CHEM1047 – Week 4 Lecture 2 – Linear least squares method

☐ *Chapters 8 and 9 of Monk and Munro, "Maths for Chemistry", 2nd edition.*

☐ *Section 21.10 of Steiner, "The Chemistry Maths Book", 2nd edition.*

1. Motivation: protein concentration measurement

The principle behind spectrophotometric determination of concentrations is the *Beer-Lambert law* – a relationship between the fraction of the passing light absorbed by the sample and the concentration of the absorbing substance. The law can be derived from basic philosophy – the change dW in the intensity of light as it passes through a solution is proportional to the intensity itself, to concentration C and to the thickness of the sample layer dx. Because the light is absorbed, the change is negative:

$$dW = -kCWdx \quad \Rightarrow \quad \frac{dW}{dx} = -kCW \tag{1}$$

where k is a proportionality constant that is determined experimentally. It is easy to check by substitution (we will learn to solve such equations in due course) that the solution is:

$$W(x) = W_0 \exp(-kCx) \tag{2}$$

where W_0 is the initial light intensity and $W(x)$ is the intensity that has made it to the point x. It is clear from Equation (2) that light intensity in the sample decays exponentially. When powers of 10 rather than e are used in Equation (2):

$$W(x) = 10^{-\varepsilon Cx} W_0 \tag{3}$$

the proportionality coefficient ε is called *molar extinction coefficient*. It depends on the nature of the dissolved substance and on the solvent.

Applied to proteins, spectrophotometric assay requires an extra step: most proteins do not absorb visible light and therefore a chemical procedure is required to convert the protein into a substance that does. The simplest such substance is ninhydrin, but more sophisticated and accurate techniques are *Bradford assay* (using a specialised dye) and *BCA assay* (using a copper complex).

Table 1. *Spectrophotometric measurement results from Bradford and BCA assays of a bovine alpha-lactalbumin sample.*

Concentration (mg/ml)	BCA assay, absorbance at 562 nm	Bradford assay, absorbance at 595 nm
0.00	0.000	0.000
0.10	0.046	0.039
0.25	0.136	0.123
0.50	0.209	0.182
0.75	0.299	0.261
1.25	0.484	0.409
2.50	0.862	0.620
M1	0.118	0.071
M2	0.119	0.064
C1	0.156	0.082
C2	0.155	0.052

A typical outcome of a protein assay experiment is shown in Table 1. The table shows the absorbance of samples with known concentration (the *calibration set*) and the absorbance of the samples being tested. The lecture today is about systematic analysis of such datasets and, in this particular case, about extracting the unknown protein concentration from spectrophotometric data.

2. Transformation to linear coordinates

Functions encountered in models of physical and chemical processes are rarely linear – a good example is Equation (2) where the dependence on concentration is exponential. Non-linear data analysis can be exceedingly difficult; fortunately, many functions may be brought into a linear form

$$f(x) = ax + b \tag{4}$$

by a suitable substitution or other mathematical transformation. There is no general recipe for this process – it relies on experience and creativity. Examples:

function	linearisation	linear plot	intercept	slope
$y = a\exp(-bx)$	$\ln(y) = \ln(a) - bx$	$\ln(y)$ against x	$\ln(a)$	$-b$
$y = \dfrac{a}{b+cx}$	$\dfrac{1}{y} = \dfrac{b}{a} + \dfrac{c}{a}x$	$\dfrac{1}{y}$ against x	$\dfrac{b}{a}$	$\dfrac{c}{a}$
$y = ax^2 + 1$	$\sqrt{y-1} = x\sqrt{a}$	$\sqrt{y-1}$ against x	0	$\sqrt{a}$

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. The parameters may then be extracted from the intercept and slope obtained using one of the procedures described below.

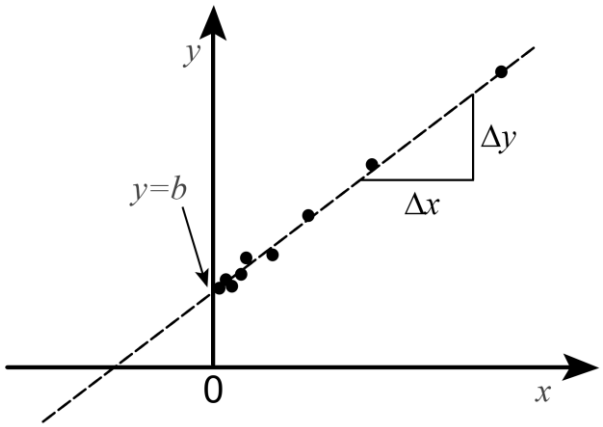
In the case of our spectrophotometric protein assay, Equation (3) is transformed into a linear form by introducing a new quantity A, called *absorbance*:

$$A = \log_{10}\frac{W_0}{W(x)} \quad \Rightarrow \quad A = \varepsilon C x \tag{5}$$

In this equation, ε is a constant. Therefore, if the length of the optical path x is also kept constant, the absorbance becomes a linear function of concentration, creating a useful analytical method.

3. 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 paper and to draw a line through it by hand:



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

$$\frac{dy}{dx} = a \approx \frac{\Delta y}{\Delta x} \tag{6}$$

(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.

4. 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 this deviation is the sum of squares of the differences between $f(x_n)$ and y_n:

$$\Omega(a,b) = \sum_n \left[f(x_n) - y_n \right]^2 \tag{7}$$

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 LBFGS, are particulary efficient with quadratic functions. After replacing $f(x)$ by its explicit linear form, we obtain the following expression for the error functional:

$$\Omega(a,b) = \sum_n \left[ax_n + b - y_n \right]^2 \tag{8}$$

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 first derivatives must be zero:

$$\begin{cases} \dfrac{\partial \Omega}{\partial a} = 0 \\ \dfrac{\partial \Omega}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \dfrac{\partial}{\partial a} \sum_n \left[ax_n + b - y_n \right]^2 = 0 \\ \dfrac{\partial}{\partial b} \sum_n \left[ax_n + b - y_n \right]^2 = 0 \end{cases} \Rightarrow \begin{cases} \sum_n \left[ax_n + b - y_n \right] x_n = 0 \\ \sum_n \left[ax_n + b - y_n \right] = 0 \end{cases} \tag{9}$$

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:

$$\begin{cases} a\sum_n x_n^2 + b\sum_n x_n = \sum_n x_n y_n \\ a\sum_n x_n + bN = \sum_n y_n \end{cases} \tag{10}$$

where N is the total number of experimental point pairs. After solving this system, we get the expressions for the data analysis technique known as the *linear least squares method*:

$$a = \frac{N\sum_{k=1}^{N} x_k y_k - \sum_{j=1}^{N} x_j \sum_{k=1}^{N} y_k}{N\sum_{k=1}^{N} x_k^2 - \left[\sum_{k=1}^{N} x_k\right]^2}; \qquad b = \frac{\sum_{j=1}^{N} y_j \sum_{k=1}^{N} x_k^2 - \sum_{j=1}^{N} x_j \sum_{k=1}^{N} x_k y_k}{N\sum_{k=1}^{N} x_k^2 - \left[\sum_{k=1}^{N} x_k\right]^2} \tag{11}$$

Practical calculations are simplified considerably by the fact that the following four quantities are occurring repeatedly in Equations (11):

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

Returning now to our spectrophotometric protein assay, we can supply the data from Table 1 to a suitable software package (Origin is much recommended) and ask it to extract the coefficients of the linear relation connecting absorbance and concentration:

$$A = \alpha C + \beta \tag{12}$$

where $\alpha = \varepsilon x$ and β is supposed to be zero, but isn't in practice because real-life instruments have unavoidable background absorption due to, for example, scratches on cuvette surfaces.

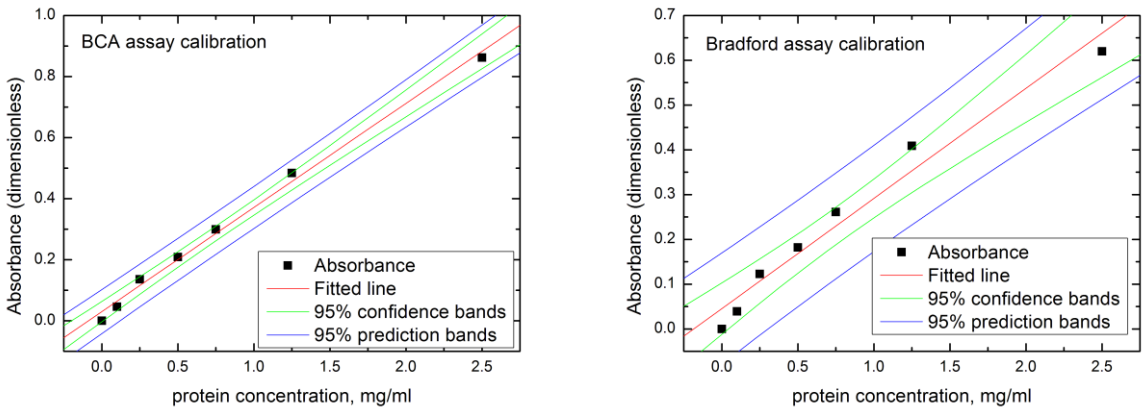


Figure 1. *Least squares fitting results for the calibration data of BCA and Bradford assays. The resulting parameters are: a= 0.34±0.01, b= 0.03±0.01 for BCA and a=0.25±0.02, b= 0.04±0.02 for Bradford.*

Figure 1 shows typical least squares fits, performed in *Origin*. We will deal with the error analysis (coloured bands) in a few lectures' time. The parameters of the fits may now be used to determine concentrations from absorbance measurements.