Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
http://www.simfit.org.uk

Curve and surface fitting aims to fit mathematical models described by equations or systems of equations to observations in order to estimate parameters that can be used to interpret the experimental data.

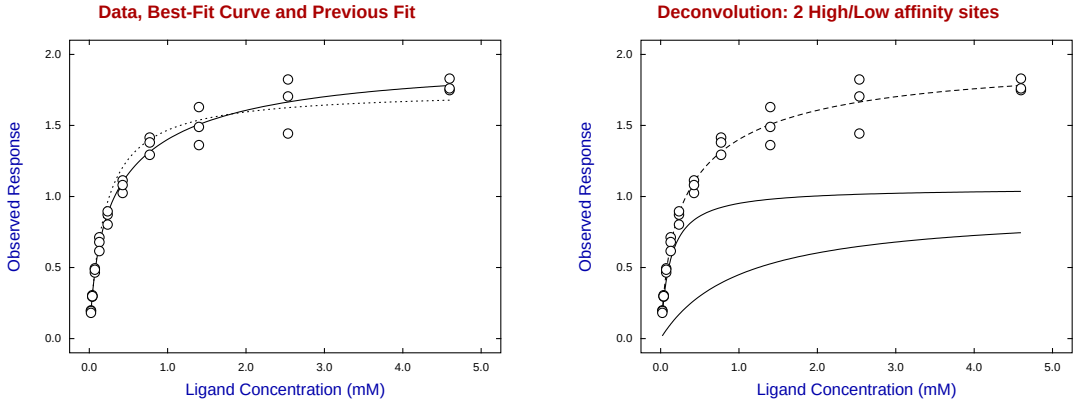
An example to illustrate the advantages of curve fitting

As a typical example consider the results from using SimFIT program **hlfit** to fit the dose response data in test file `hlfit.tf4` to first one binding site and then to two binding sites, where the aim is to decide if the assumption of two sites can be justified on statistical grounds and, if so, to estimate the parameters.

The model fitted is defined as

$$f(x) = \frac{A_1 x}{1 + Ka_1 x} + \frac{A_2 x}{1 + Ka_2 x} + C$$

but with $C = 0$, and **hlfit** displays the following plots to illustrate that fitting two sites gives significant improvement over fitting one site, and **hlfit** also provides convincing evidence of this by showing how the overall fit results as the sum of the distinct contributions from the low and high affinity sites.



Not only does the evidence support the hypothesis that there are two classes of binding sites of differing affinity and activity, but this is substantiated by the following results table for estimated parameters, their standard errors, 95% confidence ranges, and significance levels.

For the best-fit 2:2 High-Low affinity sites model using program HLFIT

Number	Parameter	Value	Std.error	Lower95%cl	Upper95%cl	p	
1	A_1	0.91175	0.2451	0.4079	1.416	0.0010	
2	A_2	1.0625	0.3055	0.4344	1.691	0.0018	
3	Ka_1	0.97501	0.6857	-0.4345	2.385	0.1669	*
4	Ka_2	8.5829	2.004	4.463	12.70	0.0002	

Apparent V_{max} (i.e. $A_1 + A_2 + \cdots + A_n$) = 1.9742
Apparent K_m (i.e. x_0 where $f(x_0) - C = V_{max}/2$) = 0.31272

Here parameters A_1 and A_2 are proportional to the responses from two populations of binding sites with binding constants Ka_1 and Ka_2, and the apparent overall response and half saturation point are calculated by numerical techniques, which removes the subjective element involved in data interpretation.

A brief survey of the nomenclature used and procedures provided by SimFIT follows.

The Data

In the simplest case an experimentalist would have N pairs of observations $x(i)$, $y(i)$, possibly together with $s(i)$, the estimated standard deviations of $y(i)$ to use as weights $w(i) = 1/s(i)^2$, as follows

$$X = x(1), x(2), x(3), \ldots, x(N)$$
$$Y = y(1), y(2), x(3), \ldots, y(N)$$
$$S = s(1), s(2), s(3), \ldots, s(N)$$

and there could be three possibilities.

1. **Case 1**
   Values of $x(i)$ are known with high accuracy, as fixed by experiment, and the error distribution of $y(i)$ is assumed to be one of constant variance. In this case it is usual to set all $s(i) = 1$.

2. **Case 2**
   Values of $x(i)$ are known with high accuracy, as fixed by experiment, and the error distribution of $y(i)$ is assumed to vary as a function of the experimental conditions, so values for $s(i)$ are required.

3. **Case 3**
   Values of $x(i)$ and $y(i)$ would both be measured, i.e. there could be error or variation in X and Y.

Of course there are endless variations on this simple scheme, for instance, X and/or Y could be multidimensional, and the model might have to be defined as an implicit function, $\Phi(x, y) = 0$, or require numerical integration of a system of nonlinear differential equations.

The weighting

It is important to realize that all curve fitting is actually weighted curve fitting. The only issue is whether the weighting is assumed to have a defined form, to be estimated from the sample, or to be estimated independently.

- **Case 1**
  This is the simplest and most used technique because it assumes that X is an independent variable and Y values result from a random error ϵ with constant variance added to an exact function value, i.e.

  $$y(i) = f(x(i)) + \epsilon(i).$$

  With this approach no separate attempt is made to estimate the variance of Y as the sample variance of Y is used. It has the great attraction of simplicity but there are two things to observe:
  a) This assumption is almost never true as, in general, error variance is an increasing function $|Y|$.
  b) It diminishes the importance of low $|Y|$ *values so that the resulting fit is dominated by large* $|Y|$ *values.*

- **Case 2**
  This is more realistic in that it accepts that the variance of Y is not constant and attempts to remedy this by providing or calculating a set of weighting factors $s(i)$. However, if the $s(i)$ are inaccurate, the resulting fit can be even more biased than with Case 1. In particular, using a weighting scheme based simply on the experimental Y values or the estimated function values can lead to a situation the reverse of Case 1, where the fit can be dominated by small values of the observations.

- **Case 3**
  This requires that the variance-covariance matrix $CV(X, Y)$ be estimated, and sample estimates for variance-covariance matrices are notoriously unreliable. For this reason this approach is often reserved for the analysis of very large samples with very simple model equations, together with prior knowledge of the covariance structure.

The model

The model to be fitted will involve parameters that have to be estimated, e.g. a parameter vector Θ as in

$$\Theta = \theta_1, \theta_2, \theta_3, \ldots, \theta_m$$

and models are usually described as linear or nonlinear.

Linear models

Linear models are of the form $f(\Theta, x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \theta_3 f_3(x) + \cdots + \theta_m f_m(x)$ and examples could be

$$\begin{aligned} \text{A simple straight line:} \quad & f(\theta, x) = \theta_1 + \theta_2 x \\ \text{A multilinear model:} \quad & f(\Theta, x) = \theta_1 x_1 + \theta_2 x_2 + \theta_3 x \cdots + \theta_m x_m \\ \text{A polynomial:} \quad & f(\Theta, x) = \theta_1 + \theta_2 x + \theta_3 x^2 + \cdots + \theta_m x^{m-1}. \end{aligned}$$

These all have partial derivatives of $f(\Theta, x)$ with respect to θ_i that are independent of Θ.

The advantage of linear models is that they can be fitted very easily and usually lead to unique solutions. Another advantage is that the assumption of normally distributed errors zero means and constant variance allows the application of convenient statistical tests for goodness of fit and parameter significance based on the χ^2, t, and F distributions. The disadvantage is that the real world is nonlinear and linear models are not based on scientific laws but are used for convenience when a meaningful mathematical model is not available.

Nonlinear models

Nonlinear models do not have have partial derivatives of $f(\Theta, x)$ with respect to θ_i that are independent of Θ. Examples could be the following.

$$\begin{aligned} \text{Michaelis-Menten functions:} \quad & f(\Theta, x) = \frac{\theta_1 x}{\theta_2 + x} + \frac{\theta_3 x}{\theta_4 + x} + \cdots + \frac{\theta_{m-1} x}{\theta_m + x} \\ \text{Exponential functions:} \quad & f(\Theta, x) = \theta_1 \exp(-\theta_2 x) + \theta_3 \exp(-\theta_4 x) + \cdots + \theta_{m-1} \exp(-\theta_m x) \\ \text{Rational functions:} \quad & f(\Theta, x) = \frac{\theta_1 x + \theta_2 x^2 + \cdots + \theta_{m/2} x^{m/2}}{1 + \theta_{m/2+1} x + \theta_{m/2+2} x^2 + \cdots + \theta_m x^{m/2}}. \end{aligned}$$

The advantage of using nonlinear models is that they may be a good approximation to reality based on scientific laws. The disadvantage is that they have to be fitted by iterative techniques which depend critically on sensible scaling, good starting estimates, and meaningful limits on parameter values. Because of this, local rather than global solutions may be located.

However it is well to remember a distinct limitation of nonlinear regression: all the numerical techniques used to fit the models, and all the statistical methods used to interpret the results, are based on the assumptions that the model can be regarded as approximately linear at a solution point, and that the weighting factors are known exactly.

The objective function

This will usually be $WSSQ$, the weighted sum of squared residuals defined as

$$WSSQ = \sum_{i=1}^{N} w(i)[y(i) - f(\Theta, x(i)]^2,$$

and the hope is that minimizing this expression with respect to parameters Θ will be equivalent to finding the maximum likelihood estimates.

Options

The SIMFIT package provides many options to prepare data, define models, fit data, and test for model discrimination, goodness of fit, and parameter redundancy as briefly summarized below.

- **Programs for linear models**

  Simple linear models can be fitted by program **linfit** which also provides techniques for orthogonal regression, generalized linear models (GLM), and partial least squares (PLS). Polynomials can be fitted by program **polnom**, which also allows inverse prediction, i.e. generating calibration curves. Several varieties of cubic splines can be fitted by program **spline**, and used to compare curves by **compare**, while program **calcurve** is dedicated to using splines to construct calibration curves followed by inverse prediction of x given y.

- **Programs for simple nonlinear regression**

  The following programs attempt to guess starting estimates then fit models and output tables of statistical results and graphs.

  - **mmfit** fits Michaelis-Menten models.
  - **hlfit** fits high and low affinity binding site models.
  - **sffit** fits cooperative ligand binding isotherms.
  - **rffit** fits positive rational functions.
  - **gcfit** fits classical nonlinear growth models.

- **Advanced nonlinear regression**

  Program **qnfit** provides the following facilities but, as it is extremely comprehensive, it requires considerable expertise and should only be used by experienced analysis.

  - Models can be functions of one or several independent variables.
  - Multiple linked or independent models can be fitted simultaneously.
  - Parameters can be constrained interactively within user-defined limits.
  - Three dimensional plots and contours of the objective function can be plotted at solution points.
  - The best fit models can be used for evaluation or inverse prediction.
  - Models can be used from a built-in library or supplied as text files.
  - Single nonlinear differential equations can be fitted.
  - Models defined as convolutions of two defined functions can be fitted.

- **Differential equations**

  Program **deqsol** allows the simulation and fitting of systems of nonlinear differential equations but, like **qnfit**, it should only be used by experts.

- **Simulation**

  An essential technique required for advanced curve fitting is the ability to simulate exact data using program **makdat** then add random error to simulate reality using program **adderr**. SIMFIT also provides numerous additional facilities to confirm the robustness of results from regression with respect to sensitivity of the results to perturbations of parameter values, change in the range of variables, nature of the error, etc.

How to interpret tables of parameter estimates

The meaning of the results generated by program **exfit** after fitting two exponentials to `exfit.tf4` will now be explained, as a similar type of analysis is generated by all the user-friendly curve fitting programs. Consider, first of all the next table listing parameter estimates which result from fitting the two exponential function

$$f(t) = A_1 \exp(-k_1 t) + A_2 \exp(-k_2 t).$$

Parameter	Value	Std.error	Lower95%cl	Upper95%cl	p
A_1	0.8526	0.0677	0.713	0.992	0.0000
A_2	1.1764	0.0747	1.023	1.330	0.0000
k_1	6.7933	0.8541	5.038	8.549	0.0000
k_2	1.1121	0.0511	1.007	1.217	0.0000
AUC	1.1834	0.0147	1.153	1.214	0.0000

AUC is the area under the curve from $t = 0$ to $t = \infty$
Initial time point $(A) = 0.03598$
Final time point $(B) = 1.611$
Area from $t = A$ to $t = B = 0.9383$
Average over range $(A, B) = 0.5958$

The first column gives the estimated values for the parameters, i.e., the amplitudes A_i and decay constants k_i, although it must be appreciated that the pairwise order of these is arbitrary. Actually program **exfit** will always try to rearrange the output so that the amplitudes are in increasing order, and a similar rearrangement will also occur with programs **mmfit** and **hlfit**. For situations where $A_i > 0$ and $k_i > 0$, the area from zero to infinity, i.e. the AUC, can be estimated, as can the area under the data range and the average function value calculated from it. The parameter AUC is not estimated directly from the data, but is a secondary parameter estimated algebraically from the primary parameters. The standard errors of the primary parameters are obtained from the inverse of the estimated Hessian matrix at the solution point, but the standard error of the AUC is estimated from the partial derivatives of AUC with respect to the primary parameters, along with the estimated variance-covariance matrix. The 95% confidence limits are calculated from the parameter estimates and the t distribution, while the p values are the two-tail probabilities for the estimates, i.e., the probabilities that parameters as extreme or more extreme than the estimated ones could have resulted if the true parameter values were zero. The windows defined by the confidence limits are useful for a quick rule of thumb comparison with windows from fitting the same model to another data set; if the windows are disjoint then the corresponding parameters differ significantly, although there are more meaningful tests. Clearly, parameters with $p < 0.05$ are well defined, while parameters with $p > 0.05$ must be regarded as ill-determined.

Expert users may sometimes need the estimated correlation matrix

$$C_{ij} = \frac{CV_{i,j}}{\sqrt{CV_{ii} CV_{jj}}},$$

where $-1 \le C_{ij} \le 1$, $C_{ii} = 1$, which is shown in the next table, and where the index i refers to the natural order of parameters, that is $i = 1, 2, 3, 4$ corresponds to A_1, k_1, A_2, k_2.

Parameter correlation matrix

1			
-0.8756	1		
-0.5961	0.8995	1	
-0.8478	0.9485	0.8199	1

How to interpret tables for goodness of fit

The next table, displaying the results from analyzing the residuals after fitting two exponentials to `exfit.tf4`, is typical of many SimFIT programs. Residuals tables should always be consulted when assessing goodness of fit.

Analysis of residuals: $WSSQ$	24.397
$P(\chi^2 \geq WSSQ)$	0.5533
R^2, cc(theory,data)^2	0.9934
Largest absolute relative residual	11.99%
Smallest absolute relative residual	0.52%
Average absolute relative residual	3.87%
Absolute relative residuals in range 0.1-0.2	3.33%
Absolute relative residuals in range 0.2-0.4	0.00%
Absolute relative residuals in range 0.4-0.8	0.00%
Absolute relative residuals > 0.8	0.00%
Number of negative residuals (n_1)	15
Number of positive residuals (n_2)	15
Number of runs observed (r)	16
P(runs $\leq r$: given n_1 and n_2)	0.5759
5% lower tail point	11
1% lower tail point	9
P(runs $\leq r$: given n_1 plus n_2)	0.6445
P(signs $\leq$ least number observed)	1.000
Durbin-Watson test statistic	1.8061
Shapiro-Wilks W statistic	0.9387
Significance level of W	0.0841
Akaike AIC (Schwarz SC) statistics	1.7979 (7.4027)

Verdict on goodness of fit: incredible

Several points should be remembered when assessing such residuals tables, where there are N observations $y(i)$, with weighting factors $s(i)$, theoretical values $f(x(i))$, residuals $r(i) = y(i) - f(x(i))$, weighted residuals $r(i)/s(i)$, and where m parameters have been estimated. Theoretical details for the statistical tests will be found in the SimFIT reference manual `w_manual.pdf`, or the appropriate tutorial documents.

- $WSSQ$
  The χ^2 test on $N - m$ degrees of freedom using $WSSQ$, the objective function at the solution point where
  $$WSSQ = \sum_{i=1}^{N} \left(\frac{y(i) - f(x_i)}{s(i)} \right)^2,$$
  is only meaningful if the weights defined by the $s(i)$ supplied for fitting are good estimates of the standard deviations of the observations at that level of the independent variable; say means of at least five replicates. Inappropriate weighting factors will result in a biased chi-square test. Also, if all the $s(i)$ are set equal to 1, unweighted regression will be performed and an alternative analysis test based on the coefficient of variation will be performed.

- R^2
  The R^2 value is the square of the correlation coefficient between data and best fit points. It only represents a meaningful estimate of that proportion of the fit explained by the regression for simple unweighted linear models, and should be interpreted with restraint when nonlinear models have been fitted.

- **Absolute relative residuals**
  The results based on the absolute relative residuals $a(i)$ defined using machine precision ϵ as

$$a_i = \frac{2|r(i)|}{\max(\epsilon, |y(i)| + |f(x(i))|)}$$

  do not have statistical relevance, but they do have obvious empirical justification, and they must be interpreted with commonsense, especially where the data and/or theoretical values are very small.

- **Run and sign tests**
  The probability of the number of runs observed given n_1 negative and n_2 positive residuals is a very useful test for randomly distributed runs, but the probability of runs given $N = n_1 + n_2$, and also the overall sign test are weak, except for very large data sets.

- **Durbin-Watson test**
  The Durbin-Watson test statistic

$$DW = \frac{\sum_{i=1}^{N-1} (r(i+1) - r(i)^2}{\sum_{i=1}^{N} r(i)^2}$$

  is useful for detecting serially correlated residuals, which could indicate correlated data or an inappropriate model. The expected value is 2.0, and values less than 1.5 suggest positive correlation, while values greater than 2.5 suggest negative serial correlation.

- **Shapiro-Wilks test**
  Where N, the number of data points, significantly exceeds m, the number of parameters estimated, the weighted residuals are approximately normally distributed, and so the Shapiro-Wilks test should be taken seriously.

- **Akaike and Schwarz criteria**
  The Akaike AIC statistic

$$AIC = N \log(WSSQ/N) + 2m$$

  and Schwarz Bayesian criterion SC

$$SC = N \log(WSSQ/N) + m \log N$$

  are only really meaningful if minimizing $WSSQ$ is equivalent to Maximum Likelihood Estimation. Note that only differences between AIC with the same data, i.e. fixed N, are relevant, as in the evidence ratio ER, defined as $ER = \exp[(AIC(1) - AIC(2))/2]$.

- **The qualitative conclusion**
  The final verdict is calculated from an empirical look-up table, where the position in the table is a weighted mean of scores allocated for each of the tests listed above. It is qualitative and rather conservative, and has no precise statistical relevance, but a good result will usually indicate a well-fitting model.

- **Residuals plots**
  As an additional measure, plots of residuals against theory, and half-normal residuals plots can be displayed after such residuals analysis, and they should always be inspected before concluding that any model fits satisfactorily.

- **Leverages**
  With linear models, SimFIT also calculates studentized residuals and leverages, while with generalized linear models, deviance residuals can be tabulated.

How to interpret tables for model discrimination results

After a sequence of models have been fitted, tables like the next one are generated.

$WSSQ$-previous	224.9
$WSSQ$-current	24.4
Number of parameters-previous	2
Number of parameters-current	4
Number of x-values	30
Akaike AIC-previous	64.44
Akaike AIC-current	1.798, $ER = 3.998E + 13$
Schwarz SC-previous	67.24
Schwarz SC-current	7.403
Mallows C_p	213.7, $C_p/m_1 = 106.9$
Numerator degrees of freedom	2
Denominator degrees of freedom	26
F test statistic (FS)	106.9
$P(F \geq FS)$	0.0000
$P(F \leq FS)$	1.0000
5% upper tail point	3.369
1% upper tail point	5.526

Conclusion based on the F test
Reject the previous model at 1% significance level
There is strong support for the extra parameters
Tentatively accept the current best fit model

First of all, note that the above model discrimination analysis is only strictly applicable for nested linear models with known error structure, and should be interpreted with restraint otherwise. Now, if $WSSQ_1$ with m_1 parameters is the previous (possibly deficient) model, while $WSSQ_2$ with m_2 parameters is the current (possibly superior) model, so that $WSSQ_1 > WSSQ_2$, and $m_1 < m_2$, then

$$F = \frac{(WSSQ_1 - WSSQ_2)/(m_2 - m_1)}{WSSQ_2/(N - m_2)}$$

should be F distributed with $m_2 - m_1$ and $N - m_2$ degrees of freedom, and the F test for excess variance can be used. Alternatively, if $WSSQ2/(N - m_2)$ is equivalent to the true variance, i.e., model 2 is equivalent to the true model, the Mallows C_p statistic

$$C_p = \frac{WSSQ_1}{WSSQ_2/(N - m_2)} - (N - 2m_1)$$

can be considered. This has expectation m_1 if the previous model is sufficient, so values greater than m_1, that is $C_p/m_1 > 1$, indicate that the current model should be preferred over the previous one. However, graphical deconvolution should always be done wherever possible, as with sums of exponentials, Michaelis-Mentens, High-Low affinity sites, sums of Gaussians or trigonometric functions, etc., before concluding that a higher order model is justified on statistical grounds.