

Curve Fitting Best Practice

Part 3: Fitting data

Regression and residuals are an important function and feature of curve fitting and should be understood by anyone doing this type of analysis. This article explores regression analysis, describing varying models that can be used to fit data, and the results produced from those particular models.

The fitting process

When performing the fitting process, observational data consists of m values and is commonly termed y values. This 'dependent' variable is subject to error values that are assumed to have a mean of zero. Systematic error may be present but its treatment is outside the scope of regression analysis. The 'independent' variable normally termed x in maths literature is always assumed to be error-free. The independent variables are also called input variables.

Regression

Regression is used to construct a model that allows the user to predict a value of the y variable given a known value of the x variable, or vice-versa. If the prediction is to be done within the range of values of the x variables used to construct the model, this is referred to as interpolation. Prediction outside the range of the data used to construct the model is known as extrapolation and it is more risky.

Regression is used to describe the relationship between two or more variables. It is often taken to be a linear relationship (i.e., a straight line when plotted) but it can also be non-linear (in the shape of a curve). When the regression relationship for the variables is known, we can predict the approximate value of one variable from the value of the other.

Assumptions underpinning regression

Regression fits a model to data by varying the parameters within that model. There are certain assumptions that underpin any type of regression performed, and the four most general are:

- The error terms must be normally distributed and independent, for example, there are no systematic errors in the data set being measured
- All experimental error in measurement is in the y values, with no error terms in the x values
- Data follows the trend of the model, so for example, when using a sigmoidal dose response model, the assumption is that the measured data set follows a sigmoidal response shape and is not, for example, exponential or linear.

Least Squares Fitting

Least squares fitting requires an iterative process to find a best fit based on four different sets of information:

- The measured raw data (normally formed from X and Y pairs but could be more)

- A set of weighting values (this is not a mandatory set of information and will be covered in more detail in the upcoming fifth article in this best practice series, Robust Fitting and Complex Models)
- A model to fit the data set
- A set of starting parameter values that give estimates for the fitting process to begin

Many different techniques for least squares fitting exist but for the most part they all use similar methods and have slightly different strengths and weaknesses. The five most common are:

- Steepest Descent
- Newton
- Gauss-Newton
- Simplex
- Levenberg-Marquardt

Most curve fitting packages will use the Levenberg-Marquardt algorithm which employs a combination of the Gauss-Newton and Steepest Descent methods. Both have advantages and disadvantages depending on whether or not the starting parameter values are close to the minimum (optimum value). The Levenberg-Marquardt method switches between both techniques depending on its point in the fitting process.

Criteria for terminating the fitting process

All least squares fitting methods use an iterative process. A least squares fitting algorithm takes a set of starting parameter values for a given model, and then uses those starting parameter values as a point at which to begin the fitting process. The fitting algorithm then alters each parameter value in an iterative process or set of cycles in order to determine the optimum solution to the problem. By changing the parameters, and so the shape of the curve, the algorithm then measures the differences in the sum of the residuals squared and will look for successive consecutive iterations where the change in the residuals is converging.

Once the convergence limit has been met, the solution is regarded as the optimum and the fitting process ends.

All least squares fitting methods use a number of criteria under which it will terminate the process, known as the iteration limit.

In general the higher the iteration limit and the smaller the convergence limit the greater the accuracy (but slower to perform). The lower the iteration limit and the higher the convergence limit the quicker the fitting (but less accurate). In most cases that use good starting parameters, a fit will always converge quickly on the optimum solution.

Pick the right model up front

The decision on which model to use is dependent on the information that the researcher wants to extract from the curve. It is important to know this information before selecting a model.

Try not to use best fit searching

Best fit searching can be used to give the researcher an idea as to the shape of the data but it is recommended that best fit searching is not used to pick the model from which decisions are to be made. When trying to perform any experimental measurements from a curve, the researcher should know the best model to represent the relationships in the measured data and understand the results that they want to extract from the data. A model is used to understand the results of the data you have. It acts as a way of generalising the relationships between the different data sets that you have and enables you to extract additional information from the data modelled.

Four Parameter Sigmoid Curves

Many four parameter models exist in literature. When looking at dose response data, the most commonly used are:

Standard 4 Parameter Logistic Model

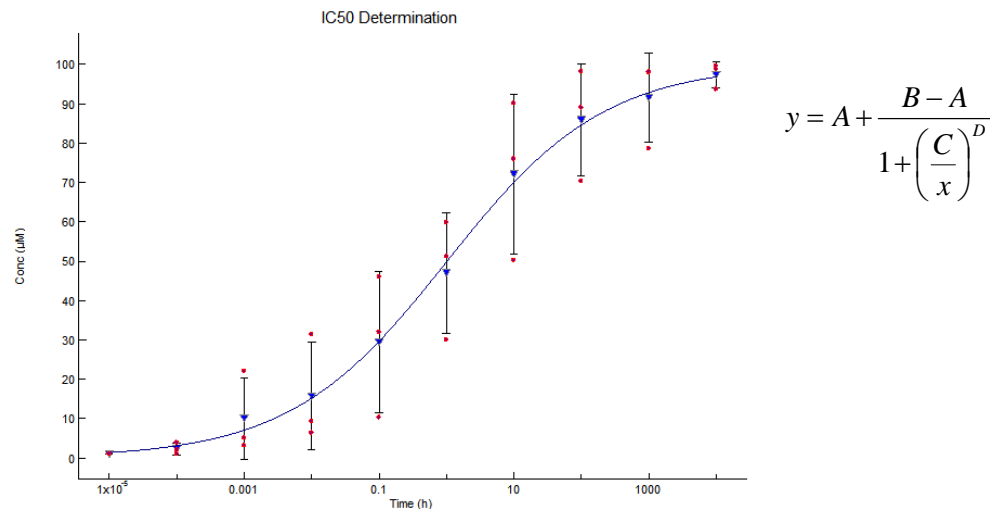


Fig 1: 4-parameter sigmoid where parameter C = EC₅₀

Standard 4 Parameter Logistic Model (Log EC50)

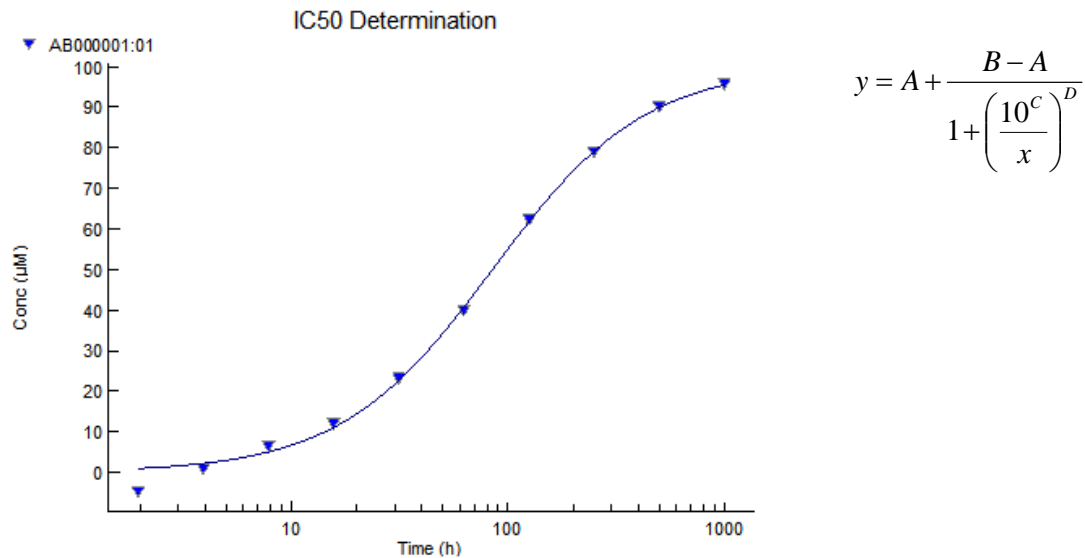


Fig 2: 4-parameter sigmoid where parameter C = Log EC₅₀ value

These two examples are interchangeable and will produce the same results for all parameters except C. Parameter C for the standard logistic model can be converted to the same value as for the Log EC₅₀ model by taking the log of C. However, this can cause issues as the results are actually different when the error values (confidence intervals) are taken into account. For this reason it is always best to use the correct model based on the results required.

One other variation of the two models above is to lock parameter D at 1. This causes the D term to drop out of this function and turns the models into three-parameter logistic curves with fixed slopes.

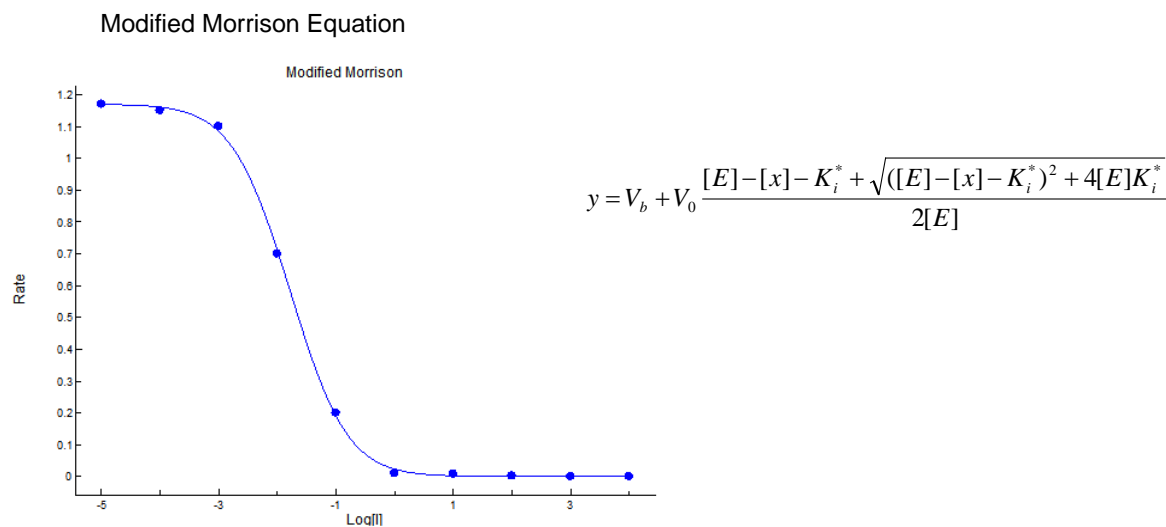


Fig 3: A Modified Morrison model is an alternative to a 4-parameter sigmoid

Three Parameter Logistic Models

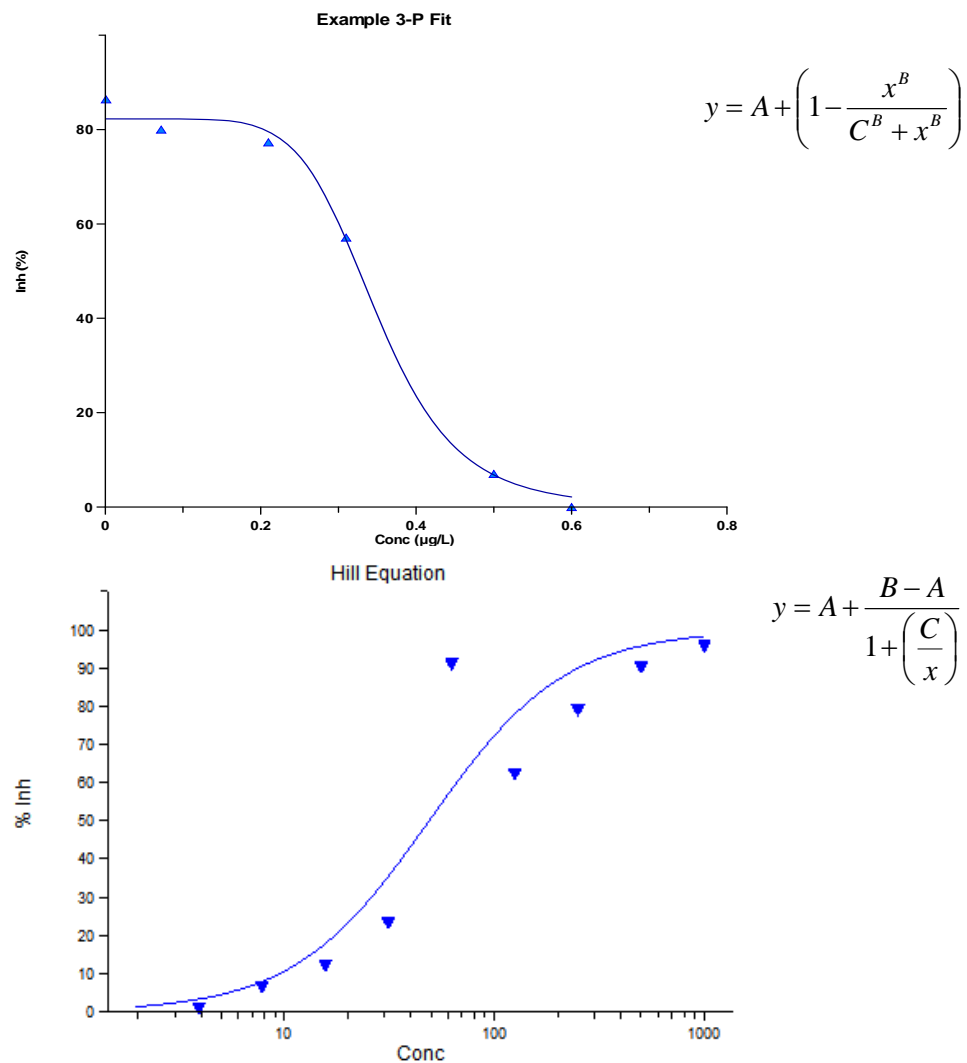


Fig 4: 3-parameter sigmoids where C = EC₅₀ value (top) and Log EC₅₀ value

Five Parameter Logistic Model

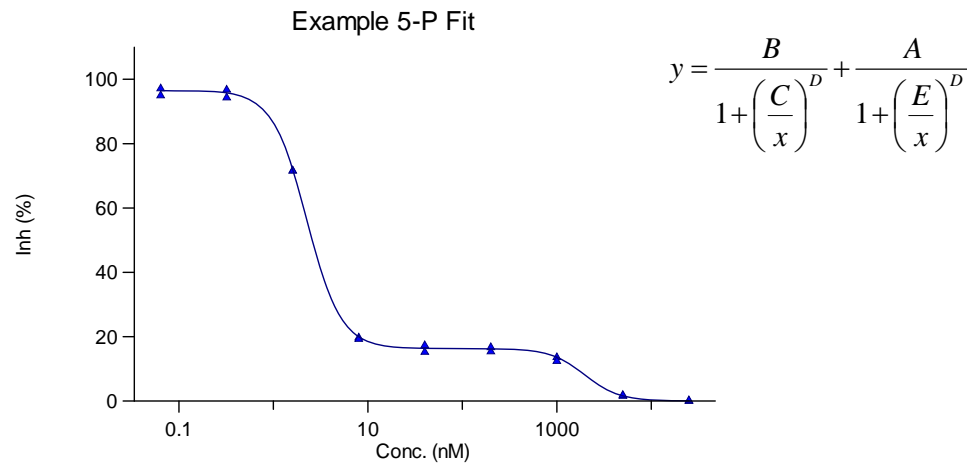


Fig 5: 5-parameter sigmoid where $C = EC_{50}$ curve 1

Summary

In general, there is no single solution for 'best-fit' of a model's parameters to the data provided, as there is in linear regression. Usually numerical optimization algorithms are applied to determine the best-fit parameters using the least squares fitting techniques mentioned earlier. Again in contrast to linear regression, there may be many local minima of the function to be optimized. In practice, estimated values of the parameters are used, in conjunction with the algorithm, to attempt to find the global minimum of a sum of squares.

The appropriate model to choose depends on the analysis to be performed. By changing the model, you change the data representation and the results generated. For each of the results to be measured and extracted from a data set, the researcher should identify the best model with the correct parameterization prior to fitting.