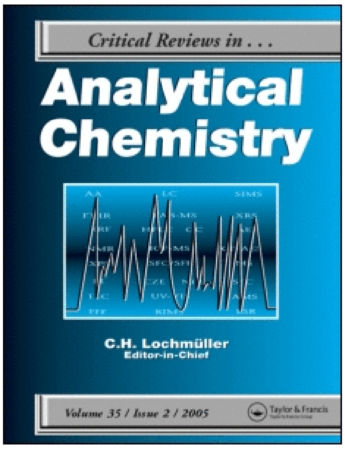


Critical Reviews in Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713400837

Curve Fitting with Least Squares

Robert de Levie

Online publication date: 03 June 2010

To cite this Article de Levie, Robert(2000) 'Curve Fitting with Least Squares', Critical Reviews in Analytical Chemistry, 30: 1, 59 — 74

To link to this Article: DOI: 10.1080/10408340091164180

URL: http://dx.doi.org/10.1080/10408340091164180

Curve Fitting with Least Squares

Robert de Levie

Department of Chemistry, Bowdoin College, Brunswick, ME 04011

I. INTRODUCTION

Least-squares analysis is based on the so-called ‘normal’ distribution of experimental uncertainties formulated by Abraham de Moivre, a friend of Newton. De Moivre was a mathematical genius who never found permanent employment, and earned a meager living in London as a math tutor for the children of the wealthy. The least squares method itself was subsequently developed by Legendre and Gauss, who realized how the ‘normal’ distribution could be used for data analysis.

For a long time, least squares analysis was the near-exclusive domain of statisticians, who used long tables to evaluate their data. The field advanced through the introduction of matrix algebra, although that tended to make it less accessible to non-statisticians. More recently, these matrix methods were implemented in specialized software and, finally, in general-purpose computational programs such as spreadsheets. Now that they are easily accessible, least squares methods have become ubiquitous in science.

In Section II we review some of the properties of the normal distribution and related concepts. We also see how multiple, small, random errors that each follow a ‘normal’ distribution lead to the least squares criterion. In Section III we then apply this method in order to fit experimental data to model expressions.

The mathematics of least squares are greatly simplified when we can assume that the experimental deviations are essentially confined to *one* parameter, the ‘dependent’ variable. (We will make this simplifying assumption throughout this review, except in Section VII.) There can be any number of ‘independent’ variables, which assumedly do not contribute to the experimental uncertainty. The mathematics are especially simple with a single ‘independent’ variable that is uniformly spaced, as in the special case of *equidistant* data discussed in Section IV. Weighted least squares are discussed in Section V. The ready availability of multi-parameter *non-linear* least squares routines in modern spreadsheets and computer-based statistical packages provides yet another approach to curve fitting, as described in Section VI. And in Section VII we finally lift the constraining assumption that only one parameter contains random experimental fluctuations.

Least-squares have found many other uses, such as to discover the presence (or absence) of linear relations between parameters (expressed in terms of the often misapplied *linear correlation coefficient r*), or to find the more subtle linear combinations of data that form the basis of near-infrared spectrometric analysis. For the latter, the reader should consult books on *partial least squares* or *principal component analysis*, since they fall outside the purview of the present review.

II. REPLICATE MEASUREMENTS

Random experimental fluctuations can be categorized in one of three classes. When the

1040-8347/00/$.50

fluctuations are discrete, and have discrete consequences (as in throwing dice, which can only be done an integer number of times, and each time leads to an integer result), they tend to follow a binomial distribution, or (if the various outcomes have unequal probabilities, as when the dice are loaded) one of its derivatives. When the fluctuations occur randomly in time, but have discontinuous consequences (as with radioactive decay, which can happen at any moment, but results in a discrete change in atomic number and/or mass), the prototypical distribution is Poissonian. Most chemical observations fall in the third category, comprising those experiments where both the experiments and the results are continuous. In this case, the 'normal' distribution is the applicable prototype. This review is restricted to the latter (third) category.

A. The Normal Distribution

In a normal distribution, the probability P of an experimental observation subject to small random fluctuations is given by an expression of the form

$$P = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(y-\bar{y})^2}{2\sigma^2}\right] \qquad (1)$$

where $\bar{y}$ is the *average* or *mean* of the observations, and σ is its *standard deviation*. The pre-exponential factor $1/\sigma\sqrt{2\pi}$ normalizes the distribution, ensuring that integration of the probabilities P over all possibilities yields the value 1. The distribution (1) has a symmetrical, bell-shaped form, with equal probabilities of positive and negative deviations of the same absolute magnitude $|y-\bar{y}|$, and with smaller deviations from the average more likely to occur than larger ones.

The normal distribution (1) can be shown to be the limit of a Poissonian distribution when the discrete steps in the latter become vanishingly small with respect to the experimental resolution. This is a common state of affairs in chemistry, where in principle many observables have a molecular, atomic, or ionic discreteness. In most chemical measurements the number of particles causing the observed effect is so large, and their individual contributions so small, that for all practical purposes the observed signals and their fluctuations are continuous. Indeed, many experimental observations in chemical analysis approximately follow a normal distribution. As a result, absent evidence to the contrary, the corresponding statistics are usually assumed to apply, hence the name 'normal'.

To illustrate the method of least squares, we consider a set of observations *y* from which we want to extract the value of its average, $\bar{y}$. We will call the experimental deviations $y-\bar{y}$ from the (as yet unspecified) average $\bar{y}$ the *residuals*. The least squares method now minimizes the sum of the squares of the residuals with respect to the unknown parameter, $\bar{y}$, i.e., $\sum(y-\bar{y})^2$. This is done by equating to zero the derivative of the sum of the squares of the residuals with respect to $\bar{y}$, i.e., through

$$\begin{aligned} 0 &= \frac{d}{d\bar{y}}\sum_{i=1}^{N}(y-\bar{y})^2 = \sum_{i=1}^{N}\frac{d(y_i-\bar{y})^2}{d\bar{y}} \\ &= -2\sum_{i=1}^{N}(y_i-\bar{y}) = -2\sum_{i=1}^{N}y_i + 2N\bar{y} \end{aligned} \qquad (2)$$

so that

$$\bar{y} = \frac{\sum_{i=1}^{N} y_i}{N} \qquad (3)$$

which indeed yields the common result for the average. Because the summation involves a finite number N of terms, the order of summation and derivative taking can be inverted.

In replicate measurements, the *standard deviation* is a measure of the *repeatability* of the observation, and is therefore used to specify the *precision* of the measurement. The *population* standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}{N}} \qquad (4)$$

is a measure of the width of the distribution, while the *sample* standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}{N-1}} \qquad (5)$$

measures the (ir)reproducibility of single data points. In chemical measurements, we almost always deal with samples rather than with the entire population, and we will therefore use sample rather than population statistics. The average of all N replicate measurements then has a standard deviation

$$\bar{s} = \frac{s}{\sqrt{N}} = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2}{N(N-1)}} \qquad (6)$$

The *variance* is defined as the square of the corresponding standard deviation.

B. The Least Squares Criterion

We now consider making a number of repeat measurements, assuming (a) that these measurements have mutually independent deviations from the true mean, and (b) that the deviations follow a single, normal distribution. Let the first observation be off from the true mean by an amount Δ_1. The probability $P(\Delta_1)$ that this will happen then follows from (1) as $P(\Delta_1) = \left(1/\sigma\sqrt{2\pi}\right)\exp\left[-\Delta_1^2/2\sigma^2\right]$. Similarly we have for the second observation $P(\Delta_2)$, $= \left(1/\sigma\sqrt{2\pi}\right)\exp\left[-\Delta_2^2/2\sigma^2\right]$, etc. The probability that we will encounter a set of deviations Δ_1, Δ_2, Δ_3, … , etc. is the product of their individual probabilities $P(\Delta_1)$, $P(\Delta_2)$,$P(\Delta_3)$, … , etc., or $P(\Delta_1,\Delta_2,\Delta_3,\ldots) = P(\Delta_1).P(\Delta_2).P(\Delta_3)\ldots = \left(1/\sigma\sqrt{2\pi}\right)\exp\left[-\left(\Delta_1^2+\Delta_2^2+\Delta_3^2+\ldots\right)/2\sigma^2\right]$.

In a normal distribution, the probability of encountering a small deviation is larger than that of encountering a larger one. Consequently, the overall probability will be smaller the larger the deviations, so that $P(\Delta_1, \Delta_2, \Delta_2,\ldots)$ goes through a maximum when $(\Delta_1^2 + \Delta_2^2 + \Delta_2^2 + \cdots)$ is minimal. Thus, in order to determine the most probable location of the maximum in the normal distribution, we determine that value for which the sum of the squares of the deviations Δ_i is minimal. That is the least-squares criterion.

III. FITTING DATA TO A MODEL EQUATION

The subject of this review is the fitting of data sets to an assumed mathematical equation. Here we can use the same principle of minimizing the sum of the squares of the residuals to optimize our fit. However, in this case the standard deviation acquires a *different meaning*: it no longer measures the repeatability of replicate measurements, but instead indicates how well the data fit the assumed model expression. This obviously blurs the distinction between precision and accuracy, because the standard deviation now reflects the amount of random noise in the data as well as the appropriateness of the assumed model expression.

A. The Proportionality

The simplest case is that of the proportionality $y = ax$, i.e., of data that, in the absence of experimental fluctuations, would lie on a straight line through the origin. Given a set of experimental data pairs x_i, y_i, where the experimental uncertainties are all located in the y_i and are assumed to follow a normal distribution, we find the deviations Δ_i as $y_i - ax_i$ because the x_i is assumed to be error-free, so that ax_i represents the correct value. The value of a is not yet known, and in fact is the parameter we want to determine. As in Equation 2, we determine a by minimizing the sum of the squares of the deviations Δ_i, i.e., by setting the derivative of $\Sigma\ \Delta_i^2 = \Sigma\ (y_i - ax_i)^2$ with respect to a equal to zero:

$$0 = \frac{d}{da}\sum_{i=1}^{N}\left(y_i - ax_i\right)^2 = \sum_{i=1}^{N}\frac{d\left(y_i - ax_i\right)^2}{da}$$

$$= -2\sum_{i=1}^{N}x_i\left(y_i - ax_i\right) = -2\sum_{i=1}^{N}x_iy_i + 2a\sum_{i=1}^{N}x_i^2 \tag{7}$$

so that can then be determined from

$$a = \sum_{i=1}^{N}x_iy_i \Bigg/ \sum_{i=1}^{N}x_i^2 \tag{8}$$

We now distinguish two variances, s_y and s_a, where the former measures the deviations of y_i from the proportionality:

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

$$= \frac{\sum_{i=1}^{N}x_i^2\sum_{i=1}^{N}y_i^2 - \left(\sum_{i=1}^{N}x_iy_i\right)^2}{(N-1)\sum_{i=1}^{N}x_i^2} \tag{9}$$

whereas s_a indicates the non-systematic uncertainty in the value of a:

$$s_a^2 = s_y^2\sum_{j=1}^{N}\left(\frac{\partial a}{\partial y_j}\right)^2 \tag{10}$$

which upon substitution of (8) yields

$$s_a^2 = s_y^2\sum_{j=1}^{N}\left(\frac{\partial a}{\partial y_j}\right)^2 = s_y^2\sum_{j=1}^{N}\left(\frac{\partial}{\partial y_j}\frac{\sum_{i=1}^{N}x_iy_i}{\sum_{i=1}^{N}x_i^2}\right)^2$$

$$= s_y^2\sum_{j=1}^{N}\left(\frac{x_j}{\sum_{i=1}^{N}x_i^2}\right)^2 = \frac{s_y^2\sum_{j=1}^{N}x_j^2}{\left(\sum_{i=1}^{N}x_i^2\right)^2} = \frac{s_y^2}{\sum_{i=1}^{N}x_i^2} \tag{11}$$

From here on we will simplify the notation by deleting the subscripts i, except when they are needed to avoid ambiguity, as in Sections III.E and VII. In order to compute a, s_y, and s_a, we need to compile the three sums Σx^2, Σxy, and Σy^2, and count the number N of data pairs entered. This is easily done in a computer program, on a spreadsheet, or even on a calculator with some memory. Spreadsheets will make the computation automatically after the data points have been entered, as long as one specifies a line through the origin. For example, Excel provides three different linear least-squares methods to fit experimental data automatically to a proportionality: (1) the rather terse but automatically self-updating LINEST function, (2) the Regression macro in the Analysis Toolpak, which yields more statistical information than you may want, and can even be set to plot the raw data, their fit, and the corresponding residuals, and (3) Trendline, which operates directly on a graph of the data, but only provides the fitting parameters.

B. The Line

A general line can be represented by the equation $y = a_0 + a_1x$ with slope a_1 and intercept a_0. In this case there are *two* parameters to be fitted to the data, and we must therefore minimize the sum of the squares of the deviations $\Delta = y - a_0 - a_1 x$, where x is again assumed to be error-free, while a_0 and a_1 are constants of yet to be determined value. We now find the minimum by setting the derivatives with respect to a_0 and a_1 equal to zero:

$$\begin{aligned} 0 &= \frac{d}{da_0}\sum\left(y - a_0 - a_1 x\right)^2 \\ &= -2\sum y + 2a_0 N + 2a_1\sum x \end{aligned} \tag{12}$$

where $\Sigma a_0 = a_0 N$ when we use N data pairs x,y, and

$$\begin{aligned} 0 &= \frac{d}{da_1}\sum\left(y - a_0 - a_1 x\right)^2 \\ &= -2\sum xy + 2a_0\sum x + 2a_1\sum x^2 \end{aligned} \tag{13}$$

Solving these two simultaneous Eqs. 12 and 13, for the two unknowns a_0 and a_1 yields

$$a_0 = \frac{\sum x^2 \sum y - \sum x \sum xy}{N\sum x^2 - \left(\sum x\right)^2} \tag{14}$$

and

$$a_1 = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - \left(\sum x\right)^2} \tag{15}$$

For the variance of the fit we find

$$s_y^2 = \frac{\sum\left(y - a_0 - a_1 x\right)^2}{N - 2} \tag{16}$$

where the residuals are $\Delta = y - a_0 - a_1 x$, and the number of degrees of freedom is $N - 2$ because the two constants, a_0 and a_1, each consume one degree of freedom. The variances in a_0 and a_1 are given by

$$\begin{aligned} s_{a_0}^2 &= s_y^2 \sum_{i=1}^{N}\left(\frac{\partial a_0}{\partial y_i}\right)^2 \\ &= \frac{\sum x^2 \sum\left(y - a_0 - a_1 x\right)^2}{(N-2)\left(N\sum x^2 - \left(\sum x\right)^2\right)} \\ &= \frac{s_y^2 \sum x^2}{N\sum x^2 - \left(\sum x\right)^2} \end{aligned} \tag{17}$$

$$\begin{aligned} s_{a_1}^2 &= s_y^2 \sum_{i=1}^{N}\left(\frac{\partial a_1}{\partial y_i}\right)^2 \\ &= \frac{N\sum\left(y - a_0 - a_1 x\right)^2}{(N-2)\left(N\sum x^2 - \left(\sum x\right)^2\right)} \\ &= \frac{N s_y^2}{N\sum x^2 - \left(\sum x\right)^2} \end{aligned} \tag{18}$$

Again, there is usually no need to evaluate the sums involved in 14 through 18 when one has access to a least-squares computer program, or to a spreadsheet. For instance, in Excel, both the function LINEST and the Regression macro in the Analysis Toolpak will fit data to a line, and will provide the corresponding standard deviations s_y, s_{a_0}, and s_{a_1}, while Trendline will fit data presented in graphical form.

C. The Polynomial

The above approach is readily extended to polynomials of the form $y = a_0 + a_1 x + a_2 x^2 + \cdots + a_j x^j$ by successively equating to zero all j derivatives $\partial\Sigma\Delta^2/\partial a_j$ where $\Delta = y -$

$(a_0 + a_1x + a_2x^2 + \ldots + a_jx^j)$. The resulting j equations in j unknowns are most readily solved in terms of determinants. For example, for fitting to a quadratic, we have $j = 2$, and the resulting expressions are

$$a_0 = \left|\begin{matrix} \sum x^4 & \sum x^3 & \sum x^2y \\ \sum x^3 & \sum x^2 & \sum xy \\ \sum x^2 & \sum x & \sum y \end{matrix}\right| \Big/ D \tag{19}$$

$$a_1 = \left|\begin{matrix} \sum x^4 & \sum x^2y & \sum x^2 \\ \sum x^3 & \sum xy & \sum x \\ \sum x^2 & \sum y & N \end{matrix}\right| \Big/ D \tag{20}$$

$$a_2 = \left|\begin{matrix} \sum x^2y & \sum x^3 & \sum x^2 \\ \sum xy & \sum x^2 & \sum x \\ \sum y & \sum x & N \end{matrix}\right| \Big/ D \tag{21}$$

with

$$D = \left|\begin{matrix} \sum x^4 & \sum x^3 & \sum x^2 \\ \sum x^3 & \sum x^2 & \sum x \\ \sum x^2 & \sum x & N \end{matrix}\right| \Big/ \tag{22}$$

While this may suggest a higher level of complexity, in practice it is not, because most software programs use matrix algebra anyway, but completely hide this fact from the user. For example, on most spreadsheets, one can just make an extra column for x^2, whereupon the regression routine will automatically, without any further ado, fit the data to $y = a_0 + a_1x + x^2$. There can be many such extra columns, which the routine will accommodate without complaints as long as they are contiguous with the column for x.

D. Multiple Parameters

As far as the mathematics are concerned, there is no difference between fitting data to $y = a_0 + a_pp + a_qq + a_r r + \ldots + a_zz$ or to $y = a_0 + a_1x + a_2x^2 + \ldots + a_j x^j$, and using the least-squares criterion for each unknown parameter leads to the required number of simultaneous equations, and ultimately to their general solution in matrix form. Again, the software usually need not be modified, and will simply return the appropriate parameters with the corresponding standard deviations. Note that the independent variables p, q, r, …, z need not be independent (they certainly are not when $q = p^2$, $r = p^3$, etc., as in the case of a polynomial) but can be any function of a known, error-free parameter, such as log(x), exp[x^{-2}], etc.

As a practical example, consider the spectral analysis of a mixture of a number of colored species. When the (absorbance or emission) spectra of the individual species making up the mixture are known, we can consider the mixture spectrum as the algebraic sum of the constituent spectra, each with an adjustable parameter reflecting the concentration of that species in the mixture. This is a straightforward problem of fitting multiple parameters that can be solved readily by least-squares programs. On a spreadsheet, e.g., make columns for the spectrum of the mixture, and for the spectra of its pure constituents, then call the least-squares routine, and feed it the mixture spectrum as the dependent variable, and the block of comparison spectra as its independent variables. The least-squares routine will then yield the concentrations in the mixture (in terms of those in the single-species spectra), as well as the associated standard deviations. It really is that simple.

The only requirement is that there are single-species spectra for all species contributing significantly to the absorbance or emittance of the mixture. If this is not the case, the residual spectrum will indicate that there is a problem. The single-species spectra may be of chemically independent species, or of species related through proton or ligand equilibria, and there may even be reference spectra of species not present in the mixture. There is no need to single out any particular wavelengths. However, for the method to work, spectra must be available for every species that significantly contributes to the spectrum of the mixture. Moreover, all spectra should cover the same wavelength range, and have the same (though not necessarily uniform) data spacing. By using the entire spectrum, the influence of noise on the final answer is much reduced.

E. Designer Least Squares

Sometimes one needs to fit data to several equations that have one or more fitting parameters in common. To take a simple example, say that we have data that should fit onto two *parallel* lines. We can of course fit the two sets of data separately, each to its own line, but then the slopes will be different, and we will have to arbitrate between their two numerical values. Below we will illustrate how one can, instead, fit these data to two intercepts but one common slope. The method is readily extended to any number of polynomials, and any number of constraints (such as curves passing through a common point), and merely requires that the matrix equations be tailored to the problem at hand.

Let the two sets of data in our example be labeled p and q, and let us fit them to $y_p = a_p + a_1x$ and $y_q = a_q + a_1x$, respectively. We write

$$\begin{aligned} 0 &= \sum_p \frac{\partial}{\partial a_p}\left(y - a_p - a_1x\right)^2 \\ &= -\sum_p y + a_pN_p + a_1\sum_p x \end{aligned} \qquad (23)$$

$$\begin{aligned} 0 &= \sum_q \frac{\partial}{\partial a_q}\left(y - a_q - a_1x\right)^2 \\ &= -\sum_q y + a_qN_q + a_1\sum_q x \end{aligned} \qquad (24)$$

$$\begin{aligned} 0 &= \sum_p \frac{\partial}{\partial a_1}\left(y - a_p - a_1x\right)^2 \\ &\quad + \sum_q \frac{\partial}{\partial a_1}\left(y - a_q - a_1x\right)^2 \\ &= -\sum_p xy - \sum_q xy + a_p\sum_p x \\ &\quad + a_q\sum_q + a_1\sum_p x^2 + a_1\sum_p x^2 \end{aligned} \qquad (25)$$

where 25 replaces the equations for the individual slopes of the two data sets when considered separately. We rewrite 23 through 25 as

$$\begin{array}{ccccc} a_pN_p & & +a_1\sum\limits_p x & = & \sum\limits_p y_p \\ & a_qN_q & +a_1\sum\limits_q x & = & \sum\limits_q y_q \\ a_p\sum\limits_p x & +a_q\sum\limits_q x & +a_1\left(\sum\limits_p x^2 + \sum\limits_q x^2\right) & = & \sum\limits_p xy_p + \sum\limits_q xy_q \end{array} \qquad (26)$$

so that the three parameters a_p, a_q, and a_1 follow from Cremer's rule as the ratios of two determinants,

$$a_p = \left|\begin{matrix} \sum\limits_p y & 0 & \sum\limits_p x \\ \sum\limits_q y & N_q & \sum\limits_q x \\ \sum\limits_p xy + \sum\limits_q xy & \sum\limits_q x & \sum\limits_p x^2 + \sum\limits_q x^2 \end{matrix}\right| \Bigg/ D \qquad (27)$$

$$a_q = \left.\begin{vmatrix} N_p & \sum\limits_p y & \sum\limits_p x \\ 0 & \sum\limits_q y & \sum\limits_q x \\ \sum\limits_p x & \sum\limits_p xy + \sum\limits_q xy & \sum\limits_p x^2 + \sum\limits_q x^2 \end{vmatrix} \middle/ D \right. \tag{28}$$

with

$$a_1 = \left.\begin{vmatrix} N_p & 0 & \sum\limits_p y \\ 0 & N_q & \sum\limits_q y \\ \sum\limits_p x & \sum\limits_q x & \sum\limits_p xy + \sum\limits_q xy \end{vmatrix} \middle/ D \right. \tag{29}$$

$$D = \begin{vmatrix} N_p & 0 & \sum\limits_p x \\ 0 & N_q & \sum\limits_q x \\ \sum\limits_p x & \sum\limits_q x & \sum\limits_p x^2 + \sum\limits_q x^2 \end{vmatrix} \tag{30}$$

Figure 1 illustrates the application of these results, and compares fitting the data to two parallel lines with fitting them to separate lines with independent slopes.

In implementing 27 through 30 we can no longer rely on ready-made software. However, on a spreadsheet, the solution is still easily obtained when we make columns for x, y, x^2, and xy, with an empty row between the two data sets p and q. We then let the spreadsheet compute the sums $\Sigma_p x$, $\Sigma_q x$, $\Sigma_p y$, $\Sigma_q y$, $\Sigma_p x^2$, $\Sigma_q x^2$, $\Sigma_p xy$, and $\Sigma_q xy$. The determinants in 27 through 30 need not frighten us, since modern spreadsheets handle them with aplomb. For example, in Excel the determinant can be evaluated with the array function MDETERM(). In Section VI we will encounter another (and often simpler) approach to designer least squares.

IV. FITTING EQUIDISTANT DATA

Data taken automatically by instruments are often **equidistant** in the independent parameter, such as time, wavelength, voltage, etc. In this context, equidistant means that the independent parameter increases by constant increments; there is no corresponding constraint on the dependent variable. When the least squares method is applied to equidistant data, it can be simplified dramatically, and can be reduced to a simple set of multipliers called **convoluting integers**, thereby completely avoiding the need for matrix algebra. The method goes back to Sheppard (1914)

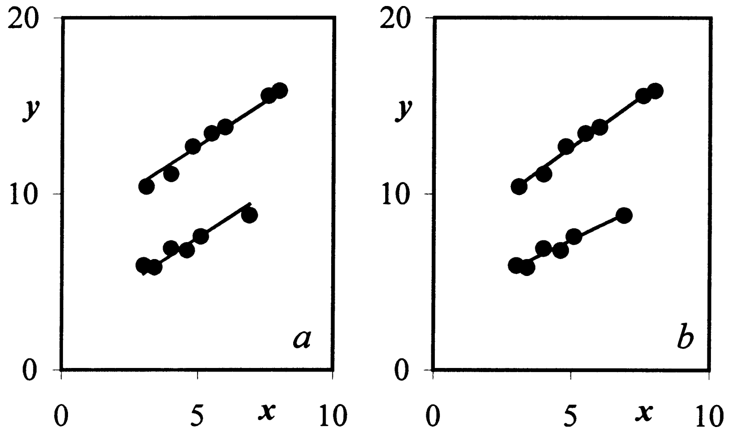


FIGURE 1. **(a):** a set of data fitted to two *parallel* straight lines. **(b):** The same data fitted to two separate straight lines.

and Sherriff (1920), and was brought to the attention of analytical chemists by Savitzky and Golay (1964) at the time that computers were beginning to become generally available. We will here use the simple example of fitting five adjacent, equidistant data points to a parabola to illustrate how this method works.

We denote the five points as x_{-2},y_{-2}, x_{-1},y_{-1}, x_0,y_0, x_1,y_1, and x_2,y_2. The x-values are equidistant, with increments δ. We now subtract x_0 from all x-values, so that $x_{-2} = -2\delta$, $x_{-1} = -\delta$, $x_0 = 0$, $x_1 = \delta$, and $x_2 = 2\delta$; such a shift in the x-axis is convenient yet is inconsequential for the final result. The experimental data pairs to be fitted then become $-2\delta,y_{-2}$, $-\delta,y_{-1}$, $0,y_0$, δ,y_1, and $2\delta,y_2$.

Explicit expressions for fitting data to a parabola were given in 19 through 22 in terms of a number of sums. We note that D as defined by 22 does not contain y, so that D can be evaluated simply on the basis of the x-values. Furthermore, because of the way we have shifted those x-values, the sums in odd powers of x must be zero. We therefore have

$$\begin{aligned}
&N = 5\\
&\sum x = 0\\
&\sum x^2 = (-2\delta)^2 + (-\delta)^2 + (0)^2 + (\delta)^2 + (2\delta)^2\\
&\qquad = 10\,\delta^2\\
&\sum x^3 = 0\\
&\sum x^4 = (-2\delta)^4 + (-\delta)^4 + (0)^4 + (\delta)^4 + (2\delta)^4\\
&\qquad = 34\,\delta^4
\end{aligned} \tag{31}$$

so that

$$\begin{aligned}
D &= \sum x^4\left[N\sum x^2 - \left(\sum x\right)^2\right]\\
&\quad + \sum x^3\left[\sum x \sum x^2 - N\sum x^3\right]\\
&\quad + \sum x^2\left[\sum x^3 \sum x - \left(\sum x^2\right)^2\right]\\
&= N\sum x^4 \sum x^2 - \left(\sum x^2\right)^3\\
&= 5 \times 34\,\delta^4 \times 10\,\delta^2 - \left(10\,\delta^2\right)^3 = 700\,\delta^6
\end{aligned} \tag{32}$$

Equations 19 through 21 contain terms in y, but none of higher order than y. We can therefore evaluate them in terms of their y-values as

$$\begin{aligned}
&\sum y = y_{-2} + y_{-1} + y_0 + y_1 + y_2\\
&\sum xy = \left(-2y_{-2} - y_{-1} + y_1 + 2y_2\right)\delta\\
&\sum x^2y = \left(4y_{-2} + y_{-1} + y_1 + 4y_2\right)\delta^2
\end{aligned} \tag{33}$$

Upon substitution into 19 through 21 we then find the following, explicit expressions for a_0, a_1, and a_2:

$$a_0 = \left(-3y_{-2} + 12y_{-1} + 17y_0 + 12y_1 - 3y_2\right)/35 \tag{34}$$

$$a_1 = \left(-2y_{-2} - y_{-1} + y_1 + 2y_2\right)/10\delta \tag{35}$$

$$a_2 = \left(2y_{-2} - y_{-1} - 2y_0 - y_1 + 2y_2\right)/14\delta^2 \tag{36}$$

With these parameters we can now determine a smoothed value for $y = a_0 + a_1x + a_2x^2$ at $x = 0$ as $y_{x=0} = a_0$. Likewise, the first derivative of y at $x = 0$ is $(dy/dx)_{x=0} = a_1$, and its second derivative is $(d^2y/dx^2)_{x=0} = 2a_2$. Consequently, we can compute these values simply by multiplying the various y-values by integer ratios: for smoothing a five-point sequence to a parabola, those integer ratios are –3, +12, +17, +12, and –3, all divided by the common denominator 35, see Equation 34. Likewise, for the first derivative, they follow from (35) as –2, –1, 0, 1, and 2, with the common denominator 10 δ, and for the second derivative from 36 as 2, –1, –2, –1, and 2, with the common denominator 7 δ^2 (because the second derivative is *twice* the value of a_2).

This method is ideally suited for a moving polynomial fit, in which a relatively short polynomial slithers along a much longer data set, in order to determine the smooth value, derivative, etc. at the mid-point of that polynomial. In order for the *x*-value of the mid-point to coincide with that of an existing data point, one always selects a polynomial with an odd number of data points.

Because the moving polynomial fits only a small portion of the data at a time, it makes only a rather weak assumption regarding the structure of the underlying data. It is readily implemented on a spreadsheet. Savitzky and Golay (1964) listed convoluting integers, but their tabulation contained many errors, which have sometimes been incorporated in commercial software. A list of corrections was given by Steinier et al. ((1972), and a table which includes the corrected data can be found, e.g., in de Levie (1997). Madden (1978) has given general formulas for the computation of the convoluting integers.

In using the moving polynomial method, the longer the moving polynomial, and the lower the order of the polynomial, the more noise is removed. On the other hand, these are also the conditions that lead to maximal signal distortion. When the method is applied to noisy data, a compromise must therefore be found between noise reduction and signal distortion. Such a compromise will, in general, depend on the properties of the data set. When a data set contains both narrow and broad features, no single choice of polynomial length and order may be optimal for the entire set. In that case, one can apply a program recently developed by Barak (1995) in which the polynomial length is fixed but its order is optimized anew at every location, each time the fitting polynomial moves one step.

V. WEIGHTED LEAST SQUARES

Least-squares analysis works fine with polynomials, and many phenomena can be represented in that form. Sometimes, however, polynomials provide awkward (i.e., unwieldy or slowly converging) representations of theoretical expressions, while simple transformations exist to bring the theory in polynomial form. An example is the exponential decay, $y = y_0 \exp[-kt]$, which is readily linearized by taking its logarithm, $\ln y = \ln y_0 - kt$. When experimental data are fitted to this transformed equation, the sum of the squares of the deviations in the transformed variable $Y = \ln y$ are minimized. This may be appropriate when the uncertainty is directly proportional to the magnitude of the signal *y*. However, when the uncertainty is essentially independent of *y*, the proper procedure would be to minimize the sum of the squares of the deviations in *y* rather than of those in *Y*. This can be handled in the following way.

Implicit in all analyses of experimental uncertainties is the assumption that these uncertainties are relatively small (in comparison to the signal), i.e., $\Delta y \ll y$ and, likewise, $\Delta Y \ll Y$. In that case we can use

$$\frac{\Delta Y}{\Delta y} \approx \frac{dY}{dy} \tag{37}$$

and we can modify the least-squares routine to minimize the sum of squares of Δy instead of ΔY by introducing weights *w* into the formalism, such that

$$w = \frac{1}{\left(dY/dy\right)^2} \tag{38}$$

Many commercial software packages for data analysis include weighted least squares. Unfortunately, spreadsheets do not (yet) do so, although macros are available (e.g., de Levie 2000) to remedy this deficiency.

We note that the need for weighting depends on the nature of the uncertainties involved. In the above example, weighting may be needed when the experimental uncertainties in the data are best represented as absolute ones, whereas it should not be used when the dominant uncertainties are relative ones, with magnitudes approximately proportional to y.

Weighted least squares are not always problem-free. Weighting factors given by (38) are often even powers of the untransformed signal, such as y^2 or y^4, which are always positive even though the corresponding y-values may have zero average. As a consequence, random noise in regions where the signal is small may contribute significantly to the sum of squares, and may then distort the analysis, as illustrated below.

Say that we want to fit a set of experimental data to a Lorentzian of the form $y = a_0/[(x - a_1)^2 + a_2]$. In this case, writing y as a power series in x is awkward, but we can simply transform the equation into the quadratic form $Y = 1/y = (a_1^2 + a_2)/a_0 - 2a_1x/a_0 + x^2/a_0$, with a corresponding weight $w = 1/(dY/dy)^2 = y^4$, see de Levie (1986). As shown in Figure 2a, we can indeed fit experimental data to a Lorentzian as long as the noise is relatively small. When the noise amplitude becomes significant, the noise in those parts of the curve where the underlying signal is insignificant will start to dominate the sums of squares (because y^4 will be positive even when y has zero mean), and the fit becomes unsatisfactor (see Figure 2b).

VI. NONLINEAR LEAST SQUARES

The examples discussed so far all involve linear least squares, in which the dependent variable y is expressed in terms of a linear function of the fitting parameters a_i (whence the name), and the analysis leads to a closed-form expression for these parameters. Here we describe another approach, which often avoids some of the pitfalls of linear least squares methods, but of course has its own limitations. The method is a sophisticated trial-and-error approach, in which the sum of squares of the residuals (or some other, appropriate parameter, as illustrated in Section VII) is minimized by varying the adjustable parameters. For example, one can combine the method of steepest descent (which is most useful when the adjustable parameters are rather crude) with a Newton-Raphson iteration (which is most efficient when the parameters are close to their final values). The most popular non-linear algorithm is that proposed by Levenberg (1944) and implemented by Marquardt (1963). The Levenberg-Marquardt routine is widely available in software packages as well as in spreadsheets; in Excel it is called Solver.

Non-linear least squares can be used in those cases where linear least squares run into difficulties. For example, after the data shown in Figure 1b have been fitted to two separate lines, yielding two slopes and intercepts to be used as guess values, we can apply a non-linear least-squares routine to fit the data to two lines with the same slope. In this case

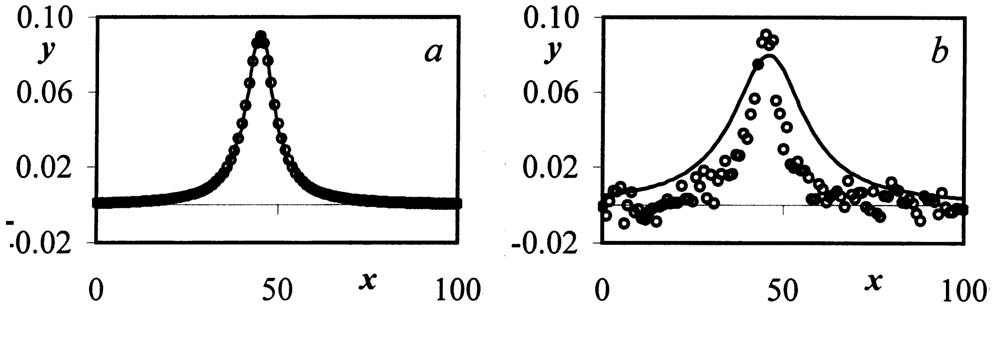


FIGURE 2. **(a):** A Lorentzian peak with noise (open circles), fitted with weighted least squares (solid line). **(b):** The same with ten times more noise, and the resulting, systematic distortion.

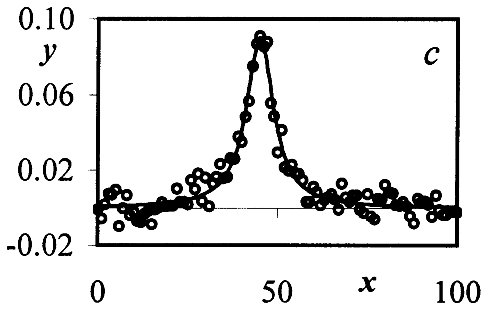


FIGURE 2C. Fitting the noisy Lorentzian of Figure 2b using non-linear least square.

the 'difficulty' we avoid is the need to derive equations such as 27 through 30.

To illustrate another type of difficulty that is readily taken care of by non-linear least squares, we can fit the data shown in Figure 2b directly (i.e., without the need for a transformation) to a Lorentzian, and thereby avoid the pernicious effect of noise. Figure 2c illustrates this.

On the other hand, the application of non-linear least squares methods has its own pitfalls. Because it looks for a minimum, it may get stuck in a local minimum, which is often caused by noise. Depending on the initial guesses made for the parameter values, the method may then lead to erroneous results. Direct observation of the signal and its fit may alert the user to the problem which, once recognized, can usually be solved simply by changing one or more of the guessed parameter values.

VII. GENERAL LEAST SQUARES

The most general least-squares method does away with the somewhat contrived distinction between dependent and independent parameters, and considers individual weights for all input parameters. Fortunately, there are relatively few instances where such general least-squares methods need be used, but even they can nowadays be handled deftly on generally available software such as spreadsheets. Here we will only consider the simplest case, in which $y = F(x_1, x_2, \ldots , x_i)$ and the parameters $x_1, x_2, \ldots , x_i$ are all *mutually independent*. If y and one or more of the parameters x_i are mutually dependent, we need information on their correlations. This requires more input information, which is usually not readily available.

For mutually independent parameters it can be shown, by extension of the arguments used by Deming (1943), that the relevant weight w is given by

$$w = \frac{1}{s_y^2 + \sum_{j=1}^{k} \left(\partial F/\partial x_j\right)^2 s_{x_j}^2} \tag{39}$$

where F is shorthand notation for $F(x_1, x_2, \ldots , x_\mathrm{j}, \ldots , x_k)$. The quantity to be minimized in the analysis of N data points then becomes

$$\sum_{i=1}^{N} w_i \left(y_i - y_{i,\,\mathrm{model}}\right)^2 = \sum_{i=1}^{N} \frac{\left(y_i - y_{i,\,\mathrm{model}}\right)^2}{s_{y_i}^2 + \sum_{j=1}^{k} \left(\partial F/\partial x_j\right)^2 s_{x_{j,i}}^2} \tag{40}$$

which can be used directly in a non-linear least squares algorithm such as Excel's Solver, where the user has full control over the minimization criterion. Again, like all the earlier results of this chapter, these equations apply only as long as the relative errors are small, so that local linearity is maintained.

For a meaningful comparison of the sum of the squares of the residuals of the weighted and unweighted least squares, it is useful to normalize the weights (such that $\Sigma_{i=1}^{N} w_i = 1$), in which case 39 must be modified to

$$w_i = \frac{\dfrac{1}{s_{y_i}^2 + \sum_{j=1}^{k} \left(\partial F/\partial x_j\right)^2 s_{x_{j,i}}^2}}{\sum_{i=1}^{N} \dfrac{1}{s_{y_i}^2 + \sum_{j=1}^{k} \left(\partial F/\partial x_j\right)^2 s_{x_{j,i}}^2}}$$

$$w_i = \frac{1}{\left\{ s_{y_i}^2 + \sum_{j=1}^{k} \left(\partial F / \partial x_j \right)^2 s_{x_{j,i}}^2 \right\} \sum_{i=1}^{N} \frac{1}{s_{y_i}^2 + \sum_{j=1}^{k} \left(\partial F / \partial x_j \right)^2 s_{x_{j,i}}^2}}$$

(41)

which, for a single *x*-parameter (i.e., $k = 1$), simplifies to

$$w_i = \frac{1}{\left\{ s_{y_i}^2 + \sum_{j=1}^{k} \left(dF / dx_j \right)^2 s_{x_{j,i}}^2 \right\} \sum_{i=1}^{N} \frac{1}{s_{y_i}^2 + \sum_{j=1}^{k} \left(dF / dx_j \right)^2 s_{x_{j,i}}^2}}$$

(42)

For the unweighted least squares we have $w_i = 1$, and we correspondingly divide the sum of the squares of the residuals by $\Sigma_{i=1}^{N} 1 = N$, where *N* is the number of data points analyzed.

Here we will illustrate the application of Eq. 42 by an example kindly provided by Prof. Jeffrey K. Nagle of this Department. Table 1 lists the rate constants *k* (corrected for diffusion effects) of the quenching, by a number of aromatic amines, of the fluorescence of the luminescent excited state of octachlorodirhennate (III), and the corresponding standard potentials E^o for the one-electron reduction of these quencers in acetonitrile containing 0.1 *M* tetrabutylammonium perchlorate, as reported by Nocera and Gray (1981). Both *k* and E^o are subject to experimental uncertainties, the magnitudes of which were estimated by Prof. Nagle. The data were then fitted, with and without weighting, to the Marcus expression (see, e.g., Marcus and Sutin 1985)

$$\ln k = \ln k_0 - \frac{\lambda F}{4RT}\left(1 + \frac{\Delta G^0 + E^0}{\lambda}\right)^2$$

(43)

where the values of ln k_0 (with k_0 in M^{-1} s^{-1}), the free energy ΔG^o (in V), and the solvent reorganization constant λ (also in V) are adjustable parameters, and $F/4RT = 9.73047$ V^{-1}. Figure 3 shows how such a computation can be made on a spreadsheet.

The specific values assigned to the experimental uncertainties can of course be questioned, but these are of no concern here, since they are used only to illustrate the computational method. The point is that we can get quite different answers depending on the assumptions made regarding the weights. Note that the 'unweighted' analysis also (though tacitly) assigns weights, namely, *unit* weights, $w_i = 1$.

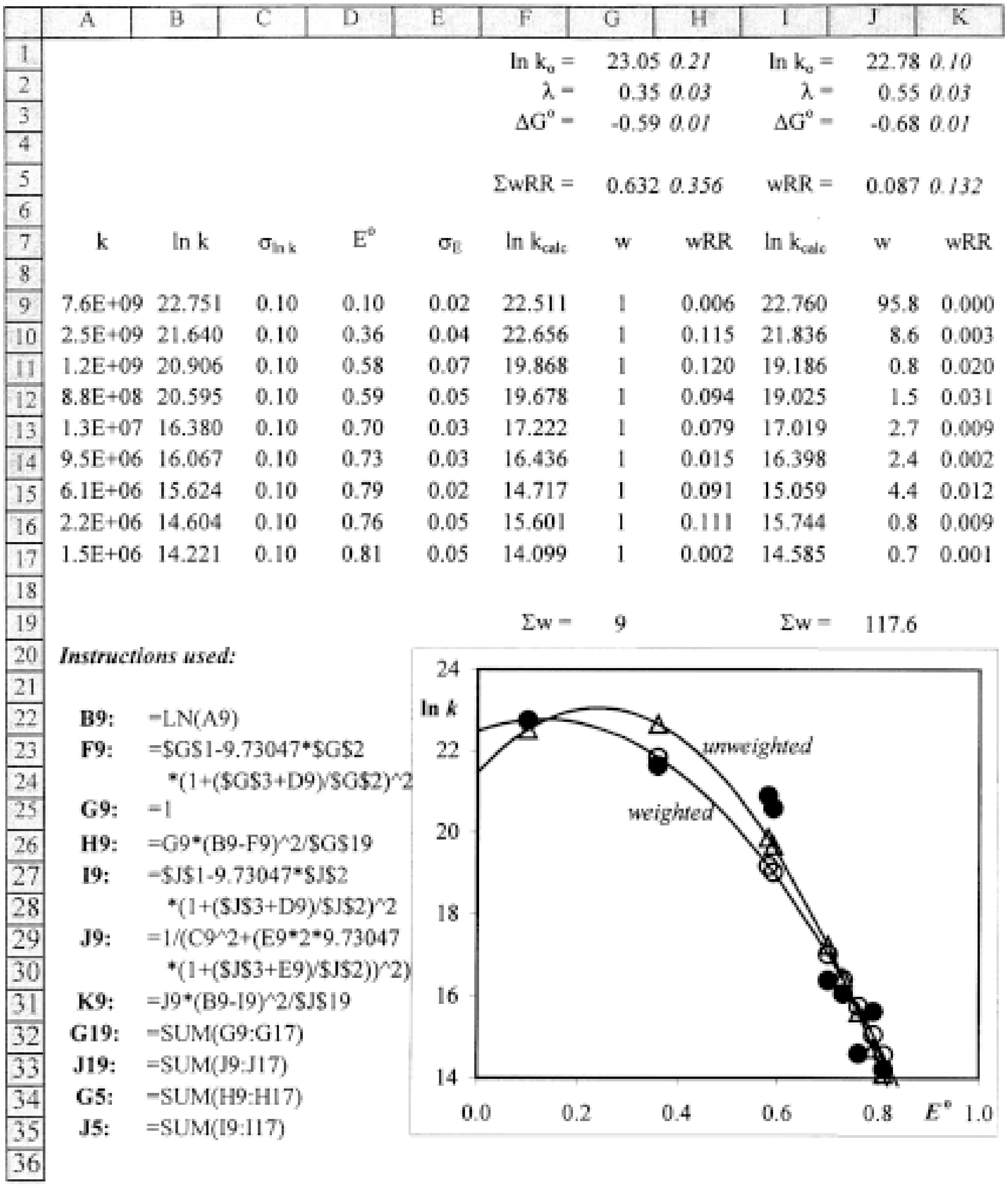
We see that, in this example, the fitting parameters can be significantly different: ln k_0 = 22.78 ± 0.10 vs. 23.05 ± 0.21, λ = 0.55 ± 0.03 V vs. 0.35 ± 0.03 V, and ΔG^o = –0.68 ± 0.01 V vs. –0.59 ± 0.01 V for the weighted and the unweighted analysis respectively. Which of these results one prefers will obviously depend on how much weight one attaches to the assigned weights.

Incidentally, these results clearly illustrate that the calculated standard deviations (e.g., ± 0.03 V for *l*, ± 0.01 V for ΔG^o) are much smaller than the differences between the parameter values in the two models (0.20 V for *l*, 0.09 V for ΔG^o), because the standard deviations indicate the precision *within a given model*, rather than the limits of reliability of the results. Even when we fit the data to the very same theoretical expression, weighting or not weighting constitute quite different models.

This example also shows the effect of the term *dF*/*dx* on the weights *w*. The first and seventh point in the above data have the same standard deviations s_y and s_x, yet the first point has a much larger weight *w* (in cell J9) than the seventh (in J15) because of the small slope *dF*/*dx* around *E* = 0.1 V.

VIII. CONCLUSIONS

The least-squares method provides a set of flexible tools to fit experimental data, even in the presence of noise. It thrives on data redundancy: the more data points, the smaller the influence of random fluctuations in the



	A	B	C	D	E	F	G	H	I	J	K
1						$\ln k_o$ =	23.05	*0.21*	$\ln k_o$ =	22.78	*0.10*
2						λ =	0.35	*0.03*	λ =	0.55	*0.03*
3						ΔG^o =	-0.59	*0.01*	ΔG^o =	-0.68	*0.01*
4											
5						ΣwRR =	0.632	*0.356*	wRR =	0.087	*0.132*
6											
7	k	ln k	$\sigma_{\ln k}$	E^o	σ_E	$\ln k_{calc}$	w	wRR	$\ln k_{calc}$	w	wRR
8											
9	7.6E+09	22.751	0.10	0.10	0.02	22.511	1	0.006	22.760	95.8	0.000
10	2.5E+09	21.640	0.10	0.36	0.04	22.656	1	0.115	21.836	8.6	0.003
11	1.2E+09	20.906	0.10	0.58	0.07	19.868	1	0.120	19.186	0.8	0.020
12	8.8E+08	20.595	0.10	0.59	0.05	19.678	1	0.094	19.025	1.5	0.031
13	1.3E+07	16.380	0.10	0.70	0.03	17.222	1	0.079	17.019	2.7	0.009
14	9.5E+06	16.067	0.10	0.73	0.03	16.436	1	0.015	16.398	2.4	0.002
15	6.1E+06	15.624	0.10	0.79	0.02	14.717	1	0.091	15.059	4.4	0.012
16	2.2E+06	14.604	0.10	0.76	0.05	15.601	1	0.111	15.744	0.8	0.009
17	1.5E+06	14.221	0.10	0.81	0.05	14.099	1	0.002	14.585	0.7	0.001
18											
19						Σw =	9		Σw =	117.6	

Instructions used:

```
20  Instructions used:
21
22    B9:   =LN(A9)
23    F9:   =$G$1-9.73047*$G$2
24            *(1+($G$3+D9)/$G$2)^2
25    G9:   =1
26    H9:   =G9*(B9-F9)^2/$G$19
27    I9:   =$J$1-9.73047*$J$2
28            *(1+($J$3+D9)/$J$2)^2
29    J9:   =1/(C9^2+(E9*2*9.73047
30            *(1+($J$3+E9)/$J$2))^2)
31    K9:   =J9*(B9-I9)^2/$J$19
32    G19:  =SUM(G9:G17)
33    J19:  =SUM(J9:J17)
34    G5:   =SUM(H9:H17)
35    J5:   =SUM(I9:I17)
36
```

FIGURE 3. A spreadsheet calculation, on Excel, for the unweighted and doubly weighted fit of the data of Nocera and Gray (1981). The experimental data are shown in columns A and D, and their estimated standard deviations in columns C and E. The equations used in row 9 are shown; those in rows 10 through 17 are obtained by copying these equations down. Solver is used to find the data in cells G1 through G3 for the unweighted case, and in J1 through J3 for weighting both axes. The uncertainty estimates shown in italics were obtained with a macro (de Levie, 1999): the deviations of log *k* in row 5. The experimental data are shown in the graph as solid points, the unweighted and weighted fits as lines with unfilled symbols.

data on the derived quantities. Convenient implementations of most least squares methods are readily available, so that it is seldom necessary to derive the corresponding equations (as we did in Section III.E) or program the computer to implement the analysis. Most linear least-squares methods are very efficient and fast. In principle, non-linear least squares are less so, because they involve iterations, and may need some manual help to guide them to the appropriate solution. Fortunately, on modern computers, even non-linear least squares methods are usually quite fast.

Most of the least-squares methods discussed in this brief review are based on the simplifying assumption that the uncertainties are fully restricted to the dependent variable. Fortunately, this is often a reasonable assumption, as discussed, e.g., by Taylor (1997). An example is provided in Section VII of how to handle cases for which that assumption might not be justified.

While the methods reviewed here are great to fit experimental data to the appropriate theoretical models, they should not be used unthinkingly. The scientific literature contains some classical illustrations (Parsons and Passeron 1966; Barclay et al., 1970) of the dangers of feeding experimental data directly into computers for analysis, without careful inspection of the input data and of the resulting data fit.

The least-squares method is designed to fit data to an assumed theoretical expression. It is *not* designed to select the best theoretical framework, and consequently is not very good at it either. Anscombe (1973) has given some dramatic examples illustrating the dangers of fitting data without visual inspection.

Least-squares methods are sometimes touted as objective, but they are so only *after* the equation to be used, and analysis method, have been selected. Whether to fit data to a proportionality or to a line, whether to include an additional term in a polynomial fit, what length to chose for a sliding polynomial, or whether the nature of the experimental uncertainties requires weighting, are all decisions that will affect the result obtained yet must still be made by the experimenter. Such decisions are typically based of information about the nature of the data that is not inherent in their numerical values, such as the existence of a theoretical expression for the observed phenomenon, or independently obtained knowledge about the most likely causes of the experimental uncertainties. The determination of the actual weights to be used in weighted least squares is often subjective as well.

While it has so far been customary to assume that errors follow a single, normal distribution, this assumption may occasionally have to be examined and, perhaps, modified. Now that the mechanics of least-squares methods are no longer a stumbling block, and the method is widely available, the focus can be expected to shift to a more critical examination of the sources and statistical nature of the errors involved.

ACKNOWLEDGMENTS

I am grateful for stimulating discussions with Prof. Jeff Nagle and Mr. Nate Harvey of this Department.

REFERENCES

Anscombe, F. J. *Am. Statist*. **1973**, *27*, 17.

Barclay, D. J., Passeron E., Anson, F. C. *Inorg. Chem*. **1970**, *9*, 1024.

Barak, P. W. *Anal. Chem*. **1995**, *67*, 2758.

Deming, W. E. *Statistical adjustment of data*, John Wiley New **1943**.

de Levie, R. *J. Chem. Educ*. **1986**, *63*, 10.

de Levie, R. *Principles of quantitative chemical analysis*, McGraw-Hill, **1997** pp. 652-657.

de Levie, R. *J. Chem. Educ*. **1999**, *76*, 1594.

de Levie, R. *How to use Excel in analytical chemistry and in general scientific data analysis*, Cambridge Univ. Press **2000**.

Levenberg, K. *Quart. Appl. Math*. **1944**, *2*, 164.

Madden, H. H. *Anal. Chem.* **1978**, *50*, 1383.

Marcus, R. A., Sutin, A. *Biophys. Biochem. Acta* **1985**, *811*, 265.

Marquardt, D. W. *J. Soc. Ind. Appl. Math* **1963**, *11*, 431.

Nocera, D. G., Gray, H. B. *J. Am. Chem. Soc.* **1981**, *103*, 7349.

Parsons, R., Passeron, E. *J. Electroanal. Chem*. **1966**, *12*, 524.

Savitzky, A., Golay, M. J. E. *Anal. Chem.* **1964**, *36*, 1627.

Sheppard, W. F. *Proc. 5th Congress of Math*., Cambridge (1912) II p. 348; *Proc. London Math. Soc*. **1914**, (*2*) *13*, 81, 97.

Sherriff, C. W. M. *Proc. Roy. Soc. Edinburgh* **1920**, *40*, 112.

Steinier, J., Termonia, Y., Deltour, J. *Anal. Chem*. **1972**, *44*, 1906.

Taylor, J. *An introduction to error analysis: the study of uncertainty in physical measurements*, 2nd ed., University Science Books **1997**, pp. 188–190.