Nonlinear regression analysis of data using a spreadsheet

BY ANGUS M. BROWN

The practice of applying curve fitting techniques to describe data is widespread in all fields of biology. The purpose of curve fitting of biological data is to describe data in the universally recognized form \( y = f(x) \), where \( y \) is the dependent variable and is measured in the experiment, and \( x \) is controlled during the experiment and is called the independent variable as its value on the \( x \)-axis is fixed \( f \) is the function used to describe the relationship between \( x \) and \( y \), and takes the form of an equation composed of one or more parameters. In general, the better the fit of the curve the more accurately it describes the data. Applying linear fits to data is a comparatively simple procedure and can be carried out easily with a pocket calculator. Describing data with nonlinear functions (nonlinear regression) is more problematical. Prior to the advent of personal computers, data was linearly transformed, then fit with a linear function. However, in the age of personal computers there is no shortage of specialist programs that will carry out nonlinear regression. Many of the programs of choice of biologists carrying out nonlinear regression analysis are expensive. Contain many redundant features. These programs do not generally handle data well and display data, graphs, curve fits, and analysis in a multitude of separate windows that may be confusing to the user. Thus, data relating to one experiment may be contained in the acquisition program, data handling program, curve fitting program, and presentation program. The track of these data files from multiple experiments can be a logistical nightmare, particularly for a laboratory head. A solution is to reduce the number of programs involved in data analysis by carrying out as much analysis as possible in one program. Excel (Microsoft, Redmond, WA) is part of the Microsoft Office suite that is usually offered as part of the computer package upon purchase. Excel offers a user-friendly interface with good data handling capabilities, built-in mathematical functions, and instantaneous graphing. In addition, the program contains the SOLVER function, which is well suited to fitting data with nonlinear functions via an iterative algorithm. In this article, a method for carrying out nonlinear regression analysis using Excel is described. Prior to the advent of personal computers, data was linearly transformed, then fit with a linear function. However, because this formula must be expressed as an array formula, press Ctrl+Shift+Enter. This encloses the whole formula within a pair of curly brackets ({}), denoting it as an array formula.

### Method

The method involves manual data entry and graphing of data, followed by curve fitting and displaying the resulting curve on top of the data. The goodness of fit is calculated so that the accuracy of fit can be assessed.

### Nonlinear regression

The description of data by a function is carried out by the process of iterative (i.e., cyclical) nonlinear regression. This process minimizes the value of the squared sum of the difference between the data and the fit.

\[
SS = \sum_{i=1}^{n} [y_i - y_{\text{fit}}]^2 \tag{1}
\]

where \( y \) is the data point, \( y_{\text{fit}} \) is the value of the curve at point \( y \), and \( SS \) is the sum of the squares of all the data points. It involves the user providing initial estimates of the parameter values upon which the first iteration calculates an initial \( SS \) value. The second iteration involves changing the parameter values by a small amount and recalculating the \( SS \) value. This process is repeated many times to ensure that the changes in the parameter values result in the smallest possible value of \( SS \). SOLVER employs the generalized reduced gradient (GRG) method of iteration.

The following example illustrates how to use the spreadsheet program’s SOLVER function to fit data with user-input nonlinear functions. The example used is the Boltzmann equation, but any nonlinear function can be used simply by substituting the relevant equation.

\[
y = \frac{1}{1 + \exp \left( \frac{V - \text{Slope}}{\text{V_slope}} \right)} \tag{2}
\]

where \( y \) is the dependent variable, \( x \) is the independent variable (Voltage), and \( V \) and Slope are the parameter values. \( V \) is the half activation voltage and Slope describes the slope at the point \( V \) and indicates the steepness of the curve. This paper does not address the critical issue of which functions are suitable to describe individual data, but this topic is discussed in detail elsewhere.

#### Configuring the spreadsheet for nonlinear regression

1. Input onto a spreadsheet the raw data in two columns, the \( x \) column containing the independent variable, and the \( y \) column containing the dependent variable. This is illustrated as Columns A and B of Figure 1a.
2. Graph the data contained in cells A2 to B20.
3. Enter labels in cells G1 to G8 to describe the contents of the adjacent cells. In cell G1 enter \( V \), which will describe the parameter in cell H1. For cell H1 select the Insert menu, choose Name then Define for cell H1. Name the cell \( V \). Similarly, for cells G2 to G8, enter Slope, Mean of \( y \), df, SE of \( y \), R2, Critical t and CI, respectively. Name cells H2 to H8, Slope, Mean of \( y \), df, SE of \( y \), RSQ, Critical t and CI, respectively.
4. Insert initial estimates of the parameters \( V \) and Slope into cells G1 and G2, respectively. Approximate estimates are 80 and 30, respectively.
5. In Column C (Boltzmann) enter the equation describing the Boltzmann function. This has been rearranged from Equation 2 into a form that the program recognizes: 

\[
(1/1+\exp(V-A2)/Slope)), \text{where } V \text{ and } \text{Slope} \text{ refer to the parameter values in cells H1 and H2.}
\]

6. Copy the equation from cell C2 down to and including C20.
7. The mean of the \( y \) values is calculated by entering the following formula in H3: 

\[
\text{=AVERAGE(B2:B20)}.
\]
8. The degrees of freedom (df) is defined as the number of data points minus the number of parameters in the function. It is calculated by entering the following formula in H4: 

\[
\text{=COUNT(B2:B20)-COUNT(H1:H2)}.
\]
9. The standard error of the \( y \) values is calculated by entering this formula in H5: 

\[
\text{=SQRT(SUM((B2:B20-Mean_of_y)^2))/df)}.
\]
10. The R² value, the correlation index or coefficient of determination, is calculated by entering the following formula in H6 and expressing it as an array formula as described above: 

\[
\text{=1-SUM((B2:B20-C2:C20)^2)/SUM((B2:B20-Mean_of_y)^2)}.
\]
11. In order for the confidence interval of the fit to be calculated, the critical \( t \) value at a significance level of 95% is calculated by entering the following formula in H7: 

\[
\text{=TINV(0.05/df)}.
\]

The confidence interval is calculated by entering the following formula in H8: 

\[
\text{=CRIT-
\]

**Figure 1. Spreadsheet template for nonlinear regression. a) Formulae used in the curve fitting procedure. The \( (y,y) \) data are entered into Columns A and B, respectively, with Column C used to generate the fit based on the parameters in Cells H1 and H2. Columns D and E calculate the 95% confidence interval around the fit. b) The solution of the fit calculated by SOLVER.**
60 / OCTOBER 2001

Enter the following formula in D2:

= C2 + CI, and copy it down to D20. Similarly, enter

= C2 - CI in E2 and copy down to E20. This calculates

the upper and lower confidence limits (95%) of the fit.

12. The SE of the y values, R^2 and CI, are automatically calculated: 0.122, 0.895, and 0.257, respectively.

13. Figure 1a illustrates the spreadsheet template with the formulas used in the fitting protocol displayed.

14. Graph Columns C, D and E versus Column A such that they are displayed as continuous lines on the graph (shown in Figure 2a). It can be seen that the initial estimate (blue line) is not a good fit of the data with large confidence limits (red line).

15. Open the SOLVER function, which can be found under the Tools menu.

16. In Set Target Cell box enter RSQ.

17. Set the Equal To option to Max. SOLVER attempts to maximize the value of R^2.

18. In By Changing Cells box, enter V, Slope.

19. Choose Solve to perform the fit. The program will iteratively cycle through the fitting routine, changing the parameter values of V and Slope until the largest value of R^2 is calculated. These changes will be displayed on the spreadsheet template, as illustrated in cells H1 and H2 of Figure 1. The optimal values of V and Slope are 99.366 and 24.388, respectively, and the maximal value of R^2 is 0.997. The blue line in Figure 2a illustrates the best fit and it is clear that it is an improvement over the fit provided by the initial parameter values. Additionally, the confidence intervals (red line) around the fit have been reduced.

20. The optimal values of V and Slope are 99.366 and 24.388, respectively, and the maximal value of R^2 is 0.997. The blue line in Figure 2b illustrates the best fit and it is clear that it is an improvement over the fit provided by the initial parameter values. Additionally, the confidence intervals (red line) around the fit have been reduced.

Conclusion

The procedure described in this paper allows the user to carry out nonlinear regression analysis of data within an Excel spreadsheet without the need of specialist curve fitting programs. The procedure involves manually entering data and graphing it. The curve fitting procedure utilizing the SOLVER function is performed and the resulting curve fit overlaid on the data. In addition, the R^2 value, an index of goodness of fit, and the 95% confidence intervals are calculated and displayed.

Once the spreadsheet template has been set up, it can be repeatedly used for new sets of data. If a new function is used to describe data, it is entered manually in Column C and the appropriate parameters are designated in Column D. Significant errors in data entry and a solution never found if inappropriate values are entered.

The spreadsheet template described in this paper is available for download from the author’s Web site at http://faculty.washington.edu/ambrown/.

References


