - p_2 \cdot x_i - p_1)^2$$

Linear result - uncertainty

- In Excel - use the function LINEST. Syntax:
  =LINEST(y-array,x-array,TRUE,TRUE)

- You'll need to look up the details.

- This routine is what EXCEL calls an "array formula", it needs to be declared as such.

- Array formulas typically require some output to be spread over multiple cells, you need to define which cells.

- Again, look up how to do this.

Summary

$$\boxed{f(x) = p_2 \cdot x + p_1}$$

N correlated pairs x_i and y_i

Intercept: $$p_1 = \frac{1}{\Delta'} \cdot \left(\sum_{i=1}^{N} x_i^2 \cdot \sum_{i=1}^{N} y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} x_i \cdot y_i \right)$$

Error in intercept: $$s_{p_1} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{\Delta'}} \cdot s$$

Slope: $$p_2 = \frac{1}{\Delta'} \cdot \left(N \cdot \sum_{i=1}^{N} x_i \cdot y_i - \sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i \right)$$

Error in slope: $$s_{p_2} = \sqrt{\frac{N}{\Delta'}} \cdot s$$

N correlated pairs x_i and y_i with individual y-errors s_i

Intercept: $$p_1 = \frac{1}{\Delta} \cdot \left(\sum_{i=1}^{N} \frac{x_i^2}{s_i^2} \cdot \sum_{i=1}^{N} \frac{y_i}{s_i^2} - \sum_{i=1}^{N} \frac{x_i}{s_i^2} \cdot \sum_{i=1}^{N} \frac{x_i \cdot y_i}{s_i^2} \right)$$

Error in intercept: $$s_{p_1} = \sqrt{\frac{1}{\Delta} \cdot \sum_{i=1}^{N} \frac{x_i^2}{s_i^2}}$$

Slope: $$p_2 = \frac{1}{\Delta} \cdot \left(\sum_{i=1}^{N} \frac{1}{s_i^2} \cdot \sum_{i=1}^{N} \frac{x_i \cdot y_i}{s_i^2} - \sum_{i=1}^{N} \frac{x_i}{s_i^2} \cdot \sum_{i=1}^{N} \frac{y_i}{s_i^2} \right)$$

Error in slope: $$s_{p_2} = \sqrt{\frac{1}{\Delta} \cdot \sum_{i=1}^{N} \frac{1}{s_i^2}}$$

$$\boxed{\Delta' = N \cdot \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2}$$

$$\boxed{s^2 = \frac{1}{N-2} \cdot \sum_{i=1}^{N} (y_i - p_2 \cdot x_i - p_1)^2}$$

$$\boxed{\Delta = \sum_{i=1}^{N} \frac{1}{s_i^2} \cdot \sum_{i=1}^{N} \frac{x_i^2}{s_i^2} - \left(\sum_{i=1}^{N} \frac{x_i}{s_i^2} \right)^2}$$

Going further - linearization

- Both the slope and intercept are determined by simple sums; no complicated iterative process is needed to get the fit results
- In many cases, experimental problems can be ***linearized*** by redefining experimental variables
- Then, linear regression offers a simple means to find a description of the data.

Example: Rydberg constant

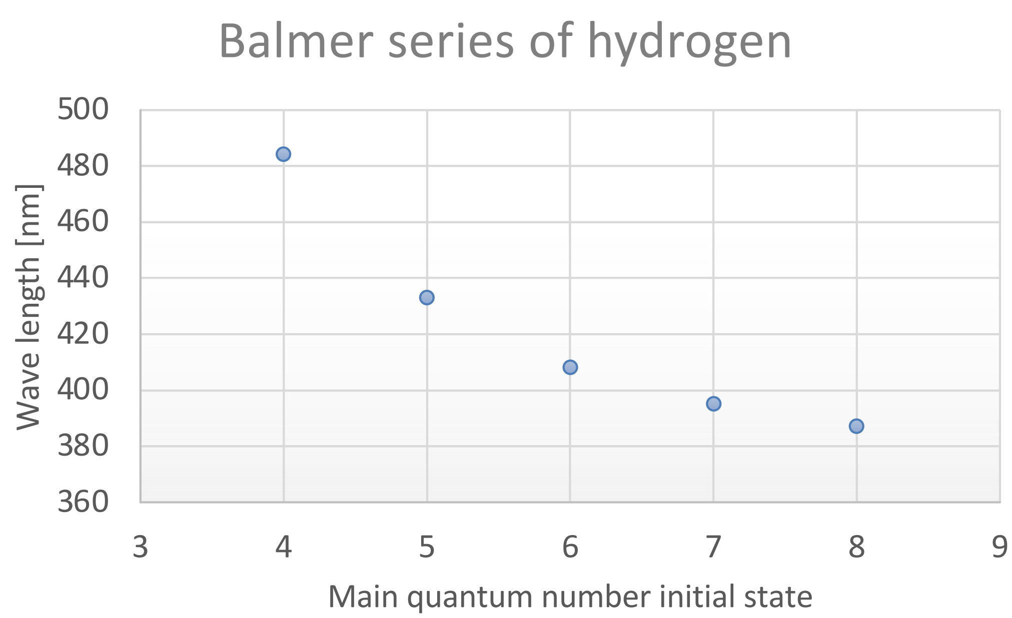
- Atomic spectra - wavelengths of transitions in the hydrogen atom (Balmer series).
- Measure the Rydberg constant, R_H, in Balmer formula
- Rydberg formula: $\frac{1}{\lambda} = R_H \cdot \left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$
- Balmer series: $n_1 = 2$ and $n_2 = 3, \dots, \infty$.
- Given n, λ data, how to extract R_H?

Example: Rydberg constant

Data from a hydrogen lamp via the Ocean Optics spectrometer.

n_2	λ (nm)	s_λ (nm)
4	484	4.0
5	433	3.0
6	408	2.5
7	395	3.0
8	387	1.0

First attempt: plot the raw data, wavelength versus main quantum number of final state.

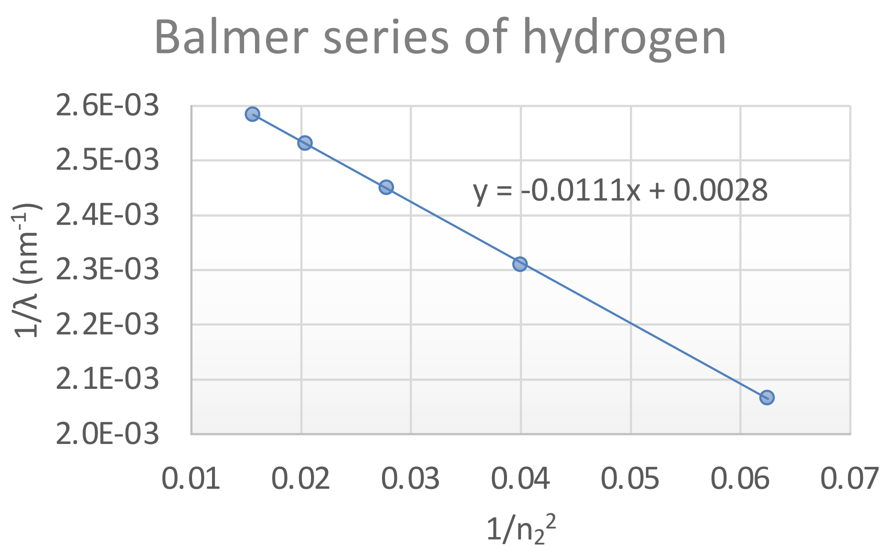


Doesn't help, the problem is obviously non-linear.

Example: Rydberg constant

Now linearize the problem with: $\frac{1}{\lambda} = -R_H \cdot \left(\frac{1}{n_2^2}\right) + \frac{R_H}{n_1^2}$.

n_2	$1/n_2^2$	λ (nm)	s_λ (nm)	$1/\lambda$ (nm^{-1})	$s_{1/\lambda}$ (nm^{-1})
4	0.063	484	4.0	2.066E-03	1.71E-05
5	0.040	433	3.0	2.309E-03	1.60E-05
6	0.028	408	2.5	2.451E-03	1.50E-05
7	0.020	395	3.0	2.532E-03	1.92E-05
8	0.016	387	1.0	2.584E-03	6.68E-06



Translate errors of y-values, from λ to $1/\lambda$. Error propagation ftw:

$$\frac{d\left(\frac{1}{\lambda}\right)}{d\lambda} = -\frac{1}{\lambda^2}$$

$$s_{1/\lambda} = \frac{s_\lambda}{\lambda^2}$$

Now apply linear regression to find the slope & Rydberg constant.

Details.

n_2	$x \equiv 1/n_2^2$	λ (nm)	s_λ (nm)	$y \equiv 1/\lambda$ (nm^{-1})	$s_{1/\lambda}$ (nm^{-1})	x_i^2/s_i^2	y_i/s_i^2	x_i/s_i^2	$x_i \cdot y_i/s_i^2$	$1/s_i^2$
4	0.0625	484	4	2.066E-03	1.71E-05	1.34E+07	7.086E+06	2.14E+08	4.429E+05	3.43E+09
5	0.0400	433	3	2.309E-03	1.60E-05	6.25E+06	9.020E+06	1.56E+08	3.608E+05	3.91E+09
6	0.0278	408	2.5	2.451E-03	1.50E-05	3.42E+06	1.087E+07	1.23E+08	3.019E+05	4.43E+09
7	0.0204	395	3	2.532E-03	1.92E-05	1.13E+06	6.848E+06	5.52E+07	1.398E+05	2.70E+09
8	0.0156	387	1	2.584E-03	6.68E-06	5.48E+06	5.796E+07	3.50E+08	9.056E+05	2.24E+10
Sum						**2.97E+07**	**9.18E+07**	**8.99E+08**	**2.15E+06**	**3.69E+10**

Perform all multiplications and products of the sums to get $p_1, p_2 \rightarrow n_1, R_H \ldots$

$p_2 = -R_H = (1.109 \pm 0.036) \cdot 10^{-7}\ m$ (y-axis is in nm, I converted to m).

n_1 from intercept: $n_1 = \sqrt{\frac{R_H}{p_1}} = 2.005$ but what's the error of n_1?

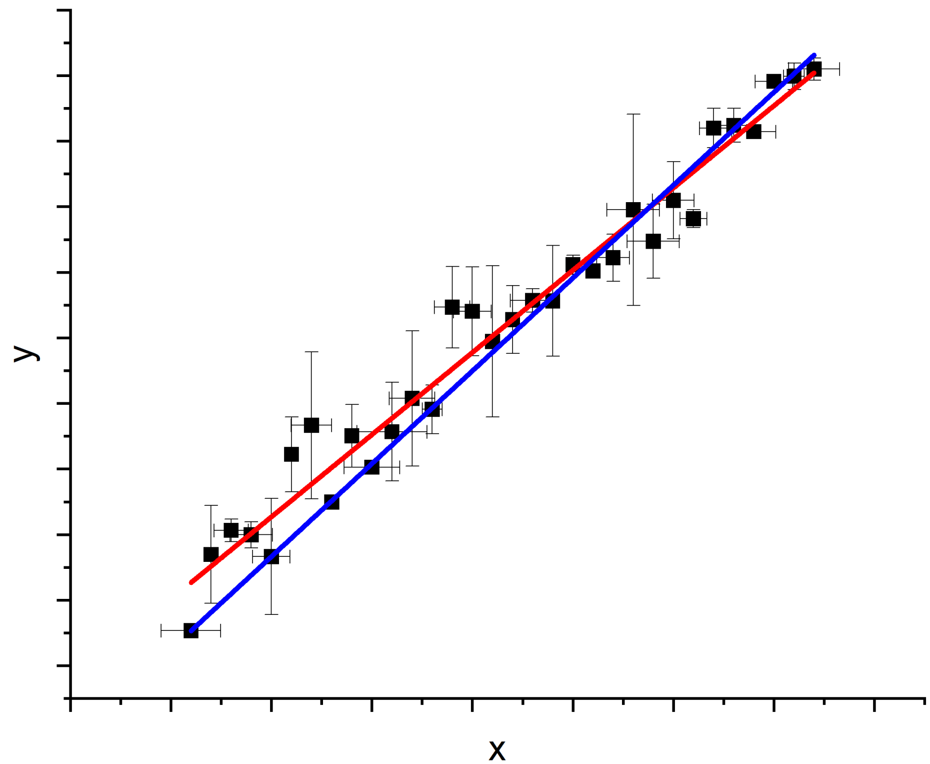
Tabulated value: $R_H = 1.097 \cdot 10^{-7}\ m$

In practice: don't have to do manually with sums, could use LINEST *in* Excel

Linear fit with uncertainties

- Red: neglect
- Blue: include
- *Huge* for intercept
- *Non-negligible for slope*

(curves with or without x error same on this scale)



Mode	Slope (err)	Intercept (err)
Ignore both	5.02 ± 0.17	0.32 ± 3.21
Include y error	5.67 ± 0.08	-15.1 ± 1.5
Include y and x errors	5.48 ± 0.12	-12.5 ± 1.8

Power laws

- "I know, let's plot it on a log scale and the power is the slope"
- "I know, we can use different power laws in different regimes"

- Uncertainty and noise floor … propagate/subtract
- Nothing is *really* a power law except in a narrow range
- Don't piece together models without (logical) glue

A better way

- If the model is $y = Ax^n$
- Then $\ln y = n \ln x + \ln A$
- It is true the slope of a ln y – ln x plot has slope n, but it is easy to fool yourself
- Better: logarithmic derivative

$$\frac{\partial \ln y}{\partial \ln x} = \frac{\partial}{\partial \ln x}(n \ln x + \ln A) = n$$

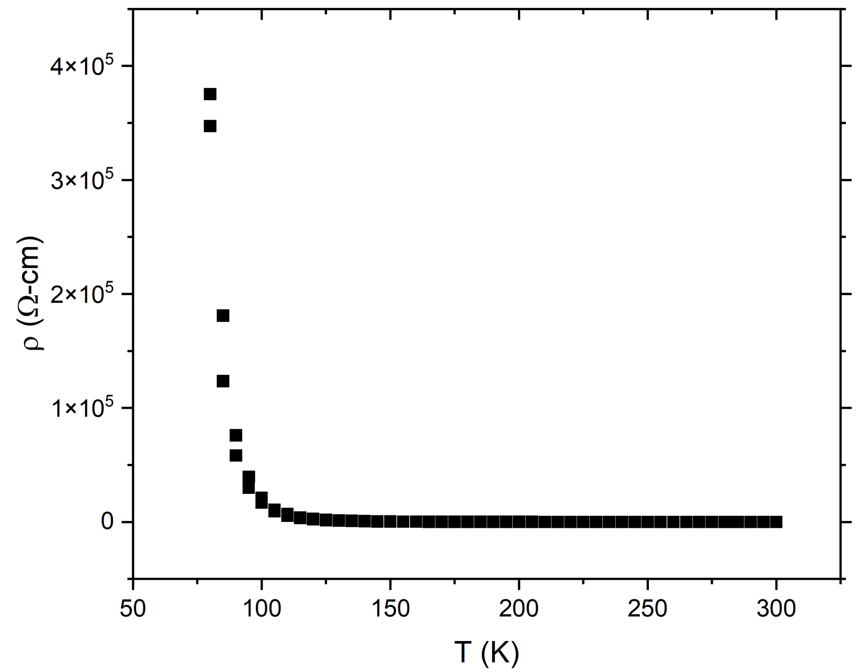
- Much easier to judge if plot is just a constant

Exercise

- From here you learn by doing
- I’ll give you data – resistivity vs temperature.
- You come up with a model and fit
- If possible – reason for model? Physics?
- Not including uncertainty for now
- Report fit parameters with uncertainty and (chi-square)/DOF (and a plot obviously)

Data and plausible models

- Will give csv file of data
- Plausible models? Many!
- Material – VO_2



$$\rho(T) = \rho_o + \rho_1 e^{-\alpha T} \quad \text{semiconductor-like}$$

$$\rho(T) = \rho_o + \rho_1 e^{\Delta/T} \quad \text{activated transport}$$

$$\rho(T) = \rho_o + \rho_1 e^{\sqrt{(\Delta/T)}} \quad \text{Mott variable-range hopping}$$

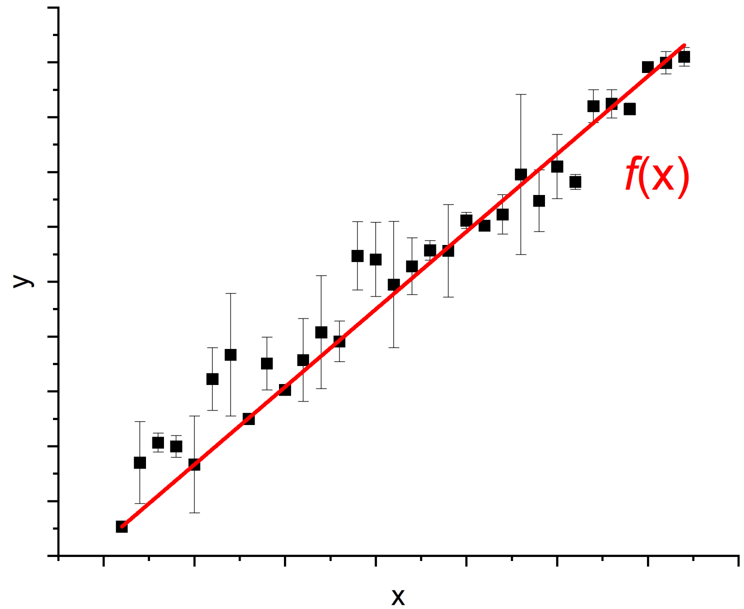
$$\rho(T) = \rho_o + \beta T^n \quad \text{power law - many models}$$

Appendices

- Further details on uncertainties of extrapolated and interpolated data
- Some derivations

Uncertainties on Extrapolated and Interpolated Values

After performing linear regression on $x - y$-data pairs, the fit line's utility is often to use it to determine the y -values for x-values you haven't directly observed during your "calibration". If these x-values are bracket by x_{min} and x_{max} such as $x_{min} \leq x \leq x_{max}$ you call this *interpolation*. If x-lies outside the "calibrated" range you call this process *extrapolation*. The question is: how do we determine the uncertainty on interpolated and extrapolated y-values?



We now know how to fit this data with the equation $y = f(x) = m \cdot x + b$. We get the values of m and b and their uncertainties s_m and s_b through linear regression.

Suppose we want to calculate the value of $\hat{y}$ at some value of x, where we did not make a measurement, using the linear regression equation. What is the uncertainty on $s_{\hat{y}}$ on $\hat{y}$?

My apology: there is a terminology change here. Before: $f(x) = p_2 \cdot x + p_1$. For the rest of this section I use: $f(x) = m \cdot x + b$. I didn't feel like changing 50+ equations…

Another reminder: m and b are correlated because

$$m = \frac{\sum_i \frac{x_i \cdot y_i}{s_i^2} - b \cdot \sum_i \frac{x_i}{s_i^2}}{\sum_i \frac{x_i^2}{s_i^2}}$$

Note: the sums all go over $i = 1, \ldots, N$. From here on the summation index will be dropped from the equations to reduce the number of symbols.

Therefore to calculate the error on $f(x)$ we need to propagate the errors on m and b, including the covariant term.

$$s_{\hat{y}}^2 = \left(\frac{\partial f(x)}{\partial m}\right)^2 \cdot s_m^2 + \left(\frac{\partial f(x)}{\partial b}\right)^2 \cdot s_b^2 + 2 \cdot s_{mb} \cdot \left(\frac{\partial f(x)}{\partial m}\right) \cdot \left(\frac{\partial f(x)}{\partial b}\right)$$

$$s_{\hat{y}}^2 = \left(\frac{\partial f(x)}{\partial m}\right)^2 \cdot s_m^2 + \left(\frac{\partial f(x)}{\partial b}\right)^2 \cdot s_b^2 + 2 \cdot s_{mb} \cdot \left(\frac{\partial f(x)}{\partial m}\right) \cdot \left(\frac{\partial f(x)}{\partial b}\right)$$

Since $f(x) = m \cdot x + b$:

$$\frac{\partial f(x)}{\partial m} = x \quad \text{and} \quad \frac{\partial f(x)}{\partial b} = 1$$

$$s_{\hat{y}}^2 = x^2 \cdot s_m^2 + s_b^2 + 2 \cdot s_{mb} \cdot x$$

The covariance in this case is

$$s_{mb} = \sum_{i=1}^{N} \left(\frac{\partial m}{\partial y_i}\right) \cdot \left(\frac{\partial b}{\partial y_i}\right) \cdot s_i^2$$

I won't show where this comes from. However, you can find this treatment in L. Lyons, *Statistics for Nuclear and Particle Physicists*, Cambridge University Press, p127f.

From before, we showed that (sums go over $i = 1, \ldots N$)

$$m = \frac{\sum \frac{x_i \cdot y_i}{s_i^2} - b \cdot \sum \frac{x_i}{s_i^2}}{\sum \frac{x_i^2}{s_i^2}}$$

and

$$b = \frac{\sum \frac{x_i \cdot y_i}{s_i^2} - m \cdot \sum \frac{x_i^2}{s_i^2}}{\sum \frac{x_i}{s_i^2}}$$

Start with the equations for m and b from linear regression

$$m = \frac{1}{\Delta} \cdot \left(\sum_i \frac{x_i \cdot y_i}{s_i^2} \cdot \sum_i \frac{1}{s_i^2} - \sum_i \frac{x_i}{s_i^2} \cdot \sum_i \frac{y_i}{s_i^2} \right)$$

$$b = \frac{1}{\Delta} \cdot \left(\sum_i \frac{x_i^2}{s_i^2} \cdot \sum_i \frac{y_i}{s_i^2} - \sum_i \frac{x_i}{s_i^2} \cdot \sum_i \frac{x_i \cdot y_i}{s_i^2} \right)$$

This is how slope and intercept depend on the primary data (x_i, y_i, and s_i). The change of the results under a variation of the data (derivatives), weighted by the amount of variability (uncertainties of individual data points) determines the variability of the result.

Now take the partial derivatives of m and b with respect to y_i or jump to result.

$$\frac{\partial m}{\partial y_i} = \frac{1}{\Delta} \cdot \left(\left[\sum_j \frac{1}{s_j^2} \right] \cdot \frac{x_i}{s_i^2} - \left[\sum_j \frac{x_j}{s_j^2} \right] \cdot \frac{1}{s_i^2} \right)$$

$$\frac{\partial b}{\partial y_i} = \frac{1}{\Delta} \cdot \left(\left[\sum_j \frac{x_j^2}{s_j^2} \right] \cdot \frac{1}{s_i^2} - \left[\sum_j \frac{x_j}{s_j^2} \right] \cdot \frac{x_i}{s_i^2} \right)$$

Remember: Δ does not depend on the y_i-values, only on the x_i and s_i-values. This means it acts as a parameter, you don't need to evaluate a complicated derivative of a ratio of functions.

Substitute this into the expression for covariance s_{mb}

$$s_{mb} = \sum_i s_i^2 \cdot \left(\frac{\partial m}{\partial y_i}\right) \cdot \left(\frac{\partial b}{\partial y_i}\right) = \frac{1}{\Delta^2} \cdot \sum_i s_i^2 \left(\left[\sum_j \frac{1}{s_j^2} \right] \cdot \frac{x_i}{s_i^2} - \left[\sum_j \frac{x_j}{s_j^2} \right] \cdot \frac{1}{s_i^2} \right) \cdot \left(\left[\sum_j \frac{x_j^2}{s_j^2} \right] \cdot \frac{1}{s_i^2} - \left[\sum_j \frac{x_j}{s_j^2} \right] \cdot \frac{x_i}{s_i^2} \right)$$

$$s_{mb} = \frac{1}{\Delta^2} \cdot \sum_i s_i^2 \left(\left[\sum_j \frac{1}{s_j^2} \right] \cdot \frac{x_i}{s_i^2} - \left[\sum_j \frac{x_j}{s_j^2} \right] \cdot \frac{1}{s_i^2} \right) \cdot \left(\left[\sum_j \frac{x_j^2}{s_j^2} \right] \cdot \frac{1}{s_i^2} - \left[\sum_j \frac{x_j}{s_j^2} \right] \cdot \frac{x_i}{s_i^2} \right)$$

$$s_{mb} = \frac{1}{\Delta^2} \cdot \sum_i s_i^2 \left[\left(\sum_j \frac{1}{s_j^2} \right) \cdot \left(\sum_j \frac{x_j^2}{s_j^2} \right) \cdot \frac{x_i}{s_i^4} - \left(\sum_j \frac{x_j}{s_j^2} \right) \cdot \left(\sum_j \frac{x_j^2}{s_j^2} \right) \cdot \frac{1}{s_i^4} \right]$$
$$+ \frac{1}{\Delta^2} \cdot \sum_i s_i^2 \left[- \left(\sum_j \frac{1}{s_j^2} \right) \cdot \left(\sum_j \frac{x_j}{s_j^2} \right) \cdot \frac{x_i^2}{s_i^4} + \left(\sum_j \frac{x_j}{s_j^2} \right) \cdot \left(\sum_j \frac{x_j}{s_j^2} \right) \cdot \frac{x_i}{s_i^4} \right]$$

Now write this out as 4 separate sums.

$$s_{mb} = \frac{1}{\Delta^2} \cdot \cancel{\left[\left(\sum_j \frac{1}{s_j^2} \right) \left(\sum_j \frac{x_j^2}{s_j^2} \right) \sum_i \frac{x_i}{s_i^2} - \left(\sum_j \frac{x_j}{s_j^2} \right) \left(\sum_j \frac{x_j^2}{s_j^2} \right) \sum_i \frac{1}{s_i^2} \right]}$$

$$+ \frac{1}{\Delta^2} \cdot \sum_i \frac{x_i}{s_i^2} \underbrace{\left[- \left(\sum_j \frac{1}{s_j^2} \right) \left(\sum_j \frac{x_j^2}{s_j^2} \right) + \left(\sum_j \frac{x_j}{s_j^2} \right)^2 \right]}_{\Delta}$$

Finally, $$s_{mb} = -\frac{1}{\Delta} \cdot \sum_i \frac{x_i}{s_i^2}$$

$$s_{mb} = -\frac{1}{\Delta} \cdot \sum_i \frac{x_i}{s_i^2} \qquad s_m^2 = \frac{1}{\Delta} \cdot \sum_i \frac{1}{s_i^2} \qquad s_b^2 = \frac{1}{\Delta} \cdot \sum_i \frac{x_i^2}{s_i^2}$$

$$s_{\hat{y}}^2 = x^2 \cdot s_m^2 + s_b^2 + 2 \cdot s_{mb} \cdot x$$

All sums are performed over all measured values, $i = 1, \dots, N$.

$$s_{\hat{y}}^2 = \frac{1}{\Delta} \cdot \left[\sum_i \frac{x_i^2}{s_i^2} + x^2 \cdot \sum_i \frac{1}{s_i^2} - 2 \cdot x \cdot \sum_i \frac{x_i}{s_i^2} \right]$$

$$\Delta = \sum_{i=1}^{N} \frac{1}{s_i^2} \cdot \sum_{i=1}^{N} \frac{x_i^2}{s_i^2} - \left(\sum_{i=1}^{N} \frac{x_i}{s_i^2} \right)^2$$

Note that the error in $\hat{y}$ depends on x. The further you extrapolate from measured values, the larger the uncertainty on the extrapolation becomes.

Check what happens in the special case that $x = 0$:

$$s_{\hat{y}}^2 = \frac{1}{\Delta} \cdot \left[\sum_i \frac{x_i^2}{s_i^2} + x^2 \cdot \sum_i \frac{1}{s_i^2} - 2 \cdot x \cdot \sum_i \frac{x_i}{s_i^2} \right] = \frac{1}{\Delta} \cdot \sum_i \frac{x_i}{s_i^2}$$

This is simply the uncertainty on the y-intercept: $s_b^2 = \frac{1}{\Delta} \sum_i \frac{x_i^2}{s_i^2}$

For the unweighted case, $s_i = s$ for all $i = 1, \dots, N$

$$s_{\hat{y}}^2 = \frac{s^2}{\Delta'} \cdot \left[\sum_i x_i^2 + N \cdot x^2 - 2 \cdot x \cdot \sum_i x_i \right]$$

$$\Delta' = N \cdot \sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i \right)^2$$

Example:

You have measured 7 linearly correlated $x - y$-data pairs (x_i, y_i) and have knowledge of their individual standard deviation s_{y_i}. What straight line fit $f(x) = m \cdot x + b$ do you obtain and what do you know about the uncertainties of the fit and its parameters m and b?

x [arbitrary units]	y [arbitrary units]	s_y [arbitrary units]
1.0	4.07	0.20
2.0	4.76	0.30
3.0	7.00	0.50
4.0	6.97	1.50
5.0	8.3	1.10
6.0	7.01	2.50
7.0	9.90	2.10

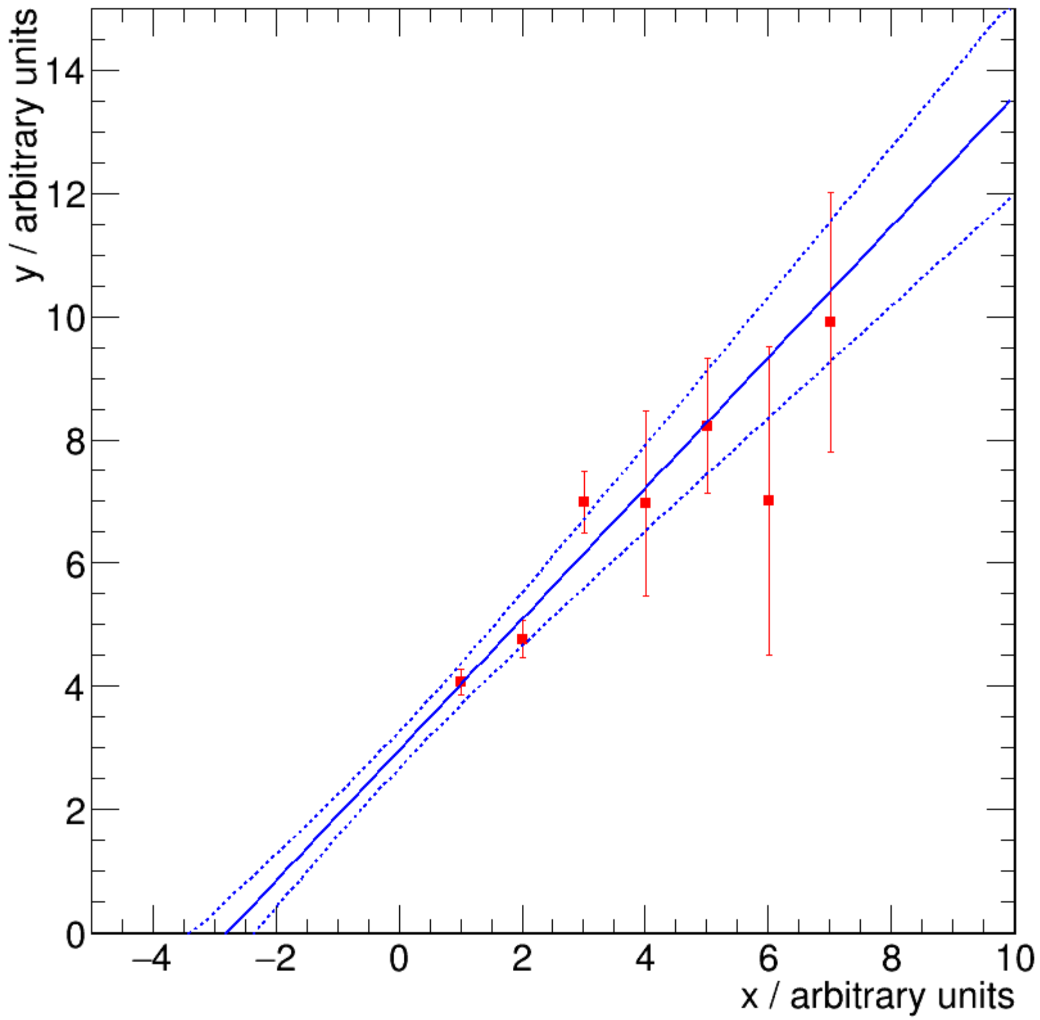
EXCEL's answer (ignoring individual point-wise uncertainties):
$m = 0.83 \pm 0.17$
$b = 3.53 \pm 0.77$
$\chi^2/NDF = 8.65/5$

Now perform linear regression we learned last class, taking into account the individual uncertainties.

$m = 1.06 \pm 0.16$
$b = 2.97 \pm 0.30$
χ^2/NDF
$= 5.04/5$

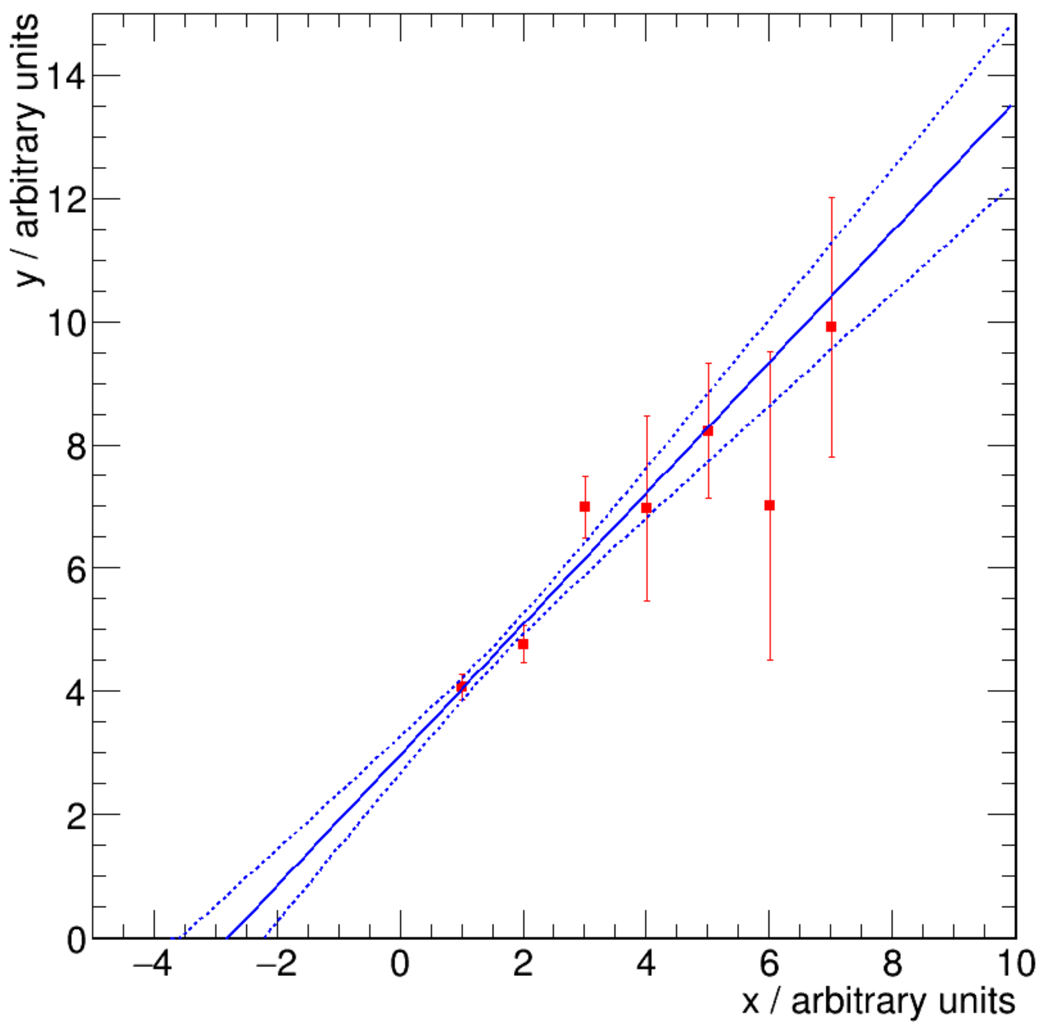
EXCEL's answer (ignoring individual point-wise uncertainties):
$m = 0.83 \pm 0.17$
$b = 3.53 \pm 0.77$
$\chi^2/NDF = 8.65/5$

The analyzed data was created with a random number generator (using the normal distribution). The "truth information" was:
$m = 1.0$
$b = 3.0$



How well do we estimate interpolated and extrapolated y-values?

In this example I just ignored the covariant error term. This model is simple but wrong.



In this calculation I utilized the covariant error term. This model is more complicated but correct.

The error boundaries are a little tighter as before, the point with minimal error corresponds to a different *x*-value.

The uncertainty is smallest where you have data (interpolation). It is smaller than the individual error bars. The uncertainty quickly grows where you have no supporting data (extrapolation).

I hope some of the material I presented sticks. These basic concepts of data treatment and estimation of certainty are essential tools for anybody in the sciences, engineering etc. who has to deal with data. In practical situations: if you can't know how sure to be about something, you need to base decisions on "feelings", "convictions", "common sense" instead of rational thought and quantifiable arguments.

There are certain situations where you have no choice because you simply don't know the quantifiable details. If one has a choice *ratio* (Latin for *reason*) is usually a good guide to decision making.

Appendix 2: derivation of the linear regression relations

Backup: derivation of linear regression formulas for slope and *y*-intercept

$$\chi^2 = \sum_i \frac{\left(y_i - y_{fi}\right)^2}{s_i^2} = \sum_i \frac{(y_i - m \cdot x_i - b)^2}{s_i^2}$$

$$\frac{\partial \chi^2}{\partial m} = -2 \sum_i \frac{(y_i - m \cdot x_i - b) x_i}{s_i^2} = 0$$

$$\frac{\partial \chi^2}{\partial b} = -2 \sum_i \frac{(y_i - m \cdot x_i - b)}{s_i^2} = 0$$

Note: in this appendix the straight line is parametrized as $f(x) = m \cdot x + b$. Therefore, $m \equiv p_2$ and $b \equiv p_1$.

$$-\sum_i \frac{y_i x_i}{s_i^2} + m\sum_i \frac{x_i^2}{s_i^2} + b\sum_i \frac{x_i}{s_i^2} = 0 \quad \Rightarrow \quad m\sum_i \frac{x_i^2}{s_i^2} + b\sum_i \frac{x_i}{s_i^2} = \sum_i \frac{y_i x_i}{s_i^2} \qquad \text{(I)}$$

$$-\sum_i \frac{y_i}{s_i^2} + m\sum_i \frac{x_i}{s_i^2} + b\sum_i \frac{1}{s_i^2} = 0 \quad \Rightarrow \quad -m\sum_i \frac{x_i}{s_i^2} - b\sum_i \frac{1}{s_i^2} = -\sum_i \frac{y_i}{s_i^2} \qquad \text{(II)}$$

Multiply (I) by $\sum_i \frac{x_i}{s^2}$ and (II) by $\sum_i \frac{x_i^2}{s^2}$

$$m\sum_i \frac{x_i^2}{s_i^2}\sum_i \frac{x_i}{s_i^2} + b\sum_i \frac{x_i}{s_i^2}\sum_i \frac{x_i}{s_i^2} = \sum_i \frac{y_i x_i}{s_i^2}\sum_i \frac{x_i}{s_i^2} \qquad \text{(I)}$$

$$-m\sum_i \frac{x_i}{s_i^2}\sum_i \frac{x_i^2}{s_i^2} - b\sum_i \frac{1}{s_i^2}\sum_i \frac{x_i^2}{s_i^2} = -\sum_i \frac{y_i}{s_i^2}\sum_i \frac{x_i^2}{s_i^2} \qquad \text{(II)}$$

Now add (I) and (II) and define $\Delta = \left[\sum_i \frac{1}{s_i^2} \sum_i \frac{x_i^2}{s_i^2} - \left(\sum_i \frac{x_i}{s_i^2}\right)^2\right]$

$$b = \frac{1}{\Delta}\left(\sum_i \frac{y_i}{s_i^2} \sum_i \frac{x_i^2}{s_i^2} - \sum_i \frac{y_i x_i}{s_i^2} \sum_i \frac{x_i}{s_i^2}\right)$$

y-intercept

Now modify the calculation from the previous slide to get the slope.

$$-\sum_i \frac{y_i x_i}{s_i^2} + m\sum_i \frac{x_i^2}{s_i^2} + b\sum_i \frac{x_i}{s_i^2} = 0 \Rightarrow m\sum_i \frac{x_i^2}{s_i^2} + b\sum_i \frac{x_i}{s_i^2} = \sum_i \frac{y_i x_i}{s_i^2} \qquad \text{(I)}$$

$$-\sum_i \frac{y_i}{s_i^2} + m\sum_i \frac{x_i}{s_i^2} + b\sum_i \frac{1}{s_i^2} = 0 \Rightarrow -m\sum_i \frac{x_i}{s_i^2} - b\sum_i \frac{1}{s_i^2} = -\sum_i \frac{y_i}{s_i^2} \qquad \text{(II)}$$

Multiply (I) by $\sum_i \frac{1}{\sigma^2}$ and (II) by $\sum_i \frac{x_i}{\sigma^2}$

$$m\sum_i \frac{x_i^2}{s_i^2}\sum_i \frac{1}{s_i^2} + b\sum_i \frac{x_i}{s_i^2}\sum_i \frac{1}{s_i^2} = \sum_i \frac{y_i x_i}{s_i^2}\sum_i \frac{1}{s_i^2} \qquad \text{(I)}$$

$$-m\sum_i \frac{x_i}{s_i^2}\sum_i \frac{x_i}{s_i^2} - b\sum_i \frac{1}{s_i^2}\sum_i \frac{x_i}{s_i^2} = -\sum_i \frac{y_i}{s_i^2}\sum_i \frac{x_i}{s_i^2} \qquad \text{(II)}$$

Now add (I) and (II) and define $\Delta = \left[\sum_i \frac{1}{s_i^2} \sum_i \frac{x_i^2}{s_i^2} - \left(\sum_i \frac{x_i}{s_i^2}\right)^2\right]$

$$m = \frac{1}{\Delta}\left(\sum_i \frac{y_i x_i}{s_i^2} \sum_i \frac{1}{s_i^2} - \sum_i \frac{y_i}{s_i^2} \sum_i \frac{x_i}{s_i^2}\right)$$ slope

Backup: derivation of error on y-intercept from linear regression fit

$$b = \frac{1}{\Delta}\left(\sum_i \frac{y_i}{\sigma_i^2}\sum_i \frac{x_i^2}{\sigma_i^2} - \sum_i \frac{y_i x_i}{\sigma_i^2}\sum_i \frac{x_i}{\sigma_i^2}\right)$$

$$\sigma_b^2 = \sum_j \left(\frac{\partial b}{\partial y_j}\right)\sigma_j^2$$

$$\frac{\partial b}{\partial y_j} = \frac{1}{\Delta}\left(\frac{1}{\sigma_j^2}\sum_i \frac{x_i^2}{\sigma_i^2} - \frac{x_j}{\sigma_j^2}\sum_i \frac{x_i}{\sigma_i^2}\right)$$

$$\sigma_b^2 = \sum_j \frac{1}{\Delta^2}\left(\frac{1}{\sigma_j^2}\sum_i \frac{x_i^2}{\sigma_i^2} - \frac{x_j}{\sigma_j^2}\sum_i \frac{x_i}{\sigma_i^2}\right)^2 \sigma_j^2$$

$$\sigma_b^2 = \sum_j \frac{\sigma_j^2}{\Delta^2}\left[\frac{1}{\sigma_j^4}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)^2 - 2\frac{x_j}{\sigma_j^4}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)\left(\sum_i \frac{x_i}{\sigma_i^2}\right) + \frac{x_j^2}{\sigma_j^4}\left(\sum_i \frac{x_i}{\sigma_i}\right)^2\right]$$

$$\sigma_b^2 = \frac{1}{\Delta^2}\sum_j \frac{1}{\sigma_j^2}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)^2 - 2\sum_j \frac{x_j}{\sigma_j^2}\left(\sum_i \frac{x_i^2}{\sigma_i^2}\right)\left(\sum_i \frac{x_i}{\sigma_i^2}\right) + \sum_j \frac{x_j^2}{\sigma_j^2}\left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2$$

$$\sigma_b^2 = \frac{1}{\Delta^2}\sum_j \frac{x_j^2}{\sigma_j^2}\cdot\underbrace{\left[\sum_j \frac{1}{\sigma_j^2}\sum_i \frac{x_i^2}{\sigma_i^2} - \left(\sum_i \frac{x_i}{\sigma_i^2}\right)^2\right]}_{\text{The term in square brackets is } \Delta}$$

$$\sigma_b^2 = \frac{1}{\Delta}\sum_j \frac{x_j^2}{\sigma_j^2}$$ Error on y-intercept

After a similar calculation

$$\sigma_m^2 = \frac{1}{\Delta}\sum_j \frac{1}{\sigma_j^2}$$ Error on slope