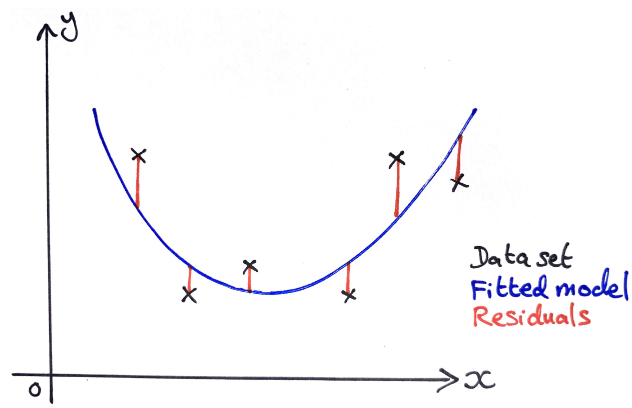
MMaD: Lecture 5 handout

Fitting a model to a data set

We can obtain a set of experimental data, and hypothesise that the relationship between the independent variable (that we control) and the dependent variable (that we measure) is described by some function (this is a **model**). Once we have decided on a general form of the relationship between the variables (e.g. linear, quadratic, expontial, power law), curve fitting is the process of **finding the set of parameter values** that best fits the set of experimental data.



Could this data set be described by a quadratic model $y = ax^2 + bx + c$, and what values should the parameters a, b, c have?

Software can decide what parameter values result in the best fit by minimising the sum of the squared differences ("residuals") between the actually-observed y-values in our data set and those predicted by our model with the current parameter choices.

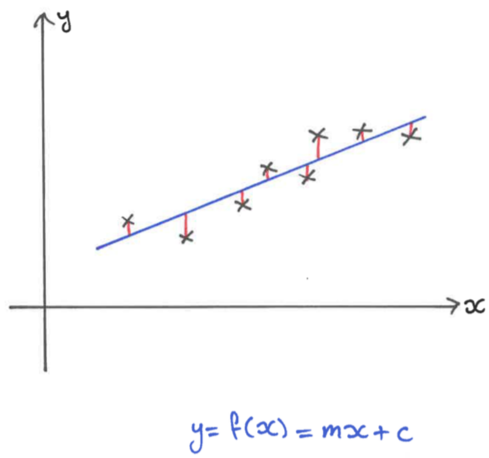
Curve fitting in EXCEL using SOLVER

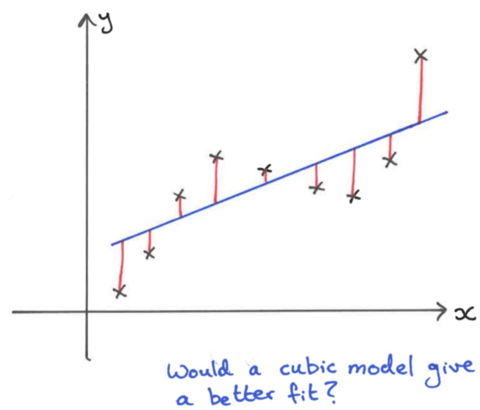
1. Begin with columns containing the observed (actual) set of data: one column each with x-values and one with y-values.
2. Create a scatter plot of the data and decide what sort of model would be suitable.
3. Input some estimate parameter values in cells away from the columns of data.
4. Create a column of y-values predicted by the model, that refer to the parameter value cells.
5. Create a column of the squares of the differences between the actual y-values and those estimated by the model. These are the "square residuals".
6. Sum all of these in a single cell.
7. Open SOLVER in the Data tab:

(a) We want to **minimise** the sum of the square residuals.
(b) By changing the variable cells containing the values of the parameters.
(c) Ensure that "Make unconstrained variables non-negative" is NOT checked.
(d) Click "Solve".

How do we determine which model to fit?

When fitting a curve to a set of data, we could choose several different models and fit the best parameter choices in each case. How would we know which model gives the best fit?





To quantify the "goodness of fit" for each, we can calculate R^2 (the coefficient of determination).

Calculating R^2

For a set of N data points (x_i, y_i), to which a model is fitted given by $y = f(x)$, we calculate R^2 using:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

where

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

and

$$SS_{tot} = \sum_{i=1}^{N} \left(y_i - \bar{y}\right)^2$$

A value of R^2 equal to 1 indicates that the curve fits the data perfectly. A smaller value means a poorer fit.