CHEM1033 - Week 10 Lecture - Model fitting

Chapters 8 and 9 of Monk and Munro, "Maths for Chemistry", 2nd edition.
Section 21.10 of Steiner, "The Chemistry Maths Book", 2nd edition.

Transformation to linear coordinates

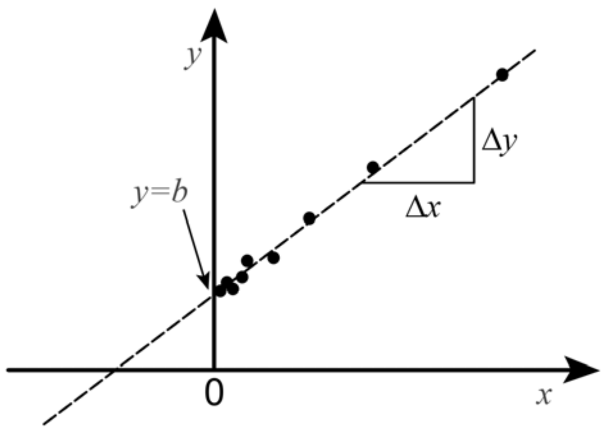
Functions encountered in models of physical and chemical processes are rarely linear, but can usually be brought into a linear form by a suitable substitution. There is no general recipe for that process – it relies on your experience and creativity. Examples:

$$\begin{aligned} y = a\exp(-bx) \quad &\Rightarrow \quad \ln(y) = \ln(a) - bx \\ y = \frac{a}{b+cx} \quad &\Rightarrow \quad \frac{1}{y} = \frac{b}{a} + \frac{c}{a}x \\ y = ax^2 + 1 \quad &\Rightarrow \quad \sqrt{y-1} = x\sqrt{a} \end{aligned} \tag{1}$$

In all three cases, plotting experimental measurements directly in $\{x, y\}$ coordinates would produce non-linear plots from which the parameters $\{a, b, c\}$ cannot be easily extracted by hand. However, plotting $\ln(y)$ against x in the first case, $1/y$ against x in the second case and $\sqrt{y-1}$ against x in the third case would produce linear plots from which the parameters may be extracted using the procedures described below.

Basic analysis of linear graphs

The most primitive method for extracting parameters from a linear graph is to plot the experimental data on a piece of graph paper and draw a line through it by hand:



For a linear function $y = ax + b$, where a and b are constant parameters, the b parameter can be extracted from the value of y at the point where $x = 0$ (the so-called *intercept*) and the a parameter may be obtained from a finite difference approximation to the derivative:

$$\frac{dy}{dx} = a \approx \frac{\Delta y}{\Delta x} \tag{2}$$

(for linear functions this approximation is actually exact). This method is slow and prone to errors, particularly in situations where data points span several orders of magnitude in one or both coordinates. It also gives no quantitative measure of uncertainty in the resulting values of the parameters. It is a good historical illustration, but modern physical sciences use faster and more accurate techniques.

On a related note, although the use of *Microsoft Excel* is widespread in undergraduate practicals, it has a peculiar psychological side effect: your tutors are seeing so many badly written student reports using *Excel,* that *Excel* itself has become firmly associated with poor quality in modern science. It would be wise to use *Origin* or *Matlab* for any research work that you would like to be taken seriously.

Linear least squares method

Given a set of data points $\{x_n, y_n\}$ and a linear model $f(x) = ax + b$, the formal statement of the problem of finding the "best" values of parameters a and b is to find the minimum of the deviation of the model from the experimental data points with respect to a and b. One possible measure of the deviation (called *error functional*) is the sum of squares of the differences between $f(x_n)$ and y_n:

$$\Omega(a,b) = \sum_n \left[f(x_n) - y_n \right]^2 \qquad (3)$$

Squares are used to make sure that negative deviations do not cancel positive ones and also because popular numerical optimisation methods, such as Newton-Raphson and BFGS, are particulary efficient with quadratic functions. After replacing $f(x)$ by its explicit linear form, we obtain the following expression for the error functional:

$$\Omega(a,b) = \sum_n \left[ax_n + b - y_n \right]^2 \qquad (4)$$

Because the values of x_n and y_n are fixed, Ω is a function of the model parameters a and b. Our task is to find the minimum of Ω with respect to these parameters. Because the function is quadratic and the coefficient of the leading term is positive, there is only one stationary point and that point is a minimum. At that minimum, the corresponding derivatives must be zero:

$$\begin{cases} \dfrac{\partial\Omega}{\partial a} = 0 \\ \dfrac{\partial\Omega}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \dfrac{\partial}{\partial a} \sum_n \left[ax_n + b - y_n \right]^2 = 0 \\ \dfrac{\partial}{\partial b} \sum_n \left[ax_n + b - y_n \right]^2 = 0 \end{cases} \Rightarrow \begin{cases} \sum_n \left[ax_n + b - y_n \right] x_n = 0 \\ \sum_n \left[ax_n + b - y_n \right] = 0 \end{cases} \qquad (5)$$

We will cover partial derivatives in more detail in a few lectures, for now they simply mean that all other variables, except the one being subject to differentiation, should be assumed to be constants. Differentiation of a quadratic expression is straightforward and, after rearranging some brackets and sums, we obtain the following system of equations for a and b:

$$\begin{cases} a\sum_n x_n^2 + b\sum_n x_n = \sum_n x_n y_n \\ a\sum_n x_n + bN = \sum_n y_n \end{cases} \qquad (6)$$

where N is the total number of experimental point pairs. After solving the second equation for b, substituting the result into the first equation, solving the first equation for a and then using the second equation to obtain b, we get the expressions for the data analysis technique that is known as the *linear least squares method*:

$$a = \frac{N\sum_{k=1}^{N} x_k y_k - \sum_{j=1}^{N} x_j \sum_{k=1}^{N} y_k}{N\sum_{k=1}^{N} x_k^2 - \left[\sum_{k=1}^{N} x_k \right]^2}; \qquad b = \frac{\sum_{j=1}^{N} y_j \sum_{k=1}^{N} x_k^2 - \sum_{j=1}^{N} x_j \sum_{k=1}^{N} x_k y_k}{N\sum_{k=1}^{N} x_k^2 - \left[\sum_{k=1}^{N} x_k \right]^2} \qquad (7)$$

Practical calculations are simplified considerably by the fact that the following four quantities are occurring repeatedly in Equations (7):

$$\sum_{k=1}^{N} x_k y_k, \qquad \sum_{k=1}^{N} x_k, \qquad \sum_{k=1}^{N} y_k, \qquad \sum_{k=1}^{N} x_k^2$$

As a practical example, consider the global average temperature masurements in the last fifteen years. The following data set may be obtained from Wolfram Alpha (*HadCRUT3GL* set for the global average temperature, used in particular, by the United Nations climate bodies):

Year	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
T/°F	57.934	58.037	58.055	58.005	58.068	57.965	57.924	57.785	57.997	58.060	57.812	57.925	58.023	58.050

Using year as the X axis and the temperature as the Y axis, we obtain the following values for the elementary sums of the linear least squares method:

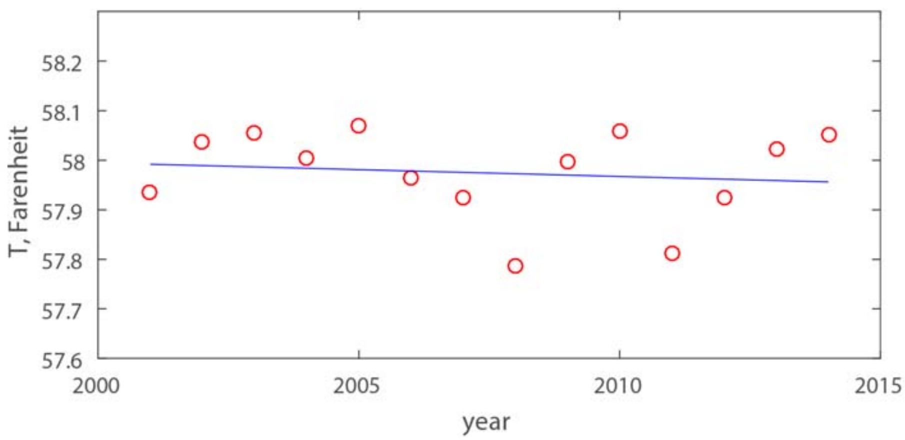
$$\sum_{k=1}^{N} x_k y_k = 1.6294 \cdot 10^6, \qquad \sum_{k=1}^{N} x_k = 28105$$

$$\sum_{k=1}^{N} y_k = 8.1164 \cdot 10^2, \qquad \sum_{k=1}^{N} x_k^2 = 56421015$$

After plugging these numbers into Equation (7), we obtain the following fitting function:

$$a = -0.00274725$$
$$b = 63.4894$$
$$\Downarrow$$
$$y = -0.00274725x + 63.4894$$

When we plot the data and the fit, the alarming global warming trend becomes immediately apparent:



Non-linear least squares method

Not all relationships between variables can be expressed as a linear function. In some cases, the error functional in Equation (3) contains a function that is non-linear:

$$\Omega(a,b,...)=\sum_n\left[f(a,b,...,x_n)-y_n\right]^2 \quad (8)$$

We will consider the case of multiple parameters in a few lectures, but for a single parameter:

$$\Omega(a)=\sum_n\left[f(a,x_n)-y_n\right]^2 \quad (9)$$

the same old Newton-Raphson optimization iteration works just fine:

$$a_{n+1}=a_n-\frac{\Omega'(a_n)}{\Omega''(a_n)} \quad (10)$$

Note that the iteration and the derivatives should all be taken with respect to the model parameter a rather than the variable x. A reasonable initial guess for a is also required and the iteration converges to the *nearest stationary point* – the task of confirming that it is indeed a minimum rests with the user.

Week 11 workshop exercises

Monk and Munro, 2nd edition: problems 8.10, 9.1, 9.4, 9.6, 9.7, 9.10.

Steiner, 2nd edition: section 21.12, problems 26(i), 27(i,ii).

Extra difficulty exercises for the brave

Reproduce the least squares fitting example from the lecture notes.