

Curve Fitting Best Practice

Part 4: Parameters and Residuals

When generating a data set, the researcher should know which measurements to take and how they want the data represented when all the results are recorded. A curve fitting model is then used to understand the data, with the parameters of that model representing the results that the researcher wants to extract.

A curve fitting model defines Y as a function of X along with one or more parameters. A set of starting values are assigned to the model's parameters to begin the fitting process and those values are varied as the process proceeds. By changing the parameter values in increments, the complete shape of a curve can be defined and altered. For example, in the four-parameter model below, the parameter values A, B, C and D essentially define the complete shape of that curve.

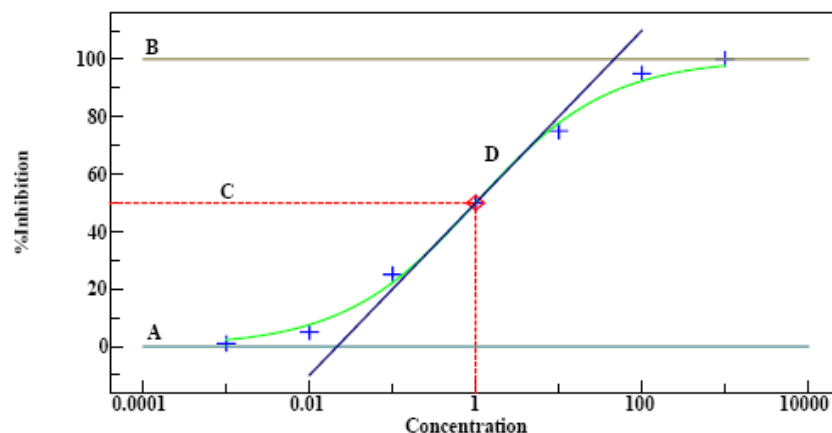


Fig 1: An example four-parameter model

On each successive cycle of the fitting process, the curve's goodness of fit is assessed and measured using the sum of each data point's residual squared (please see Residuals on page 2 of this article for more information) to give an indication as to whether the curve is converging on the optimum fit for the data. Based on this indication of goodness of fit, the fitting algorithm changes all of the curve's parameter values at each point in the cycle, defining the curve's shape and ultimately finding the best curve fit to the data.

Appropriate starting parameter values for each model are inbuilt in most curve fitting software applications to begin the fitting process. Starting parameter values are very important in determining whether a fit will converge or not. For example, in an example three-parameter sigmoid model as shown in Fig 2, if the lower asymptote (parameter A) was set at 10,000, this is significantly far away from where the fitted final value should be and so the fit may not converge. Assigning poor starting parameter values is one of the primary causes of unsuccessful fits.

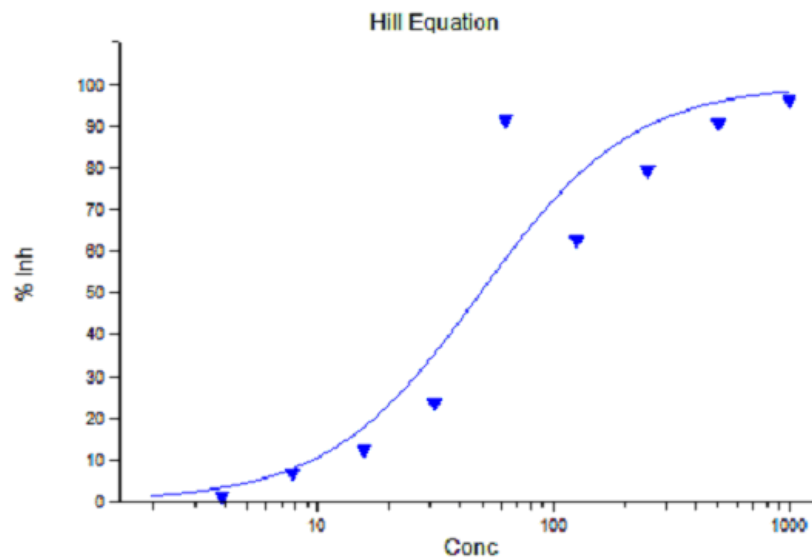


Fig 2: An example three-parameter model

Good starting parameter values can be reasonably estimated based on the data you are fitting. These estimates could be based on calculations and can be re-used for different datasets thus giving a good routine way of ensuring that a fit will converge.

Example Parameter Estimates

Taking the four-parameter sigmoidal model in Fig 1 as an example, the initial value for:

- Parameter A could be the min Y value provided
- Parameter B could be the max Y value provided
- Parameter C, the log EC₅₀, could be the middle of the X data range (e.g., $[\text{Max Conc} + \text{Min Conc}]/2$)
- Parameter D, the slope factor, could be 1 or -1 depending on the shape of the data

Note: The fitting engine provided by IDBS in all its applications uses a method called pre-fitting. Checking the pre-fit box for any of the parameters will force the math engine to create initial estimates based on calculations for the data provided. XLfit has been programmed to 'understand' what each model represents so that when pre-fitting is selected as an option, XLfit calculates suitable fitting values depending on the data set provided and applies appropriate starting parameter values.

Residuals

Goodness of fit is determined in an iterative cyclic process. When the curve is changed on each cycle, the curve fitting algorithm associated with the model takes a measurement based on that cycle to ensure that the fitting process is appropriate for the data. Goodness of fit is determined by measuring the residuals in each cycle, with curve fitting algorithms designed to minimize the sum of the residuals squared.

Residuals are key to understanding how the fitting process defines the criteria for performing a best fit for the data, as well as the analysis of outliers. To help understand the concept of residuals, the example below shows data that

has been passed to a fitting process in a basic linear fit, with an X and Y data set. Drawing a straight line through the data points at each cycle or measurement point allows us to draw a line from the measured points to the fitted line and place a dot on the line accordingly.

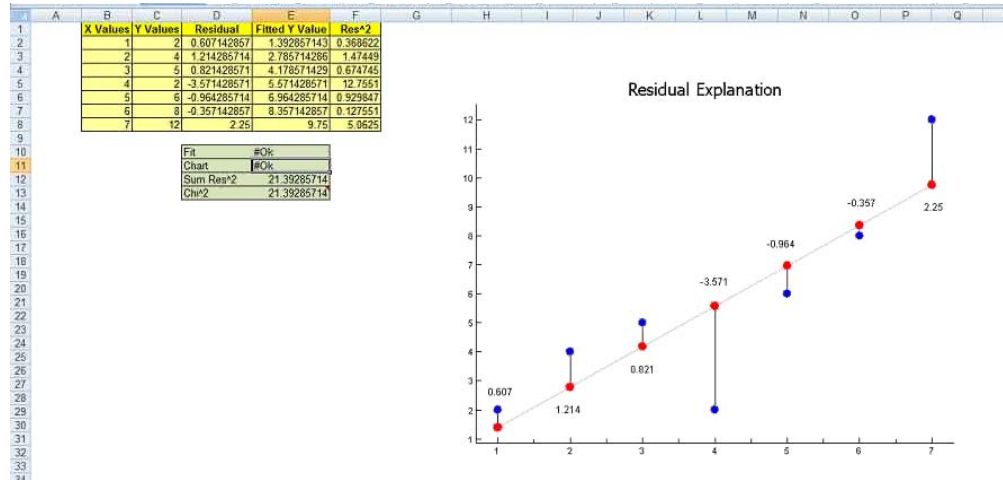


Fig 3: Measuring residuals in XLfit to identify outliers

The blue points represent the Y values for the measured data and the red points the corresponding fitted Y values. The difference between the measured value and the corresponding fitted value is known as a residual, which indicates how far the corresponding point on the fitted line is from the measured data.

If the fitted line is above our measured point, the residual is negative, and if it is below, the residual is positive. When assessing the goodness of fit, these values are always squared in order to eliminate the positive or negative nature of the residual.

Looking at the residual values in Fig 3 above, the point in the middle is -3.571 in value, which is a significant residual indicating an outlier for the data set and is a candidate for being 'knocked out'. Fig 4 below shows the fit once this outlier has been knocked out.

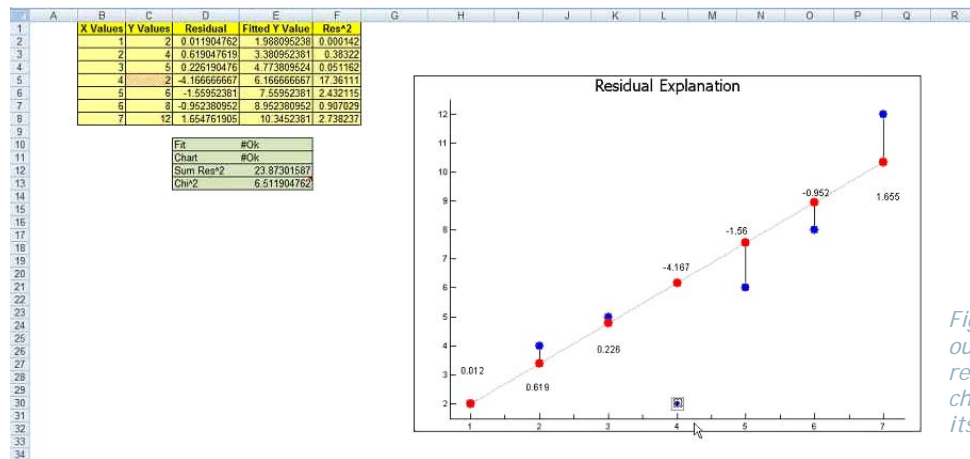


Fig 4: Knocking out outliers in XLfit recalculates residuals and changes the fit to improve its quality

With the data point knocked out, the Chi^2 value (a direct representation of scatter in a data set compared to the fit) has been re-calculated and has dropped from 23.8 to 6.5, generating a significantly better fit because the outlying data point has been removed.

Note: The Chi^2 value grows based on the number of data points provided. In order to get a better feel for the fit, the normalized Chi^2 value is used, which divides the Chi^2 by the degrees of freedom to give a more comparable value.

Summary

If you know what each parameter in a model represents, it is possible to make reasonable estimates for good starting parameter values based on the data set and an understanding of the model. By bringing parameter starting values closer to these estimates, or by employing techniques such as pre-fitting where calculations define suitable starting parameter values for a particular fit, researchers can greatly improve fitting results quality.

Fitting quality or goodness of fit can also be improved by using residuals to identify scatter in a curve and remove potentially erroneous data points, enabling the fit to reach its optimum convergence.

Calculations performed from a best fit perspective use a combination of fitted data and corresponding measured data points. By measuring residuals – the difference between a fitted value and a measured value - at each point of the fitting process, curve fitting applications such as IDBS' XLfit help researchers assess whether or not the fit is converging successfully to produce a best fit.

Squaring the sum of each data point's residual gives an indication of best fit for the data and outliers can be identified based on their residual value – the farther away from the fitted line that a data point is, the more likely it is to be an outlier. Understanding the concept of residuals brings a valuable appreciation as to why fits either fail or succeed.