JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

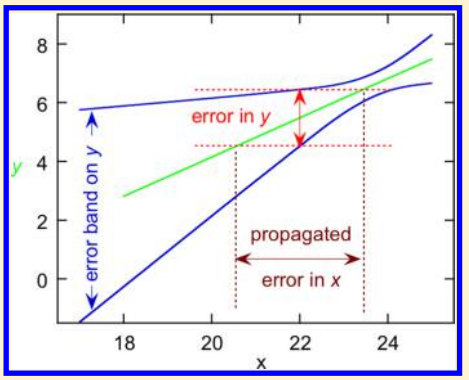
Using Least Squares for Error Propagation

Joel Tellinghuisen*

Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States

Supporting Information

ABSTRACT: The method of least-squares (LS) has a built-in procedure for estimating the standard errors (SEs) of the adjustable parameters in the fit model: They are the square roots of the diagonal elements of the covariance matrix. This means that one can use least-squares to obtain numerical values of propagated errors by defining the target quantities as adjustable parameters in an appropriate LS fit model. Often this will be an exact, weighted, nonlinear fit, requiring special precautions to circumvent program idiosyncrasies and extract the desired a priori SEs. These procedures are reviewed for several commercial programs and illustrated specifically for the KaleidaGraph program. Examples include the estimation of $\Delta H°$, $\Delta S°$, $\Delta G°$, and $K°(T)$ and their SEs from $K°$ (equilibrium constant) values at two temperatures, with and without uncertainty in T, which is included using the effective variance method, a general-purpose LS procedure for including uncertainty in independent variables. In some cases, the target quantities can be obtained from the original data analysis, by redefining the fit model to include the quantity of interest as an adjustable parameter, automatically handling correlation problems. Examples include the uncertainty in the fit function itself, line areas from spectral line profile data, and the analysis of spectrophotometric data for complex formation.



KEYWORDS: *Upper-Division Undergraduate, Graduate Education/Research, Physical Chemistry, Laboratory Instruction, Analytical Chemistry, Problem Solving/Decision Making, Calibration, Chemometrics, Thermodyamics*

Although chemistry students encounter some elements of error propagation as early as their general chemistry courses (e.g., significant figures rules), their most serious encounter with this topic is usually in the physical chemistry teaching laboratory, where they learn to use the equation

$$\sigma_f^2 = \sum \left(\frac{\partial f}{\partial \beta_i} \right)^2 \sigma_i^2 \qquad (1)$$

to compute the statistical error σ_f in a function $f(\beta)$ of the independent variables β_1, β_2, ... having statistical errors σ_1, σ_2, ... These applications include especially the analysis of the data they collect in the laboratory. There, an understanding of eq 1 should lead them to appreciate the relative importance of the several measured quantities in an experiment. For example, in bomb calorimetry, multiplication and division to the first power are the only significant mathematical operations involved in analyzing the data. Because pellet and calorimeter water masses can be measured easily to 0.1%, whereas the temperature rise ΔT is uncertain by 1%, students should recognize that ΔT will be precision-limiting and, hence, deserving of more care in the experiment.

Realizing that using eq 1 can be tedious and error-prone, several contributors to this Journal have described mathematical procedures for obtaining numerical estimates of σ_f for any defined function f.[1−6] They accomplish this by estimating the required derivatives numerically in provided algorithms. I would argue that mastery of eq 1 is of value in itself, for reasons like those stated in the bomb calorimetry example. Surely, students should learn the simpler forms this expression takes when just addition and subtraction or just multiplication and division are involved, especially because these cases lead to the separate rules for significant figures in these two cases.[7] Also, they should recognize that in single variable relations, $y = f(x)$, a 1% uncertainty in x leads to 1/2%, 1%, and 2% uncertainty in y when the functional dependence on x is to the powers $\pm 1/2$, ± 1, and ± 2, respectively; and to an uncertainty of 0.01 in $\ln(x)$. Still, there are cases more complex than these where a numerical result can be at least reassuring if not essential. To that end, I discuss below an approach that I believe has been only touched on here before:[5] using least-squares (LS).

A key virtue of the method of least-squares is that it can provide estimates of not just the adjustable parameters but also their statistical precisions. When the data error is known, these parameter standard errors (SEs) are not just estimates, they are *exact* in the case of linear LS (LLS) and exact in the limit of small data error for nonlinear LS (NSL).[8,9] In error propagation, we start with assumed known SEs for the independent variables. Thus, using LS to do the propagation requires defining the quantity f of eq 1 as an adjustable parameter in a weighted LS fit, with the uncertain independent variables β being variables in this fit, having weights equal to their reciprocal squared SEs. Usually, this fit will also be a nonlinear one. Many LS programs can handle weighted nonlinear fitting. Here, I use the KaleidaGraph program (Synergy Software) for my illustrations,[10] but programs like

DOI: 10.1021/ed500888r
J. Chem. Educ. XXXX, XXX, XXX−XXX

Origin (OriginLab Corp.), IGOR Pro (WaveMetrics), and R (r-project.org) can be used as well.

There is one nuisance complication in using LS for error propagation: Many applications involve *exact* fits, where the number of adjustable parameters is equal to the number of data points. This problem is practical rather than fundamental, as NLS algorithms can solve systems of n equations in n unknowns. But packaged routines may contain tests to determine if the number of input points is sufficient. In the case of KaleidaGraph, that test is simple: There must be at least three points. Thus, KG functions correctly in solving systems of equations having $n > 2$ (and in fact also provides solutions for underdetermined systems, like fitting a four-parameter function to three points, in which case the output depends unpredictably on the trial input parameter values). There are easy ways to get around this problem. One is just to include extra values of the points, greatly downweighted to ensure that they have no effect on the numerical outcome. Another is to include multiple values with altered SEs, for example replacing one value having SE = σ with four identical values having SE = 2σ (hence being statistically equivalent).[8] These tactics are illustrated below.

A second potentially tricky aspect to exact LS fitting is more fundamental: The sum of weighted, squared residuals (δ_i = observed − calculated y_i values)

$$S = \sum w_i \delta_i^2 = \sum (\delta_i/\sigma_i)^2 \qquad (2)$$

goes identically to zero. S is the LS minimization target ("least squares"), and it equals zero anytime all "observed" values equal their "calculated" counterparts, which must occur when the number of points equals the number of adjustable parameters (giving zero "degrees of freedom" ν). In ordinary unweighted LS, S/ν becomes an estimate of the data variance; and the estimates of the parameter SEs (called a posteriori) contain a factor of $(S/\nu)^{1/2}$, which means they must vanish (or be indeterminate) for exact fits. However, in weighted LS, when the data error is assumed to be known, the appropriate SEs are the a priori values,[8] and these do not contain this factor. This a priori mode is the default for weighted LS in KG,[10] so KG can be used directly for error propagation. The problem of invoking this mode in other programs is addressed below.

■ COMPUTATIONAL CONSIDERATIONS

In KaleidaGraph, both nonlinear and weighted fitting require the General routine.[10] The user enters x_i, y_i, and σ_i in separate columns of a given row in the data sheet. The data are displayed in an x−y plot and General is selected under the Curve Fit menu. The fit function is entered in a Define Fit box, which includes a "weight data" option. When this is checked, the user is prompted for the column containing the σ_i values; and the weights are taken as $1/\sigma_i^2$ in the computations, which use the Marquardt algorithm.[11,12] The output includes the "Value" and "Error" for each parameter, the latter being the a priori SE, which here is the desired propagated uncertainty.

Other programs handle the choice between a priori and a posteriori SEs differently. For example, the Origin program offers several ways to define weights; and in execution, it permits the user to check a box, "Scale errors by square root of reduced χ^2." (This is my S/ν, and the scaling produces the a posteriori SEs, so this box should be left unchecked in the present application.) In the FORTRAN program CURFIT provided long ago by Bevington,[12] the parameter SEs and the covariance matrix (**V**) are the desired a priori forms. This is true also for the similar subroutine MRQMIN of Press et al.[13] In R, for both linear ("lm") and nonlinear LS ("nls"), there is an output called "cov.unscaled" which is the desired a priori **V**, from which the SEs are the square roots of the diagonal elements.

In Excel, NLS is done with the Solver routine, which does not provide parameter SEs. To remedy this deficiency, de Levie provided a routine called SolverAid.[14] Present versions of this program do not accommodate weights, but the program listing is freely available,[15] so users can modify it. For general NLS applications, the minimization target for Solver will also have to be changed to the sum of weighted, squared residuals defined in eq 2. But SolverAid alone, with the weights modification, will suffice for error propagation.[16] In the current version, the definition statement for **V** (VarCovarValue) includes a factor of "Sy * Sy", which converts the desired a priori SDDInv(i, j) to the a posteriori version.

The conversion from a priori to a posteriori **V** just noted for SolverAid serves to emphasize that the latter requires an extra step beyond the former, namely the estimation of σ_y. Accordingly, users of programs like MathCad and Mathematica, who normally work "closer" to the fundamental mathematics, will compute the a priori version in any evaluation of **V**.

Some error propagation problems can be cast as LS fits of the most familiar two-dimensional $y = f(x)$ form, with only y being uncertain. However, many require allowance for uncertainty in x, or even for more than two variables. An example of the first type is the estimation of thermodynamic quantities from the T dependence of equilibrium constants, discussed in the next section. Examples requiring multiple uncertain variables include this one with inclusion of uncertainty in T, and the example in ref 2 of determining gas heat capacities from three pressure measurements, which becomes a 1-point NLS problem with three uncertain variables. Such problems are best handled with so-called *total variance* methods (also called "errors in variables" and sometimes "generalized"), which were proposed as early as 1938 by Deming,[17] and have been available in computer algorithmic form since 1972.[18,19] KaleidaGraph and similar data analysis programs cannot directly implement such methods, as they are designed to handle a single uncertain variable. However, there is a variation of total variance, called the *effective variance method*, in which the uncertainty in the "independent" variables is converted into an effective contribution to the variance in the single dependent one. This is essentially the implementation of Deming's method described long ago in this Journal by Wentworth.[20] Using Monte Carlo simulations, I have shown that there is insignificant practical difference in the two methods in a number of common fitting problems.[21] For the present application, where the fits are exact, there is no difference at all, so I will illustrate how to use effective variance to handle independent variable uncertainty in several examples. This may require the use of eq 1 to obtain expressions for the contributions to the effective variance, but it can also be done using the "functional approach",[6] keeping the treatment entirely "numerical".

■ ILLUSTRATIONS

$\Delta G°$, $\Delta H°$, and $\Delta S°$ from Temperature Dependence of $K°$

Suppose that $K°$ has been estimated at two Ts and we wish to obtain values for $\Delta G°$, $\Delta H°$, and $\Delta S°$. If values are available at only two Ts (as is often the case in the teaching laboratory), we

DOI: 10.1021/ed500888r
J. Chem. Educ. XXXX, XXX, XXX−XXX

must assume that $\Delta H°$, and hence $\Delta S°$, are independent of T. At $T = T_1$, we have

$$\Delta G° = -RT_1 \ln K_1° \tag{3}$$

whence

$$\sigma_{\Delta G°} = \frac{RT_1 \sigma_{K_1°}}{K_1°} \tag{4}$$

and similarly at $T = T_2$. A weighted LS fit of these two values to

$$\Delta G° = \Delta H° - T\Delta S° \tag{5}$$

will thus yield $\Delta H°$, $\Delta S°$, and their SEs. When (as here) the $\Delta G°$ values are obtained from equilibrium constants at multiple Ts, a useful variation of eq 5 is

$$\ln K° = -\frac{\Delta H°}{RT} + \frac{\Delta S°}{R} \tag{6}$$

If the $K°$ values are obtained with constant percent error (as is often the case[22]), we have $\sigma(\ln K°) = \sigma_{K°}/K° = $ constant; however, a weighted fit is still required to obtain the desired a priori SEs.

A classic teaching laboratory experiment for studying temperature dependence of equilibrium is the complexation of I_2 with mesitylene, as reported in a landmark paper by Benesi and Hildebrand[23] and appearing in the teaching literature as early as 1962.[24] This experiment has been revisited recently in this Journal by Baum et al.,[25] who show in their Figure 5 an analysis of 3 $K°$ values using eq 6. From their data I estimate that $\sigma(\ln K°) \approx 0.025$. To illustrate the pure error propagation capabilities of LS, I adopt this value and conduct a weighted fit of just the two extreme-T $K°$ values. Results are shown in Figure 1 (upper fit results box), where each value is quadrupled and given a σ value double its true value.

Identical results are obtained fitting to the exponential version of eq 6, for which we must now supply σ values for $K°$ for the weighted fit. From eq 1, these are 0.025 times the respective $K°$ values. Alternatively, we can fit $\Delta G°$ values using eq 5, now requiring $\sigma(\Delta G°)$, from eq 4 for the weighting. Results for both approaches are shown in Figure 2, obtained this time by adding a third downweighted value. The data sheets for both figures are shown in Figure 3.

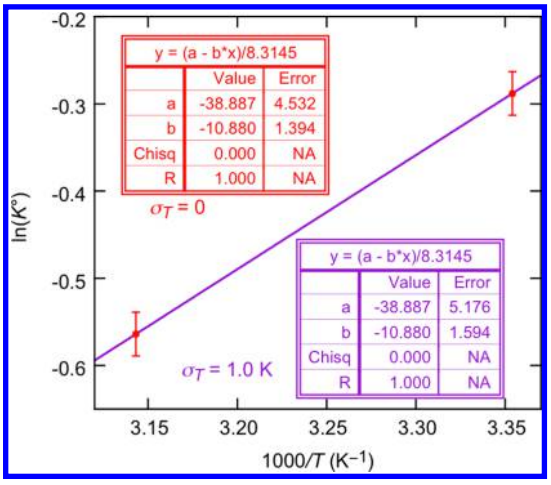


Figure 1. KaleidaGraph least-squares results for $\Delta S°$ (a, units of J mol^{-1} K^{-1}), $\Delta H°$ (b, units of kJ mol^{-1}), and their standard errors from eq 6 analysis of equilibrium constant $K°$ at 25 and 45 °C, for T taken as error-free (upper results box) and having uncertainty σ_T = 1.0 K (lower). Four ln $K°$ values are included at each T, having uncertainty $\sigma(\ln K°) = 0.050$ (2 × true) and values −0.5640 and −0.2881.

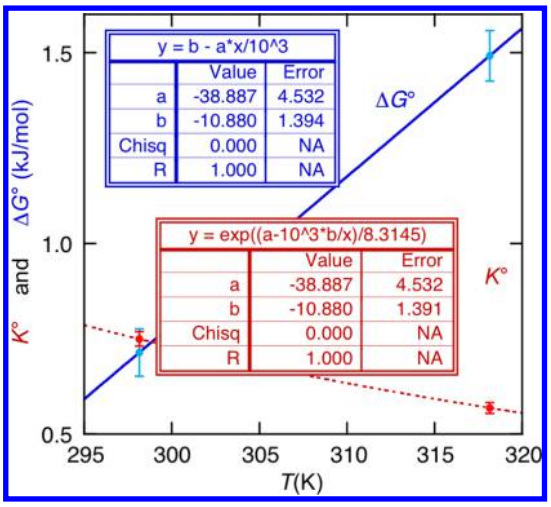


Figure 2. Alternative analyses of same data for σ_T = 0 using eq 5 (upper results) and the exponential version of eq 6 (lower, dashed curve).

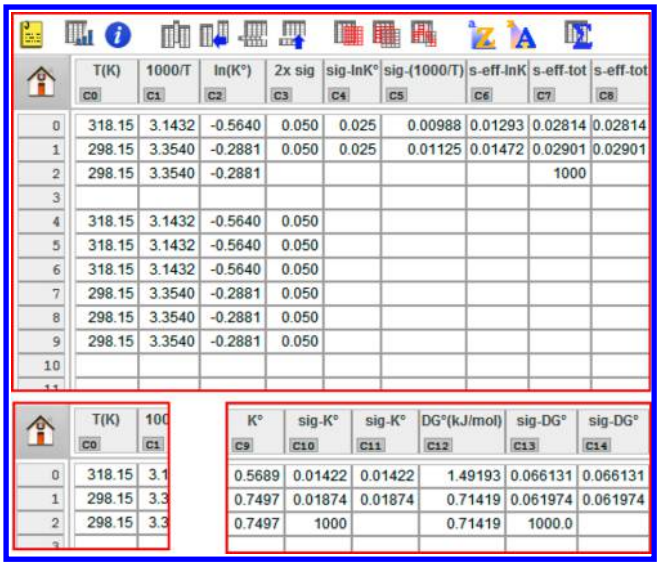


Figure 3. KaleidaGraph data sheet for Figures 1 (top) and 2 (bottom). The fit for σ_T = 0 in Figure 1 is done by masking out the third row (numbered 2) and fitting the 4 values at each T with doubled σ values (0.05). The fit for σ_T = 1.0 K is done by masking out the 6 rows numbered 4–9 and selecting the σ values in column C7 (s-eff-tot) for weighting. The displayed error bars are selected from column C4. Both fits in Figure 2 use T (C0) as independent variable and $K°$ and the adjacent σ column for one fit, and $\Delta G°$ and its adjacent σ column for the other. Columns C11 and C14 are for error bar display.

To include uncertainty in T by the effective variance treatment, we must use eq 6 or its exponential version because $\Delta G°$ is itself subject to uncertainty in T, and the effects of this and the explicit uncertainty in T are subtly correlated. Suppose the uncertainty in T is σ_T = 1.0 K. The relative error in $x \equiv 1000/T$ is the same as that in T, giving $\sigma_x = (1.0/T)x$ (column C5 in Figure 3, obtained using C5 = 1./C0*C1 in the Formula Entry window). Applying eq 1 to eq 6, the T contribution to the variance in ln$K°$ is the square of s-eff-ln $K = |\Delta H°| \times \sigma_x/R$ (C6 of Figure 3). The total effective variance in ln $K°$ is the

DOI: 10.1021/ed500888r
J. Chem. Educ. XXXX, XXX, XXX−XXX

square of s-eff-tot [C7 = sqrt (C6^2 + 0.025^2) in FE window]. This weighting yields a 14% increase in the SEs, shown in the lower fit box in Figure 1.

It is instructive to check the results of this LS approach against those obtained using eq 1. To do so, we must first express ΔH° and ΔS° in terms of K° and T. For the former, we have the van't Hoff relation

$$\ln\left(\frac{K_2^\circ}{K_1^\circ}\right) = \frac{\Delta H^\circ}{R}\left(\frac{1}{T_1} - \frac{1}{T_2}\right) \tag{7}$$

and for the latter, we obtain

$$\Delta S^\circ = R(T_2 - T_1)^{-1}[T_2\ln(K_2^\circ) - T_1\ln(K_1^\circ)] \tag{8}$$

If we neglect uncertainty in T, the error in ΔH° is proportional to that in $\ln(K_2^\circ/K_1^\circ)$, which is $0.025\sqrt{2}$. Scaling by R and the reciprocal T term yields 1.3942 kJ/mol, in agreement with results in Figure 1 for $\sigma_T = 0$. For ΔS°, the rule for addition and subtraction gives for the error in the term in brackets $0.025\,[T_1^2 + T_2^2]^{1/2}$. Scaling by the other factors in eq 8 yields agreement with the results in Figure 1.

With allowance for T uncertainty, we treat ΔH° in eq 7 by applying the results for division (squared relative error = sum of squared relative errors of numerator and denominator) after first computing the σ_s for the ln K° ratio (already done) and the difference in the reciprocal Ts. The result again agrees with the relevant one in Figure 1. Applying eq 1 to eq 8 yields four contributing terms — for the two Ts and the two Ks. The key partial derivatives are, for example

$$\frac{\partial \Delta S^\circ}{\partial T_2} = \frac{R\ln K_2^\circ - \Delta S^\circ}{T_2 - T_1} \tag{9}$$

and similarly for $\partial/\partial T_1$, but with the sign reversed. The terms for K° are as before, and the complete result again agrees with that in Figure 1, confirming the validity of the effective variance treatment. The latter, though not fully numerical, still involves a simpler use of eq 1.

The 1.0 K temperature uncertainty adds only nominally to the total errors. However, if σ_T were doubled or if the ΔT interval were halved, the effects of T uncertainty would exceed those from K uncertainty. This is because the individual variance contributions from T uncertainty are proportional to the squares of σ_T and $1/(T_2-T_1)$.

Equation 7 can be used for a different LS fit of the points at the two Ts

$$\ln(K^\circ) = \ln(K_0^\circ) + \frac{\Delta H^\circ}{R}\left(\frac{1}{T_0} - \frac{1}{T}\right) \tag{10}$$

where T_0 is specified and ΔH° and $\ln(K_0^\circ)$ (or K_0° itself) become the adjustable parameters. This fit returns the same value and SEs as before for ΔH°, for both $\sigma_T = 0$ and $\sigma_T = 1$ K. For ln K°, the error when $\sigma_T = 0$ and T_0 is set to either T_1 or T_2 is 0.025; it drops to $0.025/\sqrt{2}$ midway between these two Ts. With $\sigma_T = 1.0$ K, there is an asymmetric increase in $\sigma(K^\circ)$, from about 12% at low x (high T) to 16% at high x, averaging the previously noted 14%. Equation 10 is mathematically equivalent to the Arrhenius relation for the dependence of reaction rate constants on T, so this treatment can be used to assess the activation energy and k_T and their uncertainties as functions of T.[5]

Having obtained ΔG° and ΔH° and their SEs, one might be tempted to use eq 5 for a simpler estimation of ΔS° and its uncertainty. Although that approach is incorrect,[26] it can yield results that are close to correct. That is because $\sigma(\Delta G^\circ)$ from K° is often so much smaller than $\sigma(\Delta H^\circ)$ that $\sigma(T\Delta S^\circ) \approx \sigma(\Delta H^\circ)$.[22] In the present case, estimates of $\sigma(\Delta S^\circ)$ obtained this way exceed the true values by only 0.2% for both $\sigma_T = 0$ and $\sigma_T = 1$ K.

If data for K° are available for more than two temperatures, the problem becomes a straightforward application of weighted NLS, as noted earlier. However, one must generally allow for T-dependence in ΔH°. To do this, we start with the differential form of the van't Hoff relation

$$\left(\frac{\partial \ln K^\circ}{\partial T}\right)_P = \frac{\Delta H^\circ}{RT^2} \tag{11}$$

and integrate for an assumed T-dependent ΔH°. This case has been discussed for ΔH° that is quadratic in T (permitting ΔC_P° to be T dependent), with procedures for using KaleidaGraph to obtain $\Delta G^\circ(T)$ (dependent upon $K^\circ(T)$ alone), ΔH°, ΔS°, ΔC_P°, and their errors as functions of T.[22] Uncertainty in T was not considered but can be included through the effective variance approach.[27]

More than Two Variables

Many error propagation problems are not suitable for two-dimensional LS, because they involve more than two uncertain variables. For example, in Figure 4 of ref 4, the authors use an Excel spreadsheet to illustrate the calculation for a function of five uncertain variables, $f(y,z,u,v,w) = y + z^2 + uvw$. One *can* use KG on this problem, through its cell command, by which variables are accessed by their column numbers. One variable must first be selected as "dependent". For example, taking y to be such, the fit function can be expressed in the Define Fit box as, for example

$$\text{a} - \text{cell(x, 3)}\text{\^{}}2 - \text{cell(x, 4)*cell(x, 5)*cell(x, 6)} \tag{12}$$

in which a is the adjustable parameter equivalent to f, and z, u, v, w are placed in columns 3−6, respectively. The "independent" variable x is the row number, and again one must add at least two rows of values to satisfy the KG sufficiency test. In this case, the effective variance approach requires as much formal effort with eq 1 as just using the latter directly. Thus, there is little point in using LS here, especially because the application of eq 1 is easy. A problem similar to this but more demanding in application of eq 1 is the determination of the gaseous heat capacity C_V from three pressure measurements, discussed by Donato and Metz.[2] The KG treatment of this problem is illustrated in Supporting Information, including use of the method of ref 6 to get the contributions to the effective variance. Results from a FORTRAN program like those from refs 18 and 19 are also included for comparison.

A more appropriate target for the LS approach is the absorbance problem also discussed by Gardenier et al.[4] Here, the concentrations of two components having overlapping absorption spectra are determined from absorbance values A at two wavelengths, in accord with

$$A_i = \varepsilon_{i1}c_1 + \varepsilon_{i2}c_2 \tag{13}$$

where ε_{ij} is the molar absorptivity of component j at wavelength i. To use eq 1, we must first solve for c_1 and c_2 in terms of the two A and four ε values, all uncertain, making this a six-variable error propagation computation. The solution and error propagation are handled automatically in the LS approach,

DOI: 10.1021/ed500888r
J. Chem. Educ. XXXX, XXX, XXX−XXX

and the effective variance calculations are easy. The KG fit function might be defined as

$$\text{a*cell(x, 1)} + \text{b*cell(x, 3)} \tag{14}$$

with A taken as the dependent variable and the independent variable x again being the row number in the data sheet. The concentrations are the adjustable parameters, a and b, and C1 and C3 contain the ε values. The effective variance from each ε is of form $(\sigma_{\varepsilon_j} c_j)^2$, so the total for each A is the sum of two such terms and σ_A^2. The treatment is readily extended to more than two wavelengths (and to more than two components), with no additional labor in assessing the effective variances because they can be obtained for all wavelengths at once using column arithmetic operations in the Formula Entry window.[10] A numerical illustration of this example is included in the Supporting Information.

Designing LS Fits to Output Derived Quantities and their SEs

We have already seen how one can obtain additional information by redesigning the fit, for example, $K°(T_0)$ instead of $\Delta S°$ from $K°(T)$ in eq 10. I consider other cases here. Note that if we want to evaluate $K°$ as a function of T from the original analysis in terms of $\Delta H°$ and $\Delta S°$, we will need to take correlation into account because parameters from an LS analysis are generally correlated.[5,8,17,28−31] In matrix form, the desired expression for the first-order propagated error is

$$\sigma_f^2 = \mathbf{g}^T\mathbf{V}\mathbf{g} \tag{15}$$

where $\mathbf{g}$ is a column vector containing the partial derivatives of the function f with respect to the adjustable parameters, evaluated at the values of the variables and parameters of interest. Many available packages for numerical error propagation cannot handle correlation, or if they can,[5,32] anyway require $\mathbf{V}$ for input, which in turn requires calculations equivalent to the LS fit. The methods I discuss here yield the correct results directly from the fit, without any need for eq 15.

An important case is where f is the fit function itself, for which we may want σ_f as a function of the independent variable. For example, in classical univariate calibration, the uncertainty in the unknown x_0 can be calculated for any calibration fit function, using[33,34]

$$\sigma_{x_0}^2 = \frac{\sigma_f^2(x_0) + \sigma_{y_0}^2}{(df/dx)_0^2} \tag{16}$$

where σ_{y0} is the uncertainty in the measured response for the unknown. Consider the polynomial in x: $f(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)^2 + ...$ Using eq 15, one can show that at $x = x_0$, $f(x) = a_0$, and $\sigma_f = \sigma_0$. Furthermore, although the numerical values of all but the highest-order coefficient change with x_0, the output fit *function* is independent of x_0. Thus, by repeating the fit for various x_0, we may generate $f(x_0)$ and its σ_f as functions of x_0. Further, the derivatives and their SEs are also obtained directly this way; that is, a_1 is the first derivative at x_0, and so forth. A variation on this is the function $g(x) = b_0 + b_1(x-x_0) + b_2(x^2-x_0^2) + ...$ This approach is more widely applicable than the polynomial in $(x - x_0)$—for example,[8] to the van Deempter equation, which contains a term in x^{-1}—but the derivatives are not as easily obtained when those are of interest.

Figure 4 shows both of these functions used at the cubic level to fit thermistor calibration data, for x_0 = 25 °C. Note that the fit functions are identical, as are the a and d values and their SEs, and Chisq. Both functions remain unchanged as x_0 is altered (shown only for the polynomial centered at 25° and 30°), as do d and its Error; but all other quantities change. For any x_0, a and its Error are the values of the fit function and its uncertainty.

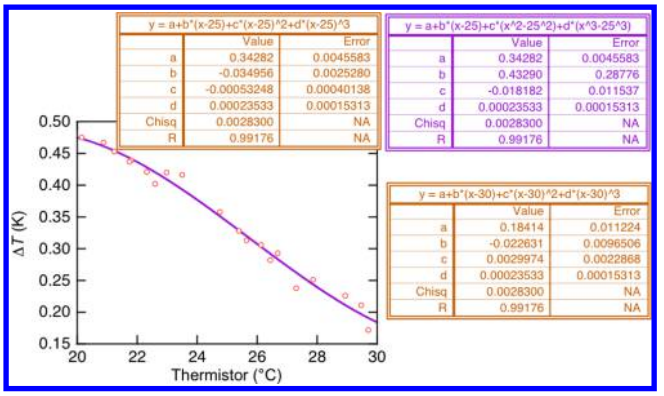


Figure 4. Cubic LS fits (unweighted) of thermistor calibration data (true−thermistor) using a polynomial in $(t - x_0)$ (upper left results for 25 °C, lower right for 30°) and the alternative form discussed in text ($g(x)$, upper right).

Among other derived properties of interest are line areas from spectrometric techniques that yield line profiles by scanning a controlled variable. If the profiles can be represented by a functional form (e.g., Gaussian), it can be useful to fit the data to such form and then compute the area under each line using the fit parameters. Intuitively, most analysts would fit a line profile using the parameters that govern its position, amplitude, and width. The area is proportional to amplitude × width, but eq 1 does not give the correct uncertainty in the area because the amplitude and width are correlated. An alternative is to fit directly to the area and obtain the correct SE directly from the fit. This case is illustrated in the Supporting Information, where the correct SE for the area is 35% smaller than computed using eq 1.

As a final example, consider the analysis of spectrophotometric data for complexation, as used in ref 25 to obtain the $K°(T)$ data I used in my first example. The reaction under study is

$$\text{M} + \text{I}_2 \leftrightarrow \text{M} \bullet \text{I}_2 \tag{17}$$

From the earliest study of this process,[23] it has been customary to take $[\text{M}]_0 \gg [\text{I}_2]_0$ and then to analyze the data using

$$\frac{[I_2]_0 L}{A_x} = \frac{1}{\varepsilon_x K[\text{M}]_0} + \frac{1}{\varepsilon_x} \tag{18}$$

where subscript 0 indicates the initial concentration in the reaction mix, A_x and ε_x are the measured absorbance and the molar absorptivity, respectively, of the complex at a suitable wavelength, and L is the cuvette path length. If these data are fitted to a straight line, $y = bx + a$, with $x = 1/[\text{M}]_0$, then $\varepsilon_x = 1/a$ and $K = a/b$. Equation 1 will correctly determine the SE for ε_x, but not for K, because it is a function of both LS parameters and they are correlated. However, by redefining the fit relation as $y = x/(a\ b) + 1/a$, we have a nonlinear LS fit to a straight line that yields $\varepsilon_x(a)$ and $K(b)$ and their SEs directly, with correlation automatically included.

DOI: 10.1021/ed500888r
J. Chem. Educ. XXXX, XXX, XXX−XXX

CONCLUSION

The method of least-squares can be used to obtain numerical values for the propagated error in any desired function of the quantities that are presumed to have known uncertainty. If the latter are themselves the results of a least-squares analysis, this fit may be modified to include the desired quantity as an adjustable parameter, in which case correlation will be taken into account automatically. In other situations, a new LS fit can be designed to represent the target quantities as parameters, with the known quantities taken as uncertain variables. This fit will usually be exact and will require weighted, nonlinear LS to obtain the a priori parameter SEs.

A numerical value for the propagated error unfortunately may not translate easily into confidence limits. First, we have completely ignored uncertainty in the "known" uncertainties; if these have been obtained from measurements, they are subject to the properties of χ^2, which means they have a relative uncertainty of $(2\nu)^{-1/2}$. Second, many quantities of interest are nonlinear functions of the knowns, which means that they will not be normally distributed even if the knowns have Gaussian error. In fact, many nonlinear estimators do not even have finite variance, so the propagated SE is at best a sort of asymptotic approximation. A simple example of this is the estimation of a reciprocal, $y = A \equiv 1/a$: If y has normal error, a has infinite variance, and when σ_A is as large as $\sim |A|/3$, this will manifest as a clear failure of the central limit theorem, meaning a and its error cannot be estimated by sampling.[35] (Of course A is well-behaved.) Confidence limits do remain defined in such situations, and they can be estimated through Monte Carlo simulations.[4,31] A hallmark of nonlinear estimators is asymmetric distributions, and a guide to when these will be problematic is a 10% rule of thumb:[9] If $\sigma_a/|a| < 0.1$, the true $\pm$ distances from the mean for 68% probability will be within 10% of σ_a. Finally, all error propagation here has been first-order. There are programs available for obtaining SDs to higher order.[32] However, when such corrections are significant, it seems likely that the asymmetry problem will anyway require Monte Carlo simulations for precise determination of confidence limits.

ASSOCIATED CONTENT

Supporting Information

Three examples: gas heat capacity from three pressure measurements,[2] determination of two component concentrations from optical absorption at two wavelengths,[4] and direct estimation of spectral line areas from line profile data. This material is available via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: joel.tellinghuisen@vanderbilt.edu.

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

I thank Andrej-Nikolai Spiess for information about working in the R programming environment and Robert de Levie and Carl Salter for helpful correspondence.

REFERENCES

(1) Ellenberger, M. R.; Nelson, R. D., Jr. Numerical error propagation with computer assistance. *J. Chem. Educ.* **1972**, *49*, 678.
(2) Donato, H., Jr.; Metz, C. A. A direct method for the propagation of error using a personal computer spreadsheet program. *J. Chem. Educ.* **1988**, *65*, 867−868.
(3) de Levie, R. On the propagation of statistical errors for a function of several variables. *J. Chem. Educ.* **2000**, *77*, 534−535.
(4) Gardenier, G. H.; Gui, F.; Demas, J. N. Error Propagation Made Easy—Or at Least Easier. *J. Chem. Educ.* **2011**, *88*, 916−920.
(5) de Levie, R. Collinearity in Least-Squares Analysis. *J. Chem. Educ.* **2012**, *89*, 68−78.
(6) Hughes, I. G.; Hase, T. P. A. Error Propagation: A Functional Approach. *J. Chem. Educ.* **2012**, *89*, 821−822.
(7) Andraos, J. On the propagation of statistical errors for a function of several variables. *J. Chem. Educ.* **1996**, *73*, 150−154.
(8) Tellinghuisen, J. Understanding Least Squares Through Monte Carlo Calculations. *J. Chem. Educ.* **2005**, *82*, 157−166.
(9) Tellinghuisen, J. A Monte Carlo Study of Precision, Bias, Inconsistency, and Non-Gaussian Distributions in Nonlinear Least Squares. *J. Phys. Chem. A* **2000**, *104*, 2834−2844.
(10) Tellinghuisen, J. Nonlinear Least-Squares Using Microcomputer Data Analysis Programs: KaleidaGraph in the Physical Chemistry Teaching Laboratory. *J. Chem. Educ.* **2000**, *77*, 1233−1239.
(11) Marquardt, D. W. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **1963**, *11*, 431−441.
(12) Bevington, P. R.; Robinson, D. K *Data Reduction and Error Analysis for the Physical Sciences*, 2nd ed.; McGraw-Hill: New York, 1992.
(13) Press, W. H.; Flannery, B. P.; Teukolsky, S. A.; Vetterling, W. T. *Numerical Recipes*; Cambridge Univ. Press: Cambridge, U. K., 1986.
(14) de Levie, R. Estimating parameter precision in nonlinear least squares with Excel's solver. *J. Chem. Educ.* **1999**, *76*, 1594−1598.
(15) Excellaneous: An ad-free, spyware-free web site for Excel users in the physical sciences. www.bowdoin.edu/~rdelevie/excellaneous/ (accessed Jan 20, 2015).
(16) The calculated values of the parameters must also be supplied to SolverAid. Solver can be used for that computation here because for exact fits, the weighting is irrelevant.
(17) Deming, W. E. *Statistical Adjustment of Data*; Dover: New York, 1964; earlier editions by Wiley: New York, 1938 and 1943.
(18) Powell, D. R.; Macdonald, J. R. Rapidly convergent iterative method for solution of generalized nonlinear least-squares problem. *Comput. J.* **1972**, *15*, 148−155.
(19) Britt, H. I.; Luecke, R. H. Estimation of parameters in nonlinear implicit models. *Technometrics* **1973**, *15*, 233−247.
(20) Wentworth, W. E. Rigorous Least Squares Adjustment. *J. Chem. Educ.* **1965**, *42*, 162−167 96−103.
(21) Tellinghuisen, J. Least-squares analysis of data with uncertainty in x and y: A Monte Carlo methods comparison. *Chemom. Intell. Lab. Syst.* **2010**, *103*, 160−169.
(22) Tellinghuisen, J. Van't Hoff analysis of $K°(T)$: How good... or bad? *Biophys. Chem.* **2006**, *120*, 114−120.
(23) Benesi, H. A.; Hildebrand, J. H. A Spectrophotometric Investigation of the Interaction of Iodine with Aromatic Hydrocarbons. *J. Am. Chem. Soc.* **1949**, *71*, 2703−2707.
(24) Daniels, F.; Williams, J. W.; Bender, P.; Alberty, R. A.; Cornwell, C. D. *Experimental Physical Chemistry*, 6th ed.; McGraw-Hill: New York, 1962.
(25) Baum, J. C.; Marzzacco, C. J.; Kendrow, C. Investigating the Thermodynamics of Charge-Transfer Complexes A Physical Chemistry Experiment. *J. Chem. Educ.* **2009**, *86*, 1330−1334.
(26) As an indication of the problem with this approach, applying eq 1 to eq 5 yields $s^2(\Delta G°) = s^2(\Delta H°) + s^2(T\Delta S°)$, while solving for $T\Delta S°$ and then using eq 1 yields $s^2(T\Delta S°) = s^2(\Delta G°) + s^2(\Delta H°)$. These two equations are inconsistent from lack of consideration of correlation, but the latter turns out to be a good approximation here.

DOI: 10.1021/ed500888r
J. Chem. Educ. XXXX, XXX, XXX−XXX

(27) For actual data, one may need to iterate the effective variance treatment because the parameters can change with the weights, and the effective variance depends on the parameters.
(28) Hamilton, W. C. *Statistics in Physical Science: Estimation, Hypothesis Testing, and Least Squares*; The Ronald Press Co.: New York, 1964.
(29) Meyer, E. F. A note on covariance in propagation of uncertainty. *J. Chem. Educ.* **1997**, *74*, 1339−1340.
(30) Salter, C. Error analysis using the variance-covariance matrix. *J. Chem. Educ.* **2000**, *77*, 1239−1243.
(31) Tellinghuisen, J. Statistical Error Propagation. *J. Phys. Chem. A* **2001**, *105*, 3917−3921.
(32) Spiess, A.-N. Propagate: Propagation of uncertainty. http://cran.r-project.org/web/packages/propagate/index.html (accessed Jan 20, 2015).
(33) Salter, C.; de Levie, R. Nonlinear fits of standard curves: A simple route to uncertainties in unknowns. *J. Chem. Educ.* **2002**, *79*, 268−270.
(34) Tellinghuisen, J. Least Squares in Calibration: Weights, Nonlinearity, and Other Nuisances. *Methods Enzymol.* **2009**, *454*, 259−285.
(35) Tellinghuisen, J. Bias and Inconsistency in Linear Regression. *J. Phys. Chem. A* **2000**, *104*, 11829−11835.

DOI: 10.1021/ed500888r
J. Chem. Educ. XXXX, XXX, XXX−XXX