Nonlinear Least Squares Curve Fitting with Microsoft Excel *Solver*

This section gives you a powerful tool for fitting data with nonlinear equations and teaches the underlying principles of least squares curve fitting.[1,2] If you master this section, you will never again wonder what the least squares algorithm does. The method requires a tool called *Solver* in the Microsoft Excel spreadsheet. Numerous other recipes for curve fitting have been described,[3-15] and this manual also gives a computer program written in C that fits equations to data.

Consider the problem of fitting the spectrophotometric protein calibration data in Table 5-2 in the textbook. Figure 1 shows that data from 0 to 20 μg of protein lie on a straight line, but data at 25 μg deviates from the line. We will fit the full set of 17 data points covering the range 0 to 25 μg with a quadratic equation:

$$y = Ax^2 + Bx + C \tag{1}$$

where y is corrected absorbance, x is μg of protein in the standard, and A, B and C are

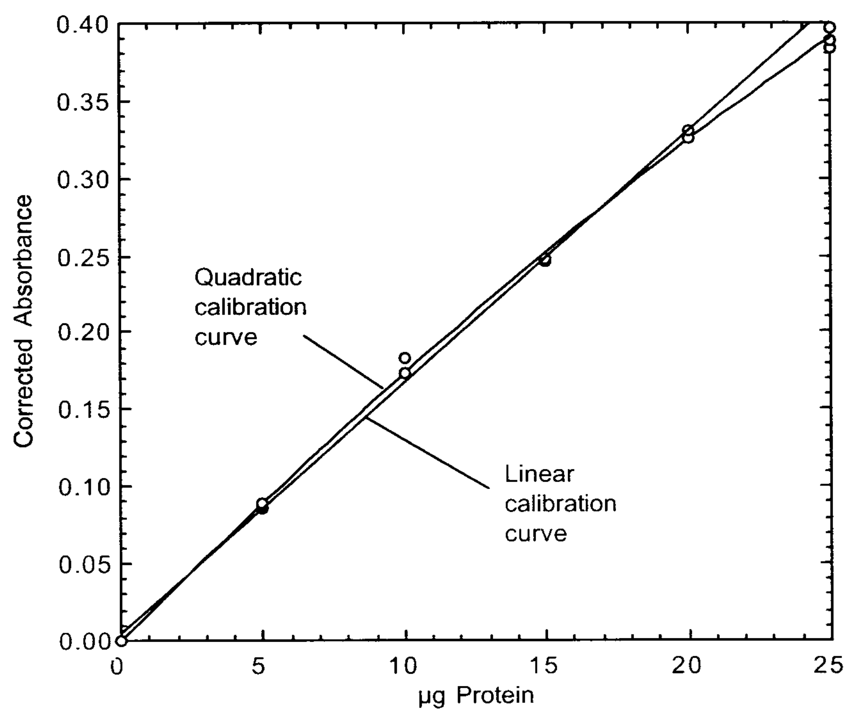

Figure 1. Calibration curves for spectrophotometric protein analysis. Circles are experimental data and the lines are least-squares fits.

constants to be found by the method of least squares. Our treatment is restricted to the common condition in which the uncertainty in y is much greater than the uncertainty in x.

We consider cases in which (1) all values of y have equal uncertainty or (2) different values of y have different uncertainty. In case (1), each datum is given equal weight for curve fitting. This procedure is the default (*unweighted*) method used when uncertainties in y are not known. Case (2) is a *weighted* least squares treatment, because more certain points are given more weight than less certain points.

Unweighted Least Squares

Experimental values of x and y are listed in the first two columns of the spreadsheet in Figure 2. The vertical deviation of the i^{th} point from the smooth curve is

$$\text{vertical deviation} = y_i\,(\text{observed}) - y_i\,(\text{calculated}) = y_i - (Ax_i^2 + Bx_i + C) \qquad (2)$$

The least squares criterion is to find values of A, B and C that minimize the sum of the squares of the vertical deviations of the points from the curve:

$$\text{sum} = \sum_{i=1}^{n} [y_i - (Ax_i^2 + Bx_i + C)]^2 \qquad (3)$$

where n is the total number of points (= 17).

Here are the steps to find values of A, B and C that minimize the sum in Equation 3:

1. Enter the measured values of x and y in columns 1 and 2 of Figure 2.
2. Temporarily assign the value 1 to A, B and C at the right side of the spreadsheet in cells F3, F4 and F5. (The labels in column E are for readability. They have no other function.)
3. In column C, calculate y from the measured value of x using Equation 1. For example, in cell C4, y is computed from the value of x in cell A4 and the values of A, B and C in cells F3, F4 and F5.
4. In column D, compute the vertical deviation with Equation 2 and then square the deviation. For example, D4 = (B4 - C4)^2.
5. In cell D22, compute the sum of the squares of vertical deviations in column D. The sum in cell D22 is the sum in Equation 3.
6. *The least squares criterion is to find values of A, B and C that minimize the sum in cell D22.* Microsoft Excel provides a tool called *Solver* that handles this problem in a manner that is transparent to the user. *Solver* is invoked in different manners by different versions of the software. At the time of this writing, Solver was found under the *Tools* menu. After invoking Solver, the screen in Figure 3 appears. If cell D22 was highlighted prior to calling *Solver*, then "D22" automatically appears in the upper left dialog box that says "Set Cell". If some other cell was highlighted, enter D22 in the Set

	A	B	C	D	E	F
1	Nonlinear least squares curve fitting					
2						
3	x	y(obs)	y(calc)	(ycalc-yobs)^2)	A =	1
4	0	-0.0003	1.0000	1.00E+00	B =	1
5	0	-0.0003	1.0000	1.00E+00	C =	1
6	0	0.0007	1.0000	9.99E-01		
7	5	0.0857	31.0000	9.56E+02		
8	5	0.0877	31.0000	9.56E+02		
9	5	0.0887	31.0000	9.56E+02		
10	10	0.1827	111.0000	1.23E+04		
11	10	0.1727	111.0000	1.23E+04		
12	10	0.1727	111.0000	1.23E+04		
13	15	0.2457	241.0000	5.80E+04		
14	15	0.2477	241.0000	5.80E+04		
15	20	0.3257	421.0000	1.77E+05		
16	20	0.3257	421.0000	1.77E+05		
17	20	0.3307	421.0000	1.77E+05		
18	25	0.3837	651.0000	4.23E+05		
19	25	0.3887	651.0000	4.23E+05		
20	25	0.3967	651.0000	4.23E+05		
21						
22			sum =	1.96E+06		
23						
24	y(calc) = F3*A4^2 + F4*A4 + F5					
25	sum = SUM(D4:D20)					

Figure 2. Initial spreadsheet for finding the best values of A, B and C in Equation. 1. Numbers in columns A and B are experimental data. Numbers in column F are initial guesses for A, B and C. The sum in cell D22 is the one to be minimized in Equation 3.

Cell box. The dollar signs are optional. Because we wish to minimize the value in cell D16, click "Min" on the second line beside "Equal to." Finally, write "F3,F4,F5" in the dialog box labeled "By Changing Cells." Now click the "Solve" button at the upper right and you have just asked the software to set the value of cell D22 to a minimum by changing values in cells F3, F4 and F5.

7. When *Solver* finishes its task in a few seconds, the values of A, B, C and sum become

$$A = 0.00369234$$
$$B = -0.1052121$$
$$C = 0.82601156$$
$$\text{Sum} = 2.47$$

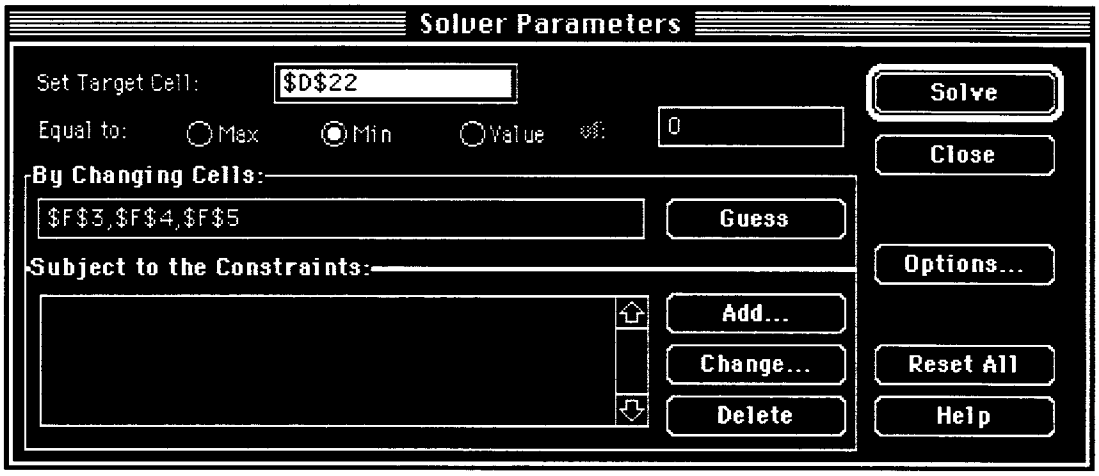

Figure 3. *Solver* screen with user input.

8. *Solver* has attempted to adjust the values in cells F3, F4 and F5 to minimize the sum in cell D22. The only problem is that if you inspect the calculated values of y in column C, you will find that they are not very good. The quality of the output depends on the quality of the original guesses for A, B and C. The values of A, B and C are now 0.00369234, -0.1052121, and 0.82601156, respectively. If you invoke *Solver* again with these starting values in cells F3, F4 and F5, the results are shown in Figure 4. The sum of the squares of the deviations has been reduced by a factor of 10^4 and comparison of y(obs) and y(calc) shows that the fit is excellent. The quadratic fit in Figure 1 comes from the values of A, B and C in Figure 4.
8. Try some different initial values for A, B and C (other than 1) to see if *Solver* finds the same solution. A given problem may have many local minima. We are seeking the best set of A, B and C to find the minimum sum in cell D22.

Weighted Least Squares

If different values of y have different uncertainties, it makes sense to force the least-squares curve to be closer to the more certain points than to the less certain points. That is, we assign a greater weight to the more certain points. If the uncertainty (standard deviation) in the measured value of y_i is s_i, then the weight assigned to point i is

$$\text{weight} = w_i = 1/s_i^2 \tag{4}$$

	A	B	C	D	E	F
1	Nonlinear least squares curve fitting					
2						
3	x	y(obs)	y(calc)	(ycalc-yobs)^2)	A =	-0.000117
4	0	-0.0003	-0.0007	1.57E-07	B =	0.01858525
5	0	-0.0003	-0.0007	1.57E-07	C =	-0.000696
6	0	0.0007	-0.0007	1.95E-06		
7	5	0.0857	0.0893	1.30E-05		
8	5	0.0877	0.0893	2.58E-06		
9	5	0.0887	0.0893	3.67E-07		
10	10	0.1827	0.1735	8.54E-05		
11	10	0.1727	0.1735	5.77E-07		
12	10	0.1727	0.1735	5.77E-07		
13	15	0.2457	0.2518	3.68E-05		
14	15	0.2477	0.2518	1.65E-05		
15	20	0.3257	0.3242	2.18E-06		
16	20	0.3257	0.3242	2.18E-06		
17	20	0.3307	0.3242	4.20E-05		
18	25	0.3837	0.3908	5.09E-05		
19	25	0.3887	0.3908	4.54E-06		
20	25	0.3967	0.3908	3.44E-05		
21						
22			sum =	2.94E-04		
23						
24	y(calc) = F3*A4^2 + F4*A4 + F5					
25	sum = SUM(D4:D20)					

Figure 4. Appearance of spreadsheet after two rounds of executing *Solver*.

In Figure 5, a set of fictitious uncertainties in y are listed in column C under the heading "error(y)." Weights computed with Equation 4 appear in column D. Columns E and F are calculated with Equations 1 and 2, just as they were in Figure 2. Column G contains weighted, squared residuals, obtained by multiplying the squared residuals in column F times the weights in column D. Cell G22 contains the sum of weighted residuals. *Solver* is then invoked to vary the values of A, B and C (in cells G24, G25 and G26) to minimize the sum of weighted residuals in cell G22. The final values of A, B and C are slightly different from the final values in the unweighted procedure.

	Unweighted fit:	Weighted fit:
A =	-0.000 117	-0.000 105
B =	0.018 585	0.018 327
C =	-0.000 696	-0.000 156

	A	B	C	D	E	F	G
1	Weighted nonlinear least squares curve fitting						
2							
3	x	y(obs)	error(y)	weight (w)	y(calc)	(ycalc-yobs)^2)	w*(ycalc-yobs)^2
4	0	-0.0003	0.002	2.50E+05	0.000	2.08E-08	5.20E-03
5	0	-0.0003	0.002	2.50E+05	0.000	2.08E-08	5.20E-03
6	0	0.0007	0.002	2.50E+05	0.000	7.32E-07	1.83E-01
7	5	0.0857	0.003	1.11E+05	0.089	9.97E-06	1.11E+00
8	5	0.0877	0.003	1.11E+05	0.089	1.34E-06	1.49E-01
9	5	0.0887	0.003	1.11E+05	0.089	2.47E-08	2.75E-03
10	10	0.1827	0.004	6.25E+04	0.173	1.02E-04	6.35E+00
11	10	0.1727	0.004	6.25E+04	0.173	5.80E-09	3.62E-04
12	10	0.1727	0.004	6.25E+04	0.173	5.80E-09	3.62E-04
13	15	0.2457	0.005	4.00E+04	0.251	2.96E-05	1.19E+00
14	15	0.2477	0.005	4.00E+04	0.251	1.19E-05	4.74E-01
15	20	0.3257	0.006	2.78E+04	0.324	1.64E-06	4.56E-02
16	20	0.3257	0.006	2.78E+04	0.324	1.64E-06	4.56E-02
17	20	0.3307	0.006	2.78E+04	0.324	3.95E-05	1.10E+00
18	25	0.3837	0.007	2.04E+04	0.392	7.65E-05	1.56E+00
19	25	0.3887	0.007	2.04E+04	0.392	1.40E-05	2.86E-01
20	25	0.3967	0.007	2.04E+04	0.392	1.81E-05	3.69E-01
21							
22	error(y) = estimated uncertainty in y					sum =	9.55E+00
23	weight = 1/(error^2)						
24						A=	-0.000105
25						B=	0.018327
26						C=	-0.000156

Figure 5. Spreadsheet for weighted least-squares fit to Equation 1 after invoking *Solver*. Measured uncertainties in y are listed in column C and weights computed with Equation 4 appear in column D. The parameters A, B and C are listed at the lower right.

Estimating Uncertainties in the Least-Squares Parameters

Uncertainties in A, B and C are as important as the values of the parameters themselves. Small uncertainties mean that Equation 1 fits the experimental data well. Large uncertainties mean that there is considerable error in the measured points (x,y) or that the model is inappropriate.

Here is how to estimate uncertainties in A, B and C of Figure 2 by the "jackknife" procedure:[16,17]

1. Delete the first row of data in Figure 2 and then use *Solver* to find the least-squares parameters A, B and C. For this purpose, the initial guesses for A, B and C should be those found by *Solver* in the previous run. Copy and paste the values of A, B and C

into the first row of the new spreadsheet in Figure 6. Restore the first row of data and delete the second row to generate a second solution. Paste this solution into the next line of Figure 6. Repeat this process a total of *n* times and paste each result into Figure 6. It is not necessary to actually delete data from the spreadsheet. You can write the sum in cell D22 in the form D22 = D4 + D5 + D6 + D7 + D8 + D9 + D10 + D11 + D12 + D13 + D14 +D15 + D16 + D17 + D18 + D19 + D20. Delete one term in the sum each time to generate the 17 lines of Figure 6. It took approximately 10 minutes of work to generate the data for Figure 6.

2. For each column in Figure 6, compute the standard deviation with the function STDEV.
3. Find the standard error for each parameter (A, B and C) by multiplying its standard deviation times $(n-1)/\sqrt{n}$, where n is the number of data points (= 17). Standard errors are estimates of uncertainty in the least-squares parameters. The final result is

$$A = -0.000\,117 \pm 0.000\,021$$
$$B = 0.018\,585 \pm 0.000\,461$$
$$C = -0.000\,696 \pm 0.001\,006$$

The same process can be carried out for the weighted least squares procedure in Figure 5 by deleting one data point at a time.

	A	B	C
1	Jackknife procedure for uncertainties in		
2	y = Ax^2 + Bx + C		
3	A	B	C
4	-0.000118	0.018608	-0.000847
5	-0.000118	0.018608	-0.000847
6	-0.000119	0.018664	-0.001229
7	-0.000118	0.018593	-0.000276
8	-0.000118	0.018589	-0.000509
9	-0.000117	0.018587	-0.000625
10	-0.000105	0.018314	-0.000629
11	-0.000118	0.018608	-0.000701
12	-0.000118	0.018608	-0.000701
13	-0.000125	0.018793	-0.001080
14	-0.000122	0.018724	-0.000953
15	-0.000116	0.018562	-0.000629
16	-0.000116	0.018562	-0.000629
17	-0.000115	0.018482	-0.000402
18	-0.000105	0.018387	-0.000345
19	-0.000113	0.018526	-0.000591
20	-0.000127	0.018749	-0.000985
21	Standard deviations =		
22	0.000005	0.000119	0.000259
23	Standard error = Std. dev. * 16/Sqrt(17)		
24	0.000021	0.000461	0.001006

Figure 6. Jackknife procedure for estimating uncertainties in parameters A, B and C.

References

1. S. Walsh and D. Diamond *Talanta* **1995**, *42*, 561.
2. D. C. Harris, *J. Chem. Ed.* **1998**, *75*, 119; E. J. Billo, *Microsoft Excel for Chemists* (New York: Wiley, **1997**), Chapter 17.
3. R. de Levie, *Principles of Quantitative Chemical Analysis* (New York: McGraw-Hill, **1997**).
4. W. E. Wentworth, *J. Chem. Ed.* **1965**, *42*, 96, 162.
5. D. York, *Can. Y. Phys.* **1966**, *44*, 1079.
6. J. A.Irvin and T. I. Quickenden, *J. Chem. Ed.* **1983**, *60*, 711.
7. R. de Levie, *J. Chem. Ed.* **1986**, *63*, 10.
8. R. T. O'Neill and D. C. Flaspohler, *J. Chem. Ed.* **1990**, *67*, 40.
9. P. J. Ogren and J. R. Norton, *J. Chem. Ed.* **1992**, *69*, A130.
10. E. Heilbronner, *J. Chem. Ed.* **1979**, *56*, 240.
11. M. D. Pattengill and D. E. Sands, *J. Chem. Ed.* **1979**, *56*, 244.
12. A. H. Kalantar, *J. Chem. Ed.* **1987**, *64*, 28.
13. S. D. Christian and E. E. Tucker, *J. Chem. Ed.* **1984**, *61*, 788.
14. S. D. Christian, E. H. Lane and F. Garland, *J. Chem. Ed.* **1974**, *51*, 475.
15. T. Trindle, *J. Chem. Ed.* **1983**, *60*, 566.
16. Caceci, M. S. *Anal. Chem.* **1989**, *61*, 2324.
17. Bradley, E; Gong, G. *Am. Statistician* **1983**, *37*, 36.