edited by
Steven D. Gammon
University of Idaho
Moscow, ID 83844

Nonlinear Least-Squares Curve Fitting with Microsoft Excel Solver

Daniel C. Harris
Chemistry & Materials Branch, Research & Technology Division, Naval Air Warfare Center,China Lake, CA 93555

A powerful tool that is widely available in spreadsheets provides a simple means of fitting experimental data to nonlinear functions. The procedure is so easy to use and its mode of operation is so obvious that it is an excellent way for students to learn the underlying principle of least-squares curve fitting. The purpose of this article is to introduce the method of Walsh and Diamond (*1*) to readers of this *Journal*, to extend their treatment to weighted least squares, and to add a simple method for estimating uncertainties in the least-square parameters. Other recipes for curve fitting have been presented in numerous previous papers (*2–16*).

Consider the problem of fitting the experimental gas chromatography data (*17*) in Figure 1 with the van Deemter equation:

$$y = Ax + B/x + C \qquad (1)$$

where y is plate height (mm), x is flow rate (mL/min), and A, B, and C are constants to be found by the method of least squares. This paper is restricted to the situation in which the uncertainty in y is much greater than the uncertainty in x.

We treat cases in which (i) all values of y have equal uncertainty or (ii) different values of y have different uncertainty. In case i, 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 ii is a *weighted* least squares treatment, because more certain points are given more weight than less certain points.

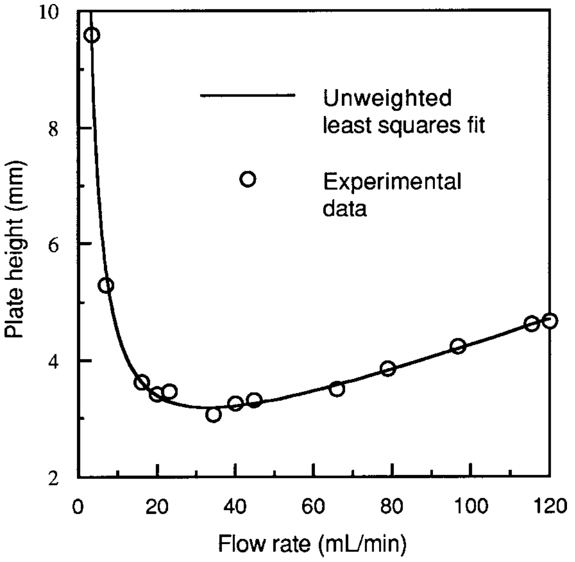


Figure 1. Plate height versus flow rate in a gas chromatography experiment. Circles are experimental data (*17*) and the solid line is the best fit to eq 1 by the method of least squares in Figures 2–4.

Unweighted Least Squares

Experimental values of x and y from Figure 1 are listed in the first two columns of the spreadsheet in Figure 2. The vertical deviation of the ith point from the smooth curve is

$$\begin{aligned} \text{vertical deviation} &= y_i\ (\text{observed}) - y_i\ (\text{calculated}) \\ &= y_i - (Ax_i + B/x_i + C) \end{aligned} \qquad (2)$$

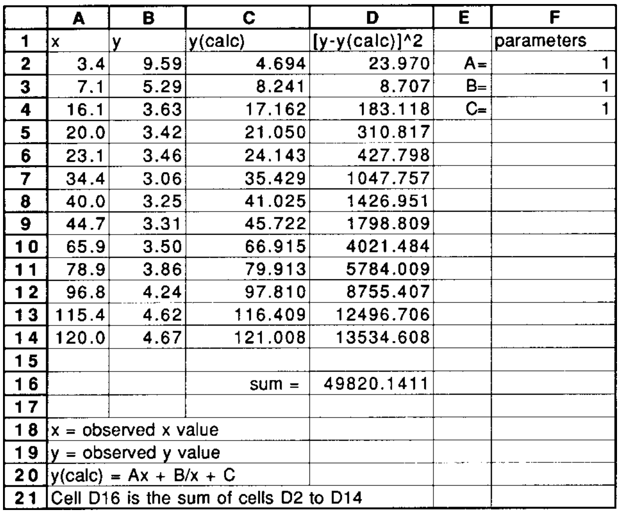
The least squares criterion is to find values of A, B, and C in eq 1 that minimize the sum of the squares of the vertical deviations of the points from the curve:

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

where n is the total number of points (= 13 in Fig. 1).

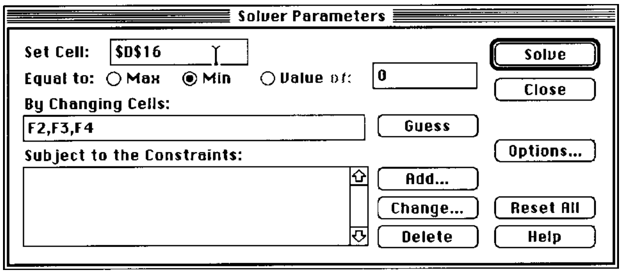
Here are the steps to find the best values of A, B, and C that minimize the sum in eq 3:

1. List 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 F2, F3, and F4. (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 (eq 1). For example, in cell C2, y is computed from the value of x in cell A2 and the values of A, B, and C in cells F2, F3, and F4.



	A	B	C	D	E	F
1	x	y	y(calc)	[y-y(calc)]^2		parameters
2	3.4	9.59	4.694	23.970	A=	1
3	7.1	5.29	8.241	8.707	B=	1
4	16.1	3.63	17.162	183.118	C=	1
5	20.0	3.42	21.050	310.817		
6	23.1	3.46	24.143	427.798		
7	34.4	3.06	35.429	1047.757		
8	40.0	3.25	41.025	1426.951		
9	44.7	3.31	45.722	1798.809		
10	65.9	3.50	66.915	4021.484		
11	78.9	3.86	79.913	5784.009		
12	96.8	4.24	97.810	8755.407		
13	115.4	4.62	116.409	12496.706		
14	120.0	4.67	121.008	13534.608		
15						
16			sum =	49820.1411		
17						
18	x = observed x value					
19	y = observed y value					
20	y(calc) = Ax + B/x + C					
21	Cell D16 is the sum of cells D2 to D14					

Figure 2. Initial spreadsheet for finding the best values of *A*, *B*, and *C* in eq 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 D16 is the one to be minimized in eq 3.



Set Cell: $D$16

Equal to: ○ Max ● Min ○ Value of: 0

By Changing Cells: F2,F3,F4

Subject to the Constraints:

Figure 3. Solver screen with user input.

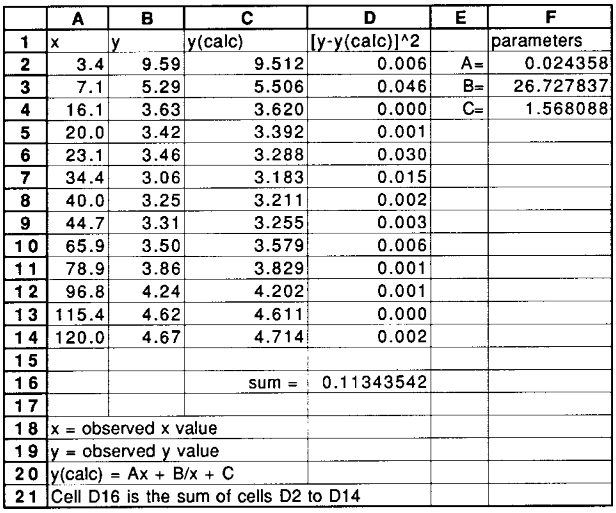
4. In column D, compute the vertical deviation in eq 2 and then square the deviation. For example, D2 = (B2 – C2)^2.
5. In cell D16, compute the sum of the squares of vertical deviations in column D. The sum in cell D16 is the sum in eq 3.
6. *The least squares criterion is to find values of* A, B, *and* C *that minimize the sum in cell D16.* 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. In version 5, Solver is found under the Tools menu. After invoking Solver, the screen in Figure 3 appears. If cell D16 was highlighted prior to calling Solver, then "$D$16" automatically appears in the upper left dialog box that says "Set Cell". If some other cell was highlighted, enter D16 in the Set 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 "F2,F3,F4" 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 D16 to a minimum by changing values in cells F2, F3, and F4.
7. When Solver finishes its task in a few seconds, the spreadsheet will appear as in Figure 4. Solver has adjusted the values in cells F2, F3, and F4 to minimize the sum in cell D16. The values of *A*, *B*, and *C* in cells F2, F3, and F4 were used to plot the curve in Figure 1.[1]
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 lowest minimum sum in cell D16.

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

$$\text{weight} = w_i = 1/s_i^2 \qquad (4)$$

In Figure 5, measured uncertainties in *y* are listed in column C under the heading "error(y)". Weights computed with eq 4 appear in column D. Columns E and F are calculated with eqs 1 and 2, just as they were in Figure 2. Column G contains weighted square residuals, obtained by multiplying the square residuals in column F times the weights in column D. Cell G16 contains the sum of weighted residuals. Solver is then invoked to vary the values of *A*, *B*, and *C* (in cells G20, G21, and G22) to minimize the sum of weighted residuals in cell G16. The final values of *A*, *B*, and *C* are somewhat different from the final values in the unweighted procedure in Figure 4.



	A	B	C	D	E	F
1	x	y	y(calc)	[y-y(calc)]^2		parameters
2	3.4	9.59	9.512	0.006	A=	0.024358
3	7.1	5.29	5.506	0.046	B=	26.727837
4	16.1	3.63	3.620	0.000	C=	1.568088
5	20.0	3.42	3.392	0.001		
6	23.1	3.46	3.288	0.030		
7	34.4	3.06	3.183	0.015		
8	40.0	3.25	3.211	0.002		
9	44.7	3.31	3.255	0.003		
10	65.9	3.50	3.579	0.006		
11	78.9	3.86	3.829	0.001		
12	96.8	4.24	4.202	0.001		
13	115.4	4.62	4.611	0.000		
14	120.0	4.67	4.714	0.002		
15						
16			sum =	0.11343542		
17						
18	x = observed x value					
19	y = observed y value					
20	y(calc) = Ax + B/x + C					
21	Cell D16 is the sum of cells D2 to D14					

Figure 4. Appearance of spreadsheet from Figure 2 after Solver has finished its operation.

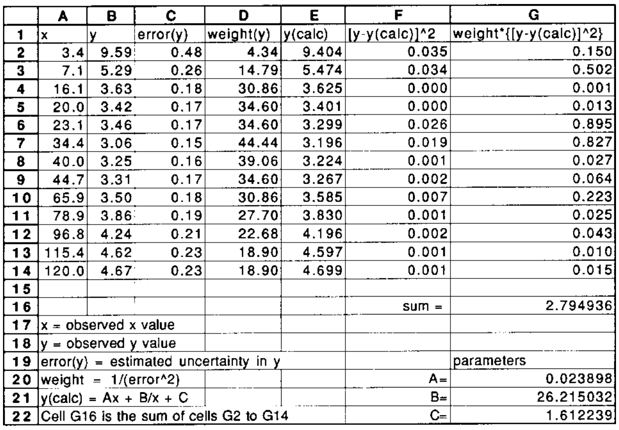
Estimating Uncertainties in the Least-Squares Parameters

Uncertainties in *A*, *B*, and *C* are as important as the values of the parameters themselves. Small uncertainties mean that the model (eq 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.

Figure 6 shows how to estimate uncertainties in *A*, *B*, and *C* of Figure 4 by the "jackknife" procedure (*18, 19*). Here are the steps:

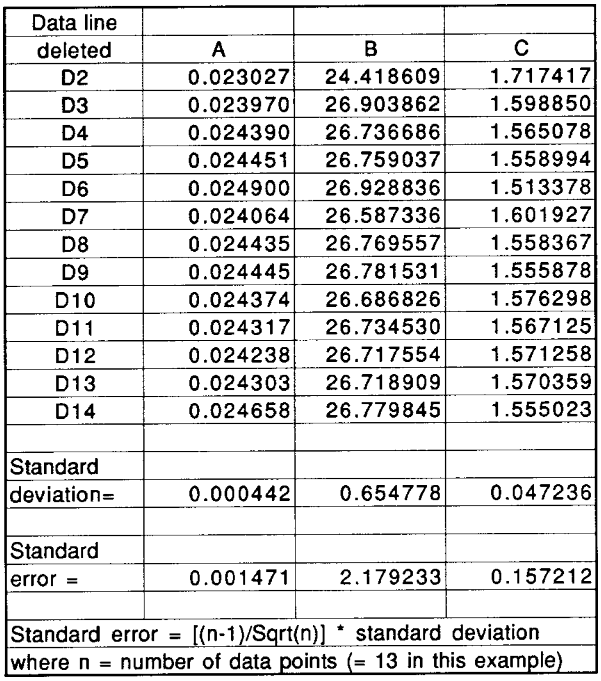
1. Delete the first row of data in Figure 4 (cells A2, B2, C2, and D2) and then use Solver to find the least-squares parameters *A*, *B*, and *C*. For this purpose, the initial values of *A*, *B*, and *C* in cells F2–F4 should be those found by Solver in the previous run. Copy and paste the values of *A*, *B*, and *C* into the first row of Figure 6. Restore the first row of Figure 4 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 D16 of Figure 4 in the form D16 = D2 + D3 + D4 + D5 + D6 + D7 + D8 + D9 + D10 + D11 + D12 + D13 + D14. Delete one term in the sum each time to generate the 13 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 (= 13 in Fig. 6). Standard errors are estimates of uncertainty in the least-squares parameters. The final result for Figures 4 and 6 is therefore[2]

Parameter	Raw result	Rounded result
A	0.024358 ± 0.001471	0.0244 ± 0.0015
B	26.727837 ± 2.179233	26.7 ± 2.2
C	1.568088 ± 0.157212	1.57 ± 0.16



	A	B	C	D	E	F	G
1	x	y	error(y)	weight(y)	y(calc)	[y-y(calc)]^2	weight*{[y-y(calc)]^2}
2	3.4	9.59	0.48	4.34	9.404	0.035	0.150
3	7.1	5.29	0.26	14.79	5.474	0.034	0.502
4	16.1	3.63	0.18	30.86	3.625	0.000	0.001
5	20.0	3.42	0.17	34.60	3.401	0.000	0.013
6	23.1	3.46	0.17	34.60	3.299	0.026	0.895
7	34.4	3.06	0.15	44.44	3.196	0.019	0.827
8	40.0	3.25	0.16	39.06	3.224	0.001	0.027
9	44.7	3.31	0.17	34.60	3.267	0.002	0.064
10	65.9	3.50	0.18	30.86	3.585	0.007	0.223
11	78.9	3.86	0.19	27.70	3.830	0.001	0.025
12	96.8	4.24	0.21	22.68	4.196	0.002	0.043
13	115.4	4.62	0.23	18.90	4.597	0.001	0.010
14	120.0	4.67	0.23	18.90	4.699	0.001	0.015
15							
16						sum =	2.794936
17	x = observed x value						
18	y = observed y value						
19	error(y) = estimated uncertainty in y						parameters
20	weight = 1/(error^2)					A=	0.023898
21	y(calc) = Ax + B/x + C					B=	26.215032
22	Cell G16 is the sum of cells G2 to G14					C=	1.612239

Figure 5. Spreadsheet for weighted least-squares calculation.



Data line			
deleted	A	B	C
D2	0.023027	24.418609	1.717417
D3	0.023970	26.903862	1.598850
D4	0.024390	26.736686	1.565078
D5	0.024451	26.759037	1.558994
D6	0.024900	26.928836	1.513378
D7	0.024064	26.587336	1.601927
D8	0.024435	26.769557	1.558367
D9	0.024445	26.781531	1.555878
D10	0.024374	26.686826	1.576298
D11	0.024317	26.734530	1.567125
D12	0.024238	26.717554	1.571258
D13	0.024303	26.718909	1.570359
D14	0.024658	26.779845	1.555023
Standard			
deviation=	0.000442	0.654778	0.047236
Standard			
error =	0.001471	2.179233	0.157212
Standard error = [(n-1)/Sqrt(n)] * standard deviation			
where n = number of data points (= 13 in this example)			

Figure 6. Estimating uncertainties in least-squares parameters of Figure 4 by the jackknife procedure (*18, 19*).

The same process can be carried out for the weighted least squares procedure in Figure 5 by deleting one data point at a time to generate 13 “jackknifed” data sets for input to Figure 6.

Acknowledgments

We are grateful to Joe Roberts and Kelvin Higa for helpful suggestions that improved this manuscript.

Notes

1. The smooth curve in Figure 1 was obtained from the values of *A*, *B*, and *C* generated by Solver in Figure 4. In a fresh column of a spreadsheet, values of *x* from 3 to 120 were entered. For each *x*, a value of *y* was calculated in the next column from the equation $y = Ax + B/x + C$. A graphing program was then used to plot the calculated points and to draw a smooth curve between the points. Discrete experimental values of *x* and *y* listed in columns A and B of Figure 2 were superimposed on the same graph.

2. We retain an extra, nonsignificant digit for *A*, *B*, and *C* to reduce future roundoff errors if these parameters are used in subsequent computations. In the spreadsheets, the number of decimal places chosen for display was selected arbitrarily for readability. The spreadsheet retains more precision than the number of digits displayed.

Literature Cited

1. Walsh, S.; Diamond, D. *Talanta* **1995,** *42*, 561.
2. Wentworth, W. E. *J. Chem. Educ.* **1965,** *42*, 96, 162.
3. York, D. *Can. Y. Phys.* **1966,** *44*, 1079.
4. Irvin, J. A.; Quickenden, T. I. *J. Chem. Educ.* **1983,** *60*, 711.
5. de Levie, R. *J. Chem. Educ.* **1986,** *63*, 10.
6. O’Neill, R. T.; Flaspohler, D. C. *J. Chem. Educ.* **1990,** *67*, 40.
7. Ogren, P. J.; Norton, J. R. *J. Chem. Educ.* **1992,** *69*, A130.
8. Heilbronner, E. *J. Chem. Educ.* **1979,** *56*, 240.
9. Pattengill, M. D.; Sands, D. E. *J. Chem. Educ.* **1979,** *56*, 244.
10. Kalantar, A. H. *J. Chem. Educ.* **1987,** *64*, 28.
11. Christian, S. D.; Tucker, E. E. *J. Chem. Educ.* **1984,** *61*, 788.
12. Christian, S. D.; Lane, E. H.; Garland, F. *J. Chem. Educ.* **1974,** *51*, 475.
13. Trindle, T. *J. Chem. Educ.* **1983,** *60*, 566.
14. Machuca-Herrera, J. O. *J. Chem. Educ.* **1997,** *74*, 448.
15. Zielinski, T. J.; Allendoerfer, R. D. *J. Chem. Educ.* **1997,** *74*, 1001.
16. Lieb, S. G. *J. Chem. Educ.* **1997,** *74*, 1009.
17. Moody, H. W. *J. Chem. Educ.* **1982,** *59*, 290.
18. Caceci, M. S. *Anal. Chem.* **1989,** *61*, 2324.
19. Bradley, E; Gong, G. *Am. Statistician* **1983,** *37*, 36.