Mathematical Modelling
Lecture 4 – Fitting Data

Phil Hasnip
phil.hasnip@york.ac.uk

Overview of Course

- Model construction $\longrightarrow$ dimensional analysis
- Experimental input $\longrightarrow$ fitting
- Finding a 'best' answer $\longrightarrow$ optimisation
- Tools for constructing and manipulating models $\longrightarrow$ networks, differential equations, integration
- Tools for constructing and simulating models $\longrightarrow$ randomness
- Real world difficulties $\longrightarrow$ chaos and fractals

A First Course in Mathematical Modeling by Giordano, Weir & Fox, pub. Brooks/Cole. Today we're in chapter 3.

Aim

There are two main aims:

- To fit a model to experimental data, or to choose which model best fits the data $\longrightarrow$ Model fitting.
- To use given experimental data with a model to predict other experimental results $\longrightarrow$ Model interpolation.

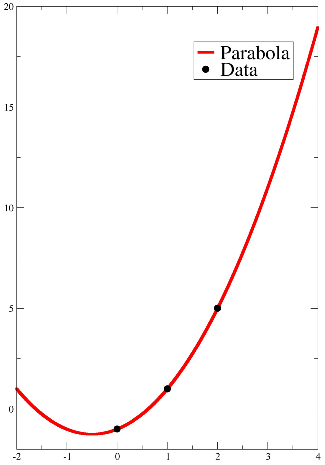
Aim

The difference between these two aims is one of emphasis:

- Model fitting: we expect some scatter in the experimental data, we want the best model of a given form – ‘theory driven’
- Model interpolation: the existing data is good, model is less important – ‘data driven’

Today we’ll be focussing on the first aim: model fitting.

Model fitting

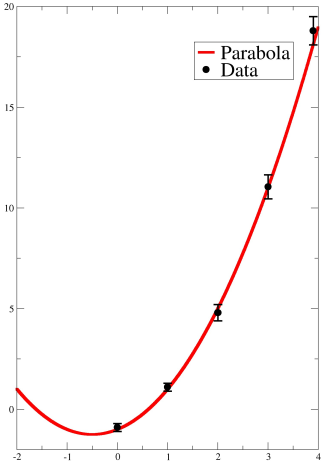


What do we mean by model fitting? Suppose we know

$$f(x) = a + bx + cx^2$$

If we knew $f(x)$ at three different points precisely then we could compute a, b and c.

Model fitting



In practice there is always experimental error, so we make several measurements and try to find the values of a, b and c that fit the data best. How do we do that?

Least-squares

We define the residual R_i as the difference between the data y_i and our model's prediction $f(x_i)$,

$$R_i = y_i - f(x_i)$$

Choose the coefficients of the model so as to minimise the sum of the squared residuals of model from data.

i.e. minimise

$$S = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

Least-squares

Suppose our model is a straight line: $f(x) = mx + c$.

$$S = \sum_{i=1}^{N} (y_i - mx_i - c)^2$$

And at the minimum of S we have

$$\begin{aligned} \frac{\partial S}{\partial m} &= 0 \\ \frac{\partial S}{\partial c} &= 0 \end{aligned}$$

Least-squares

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

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

Similar process for other forms of $f(x_i)$, though more parameters!

Least-squares

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

S measures the absolute error, but we could also measure the relative error:

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

These are both closely related to χ^2, another measure of 'goodness of fit':

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

Data transformations

What about transforming the data? E.g.

$$\begin{aligned} y &= \alpha e^{\beta x} \\ \Rightarrow \ln y &= \ln \alpha + \beta x \end{aligned}$$

we could then fit a straight line to $\ln y$.

Not a good idea! See spreadsheet...

Goodness of fit

We've already mentioned some ways to measure how well a model fits the experimental data.

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

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

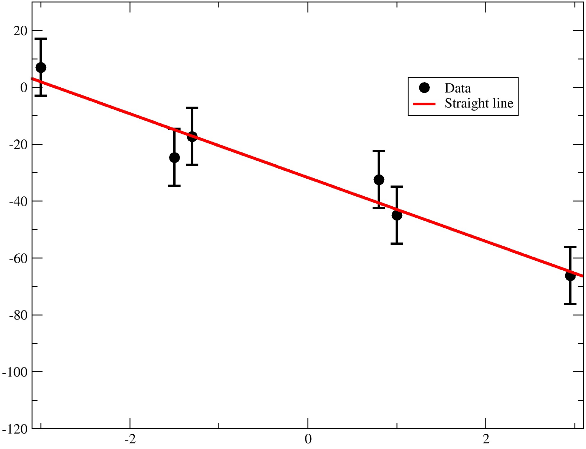
There are many others. One interesting method is to just look at the *maximum* deviation of the model.

Different models

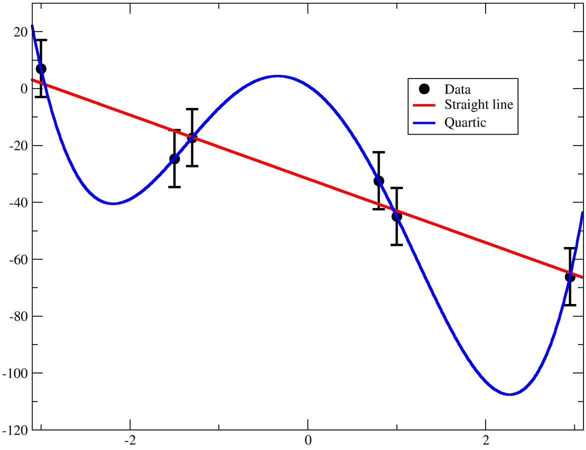
Once we have decided on our measure of ‘goodness of fit’, we can decide which of several models is the best.

BUT we need to be careful...

Different models



Different models



Different models

A model with more parameters is much more likely to fit the data well, regardless of whether it is actually better or not.

- Adding another term to a model usually improves the fit
- Is this improvement ‘real’, or chance?
- Is it worth adding the extra parameter?
- Occam’s razor $\longrightarrow$ simpler is better!

Degrees of freedom

If N data points, and p model parameters, then can think of the fitting process as:

- Use first p data points to determine model parameters
- Use remaining $N - p$ points to calculate error

The $N - p$ points represent the freedom we have in fitting a model of this form. We say there are *$N - p$ degrees of freedom*.

F-test

We look at the fractional improvement in goodness of fit, and we do this by calculating F,

$$F = \frac{\chi_2^2}{\chi_1^2}$$

(label models such that $F \geq 1$).

What F could just be chance? Decide what probability to reject: e.g. if probability of F by chance is $\leq 5\%$ then it is unlikely to happen accidentally, so decide model 2 is better than model 1.

The probability we choose to reject (e.g. 5%) is called the *significance level* – we usually use 5% or 1%.

Critical F-values

The maximum likely improvement of F due to chance at various significance levels can be found in tables of F values. It depends on the degrees of freedom of each model, so our procedure for testing is:

- Work out *degrees of freedom* for each model
- Decide significance level (usually 5% or 1%)
- Consult a table to find critical F-value, F_c
- If $F \geq F_c$ then the addition of extra parameters in model 2 is worth it

Critical F-values at 5% level

	$N-p_1$				
$N-p_2$	1	2	3	4	5
1	161.448	199.500	215.707	224.583	230.162
2	18.513	19.000	19.164	19.247	19.296
3	10.128	9.552	9.277	9.117	9.013
4	7.709	6.944	6.591	6.388	6.256
5	6.608	5.786	5.409	5.192	5.050
6	5.987	5.143	4.757	4.534	4.387
7	5.591	4.737	4.347	4.120	3.972
8	5.318	4.459	4.066	3.838	3.687
9	5.117	4.256	3.863	3.633	3.482
10	4.965	4.103	3.708	3.478	3.326
11	4.844	3.982	3.587	3.357	3.204
12	4.747	3.885	3.490	3.259	3.106

Back to the drawing board

Sometimes we find the model works significantly better under some circumstances than others. Examine the residuals

$$R_i = y_i - f(x_i)$$

Are there points a long way from the model prediction?

- Suspect data – measure again
- Suspect model – fit again, or re-check assumptions

Errors

Two main kinds of experimental error:

- Systematic
  e.g. your tape measure has stretched over time
- Random
  Measure several times, get slightly different results

Errors

The model can also introduce errors:

- Formulation
  Assumptions made in model may not be strictly correct
- Truncation
  Might make approximations to series, e.g.
  $\cos(x) \approx 1 - \frac{1}{2}x^2$
- Round-off
  Computers, calculators etc. can't represent numbers exactly

Errors

Want fitting procedure to care less about data points with greater error, so could use

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

(where δy_i is error in measurement y_i)

Summary

When fitting models to experimental data:

- Choose a measure of the difference between the model prediction and the experimental data
  - Absolute residual-squared
  - Relative residual-squared
  - χ^2
  - Worst error (Chebyshev)
  - Divided by experimental error
- For similar models, choose the one that minimises your measure of difference
- Only choose more complex models if the improvement is worth it (F-test)