Simfit

Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.
https://simfit.org.uk
*https://simfit.silverfrost.com*

Goodness of fit analysis is always required before a model can be considered to be satisfactory, otherwise using parameter estimates to characterize experimental observations can lead to false interpretation.

As an example consider the use of program **gcfit** to fit nonlinear growth models to data in the test file gcfit.tf2 as shown in the next figure. A typical situation would be when an experimentalist would want to fit growth curves to data with the main aim being to estimate parameters like the maximum growth rate, the time at which this was achieved, and the final size attained in order to characterize a group under observation, say bacterial colonies of several species incubated with alternative antibiotics.

Fitting Alternative Growth Models
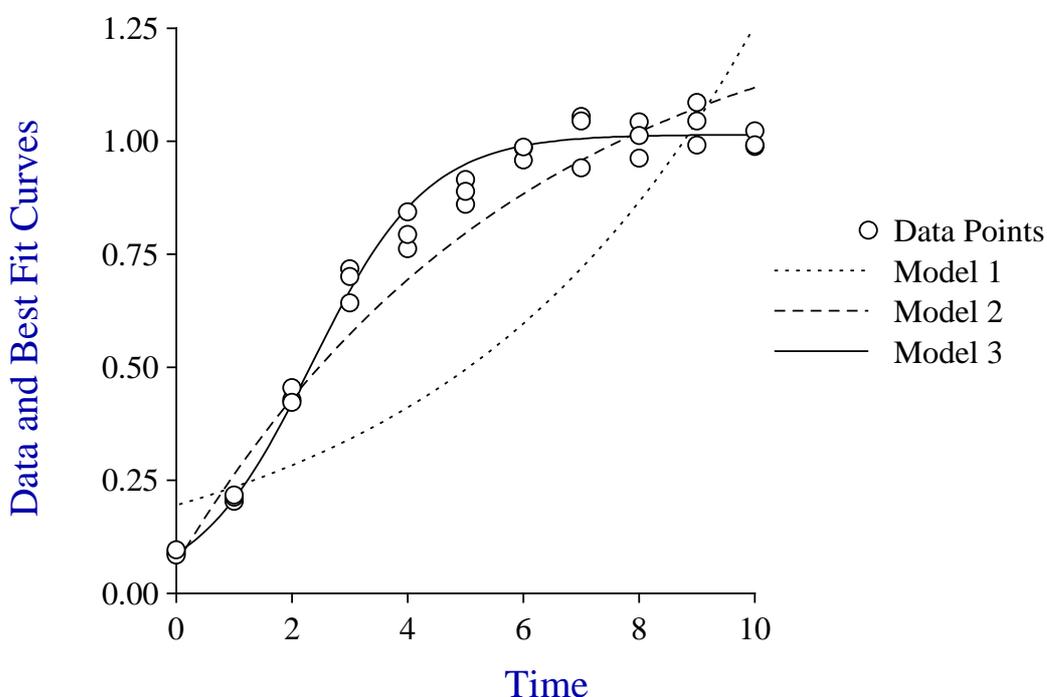
The models fitted were Model 1 (exponential), Model 2 (monomolecular) and Model 3 (logistic) as follows.

$$\text{Model 1: } f_1(t) = A_1 \exp(k_1 t)$$
$$\text{Model 2: } f_2(t) = A_2(1 - \exp(-k_2 t))$$
$$\text{Model 3: } f_3(t) = \frac{A_3}{1 + B \exp(-k_3 t)}$$

It is perfectly clear in this case that Model 1 is completely unsatisfactory, Model 2 would give a rough estimate for the final asymptotic size, while Model 3 would accurately fit all features of the data set. SimFIT program **gcfit** also displays the following summary.

Model	$WSSQ/NDOF$	$P(C \geq W)$	$P(\text{Runs} \leq r)$	$N > 10\%$	$N > 40\%$	$Av.r\%$	Verdict
1	1.52E+02	0.000	0.000	29	17	40.03	Very bad
2	1.81E+01	0.000	0.075	20	0	12.05	Very poor
3	1.32E+00	0.113	0.500	0	0	3.83	Incredible

The way to interpret this goodness of fit summary table will now explained, but it must always be remembered that the only situations where statistical tests are justified, and when minimizing $WSSQ$ is equivalent to maxmimum likelihood, are when the following four conditions are met.

1. The values $x(i)$ must be known exactly, and not subject to errors of estimation or natural variation. In other words, X can be regarded as an independent variable and not a covariate.

2. The error of measurement $\epsilon(i)$ must be normally distributed with mean zero.

3. The variance of $\epsilon(i)$ has one of two forms.

    (a) The homoscedastic case where weighting factors $s(i)$ are all equal to one and $WSSQ/NDOF$ estimates the constant variance.

    (b) The heteroscedastic case where the variance of $\epsilon(i)$ is a function of X and/or Y and exact values for the standard deviation of the $\epsilon(i)$ are supplied as $s(i)$. In other words, values of $s(i)$ are supplied to reduce this case to the homoscedastic case with error variance = 1.

4. The model is correct and linear.

As experimental errors are more like a Cauchy distribution than a normal distribution, variance of the experimental error is usually an increasing function of the absolute value of the observations, values of $s(i)$ supplied are at best only determined with limited precision from independent studies or at worst are determined from replicates, and the model is nonlinear and often only an approximation anyway, such results tables must be interpreted with restraint.

- *$WSSQ/NDOF$*
  This is the objective function estimated by SimFIT and $WSSQ/NDOF$ should be approximately one at the solution point, as the expectation of a chi-square variable is the number of degrees of freedom.

- *$P(C \geq W)$*
  This is the very approximate result of a performing a χ^2 test on the weighted sum of squared residuals. An alternative test is usually done by SimFIT based on the estimated coefficient of variation here and for the previous result when all $s(i) = 1$.

- *$P(Runs \leq r)$*
  This is the probability of runs less than the number obtained, given the number of negative and positive residuals.

- *$N > 10\%$*
  This is the number of data points where the ratio of absolute residual to absolute value of observation exceeds 0.1.

- *$N > 40\%$*
  This is the number of data points where the ratio of absolute residual to absolute value of observation exceeds 0.4.

- *$Av.r\%$*
  This is the average of absolute residual divided by absolute observation as a percentage.

- Verdict
  This is a somewhat arbitrary decision based on a formula involving all of these, and also some other factors.

Analysis of residuals

Analysis of residuals and/or weighted residuals is a very important way to judge goodness of fit, especially when there is only one independent variable, and SіMFιT provides numerous ways to do this as follows.

1. Tables of residuals
   These highlight residuals which indicate poor fit by colour changes and stars.

2. Tables summarizing goodness of fit based on residuals

3. Test for runs and serial correlations
   These rely on the residuals being in a systematic order, such as in order of the independent variable.

4. Test for a normal distribution
   Residuals cannot be normally distributed due to correlations induced by parameter estimation, nevertheless the Shapiro-Wilks test is quiet robust when the number of observations greatly exceeds the number of parameters estimated.

5. Methods for plotting residuals

   (a) Residuals plotted against the independent variable
   (b) Residuals plotted against the observations
   (c) Residuals plotted against the best-fit model
   (d) Normal and half-normal plots

   Probably option (a) is the easiest to interpret and residuals have to deviate wildly from normality before option (d) picks this up. Unfortunately this is the only option available when there are multiple independent variables.

Residuals and/or weighted residuals should be scattered randomly about zero, and the next plot shows very clearly that with Model 1 there is a systematic nonlinear drift which is much less with model 2, while model 3 shows a much more acceptable pattern

Residuals from GCFIT with Models 1, 2, and 3
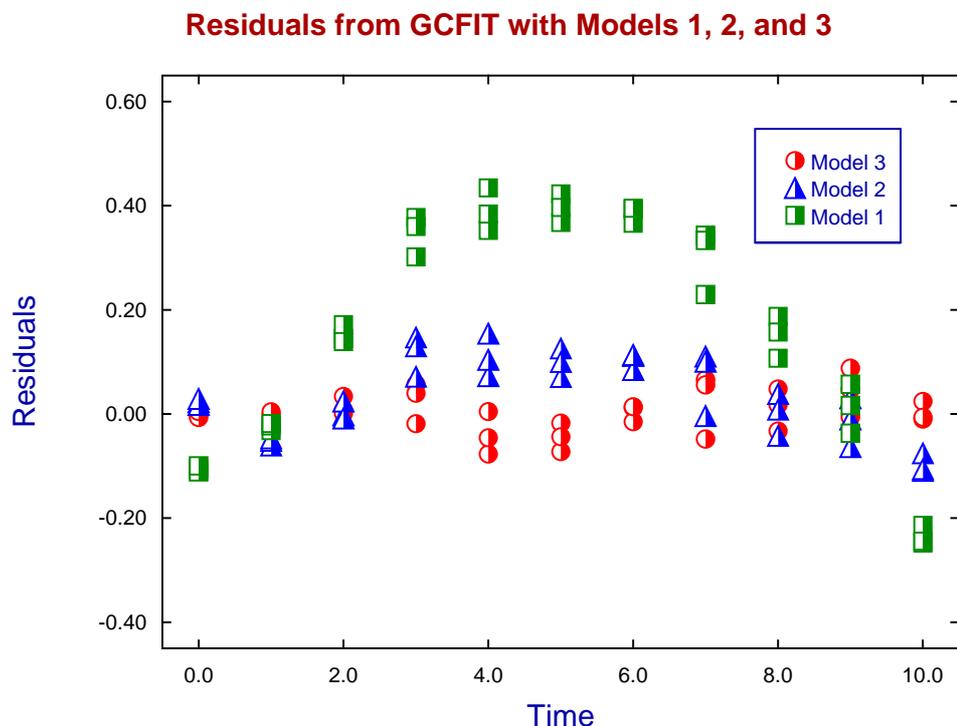
A plot of the n ordered residuals or weighted residuals as Y against normal order statistic medians as X should be very close to linearity, since residuals should be approximately normally distributed when the number of points is much greater than the number of parameters estimated. The medians are approximated using

$$x_i = \Phi^{-1}(z_i) \text{ where } z_i = (i - 3/8)/(n + 1/4) \text{ for } i = 1, 2, \cdots, n.$$

Here Φ^{-1} is the inverse standard normal distribution function. It is also possible to create a half-normal plot where Y are the ordered absolute residuals, and X values are calculated by a similar approximation but using $z_i = (n + 0.5 + i)/(2n + 9/8)$ to allow for the wrapping round of the negative residuals. If correctly weighted residuals are plotted, as in the next figure for the fitted logistic model, the Y values should be in the range -3 to 3 for the normal plot, but 0 to 3 for the half normal plot.

Normal Plot: r = 0.9906

Half-Normal Plot: r = 0.9911
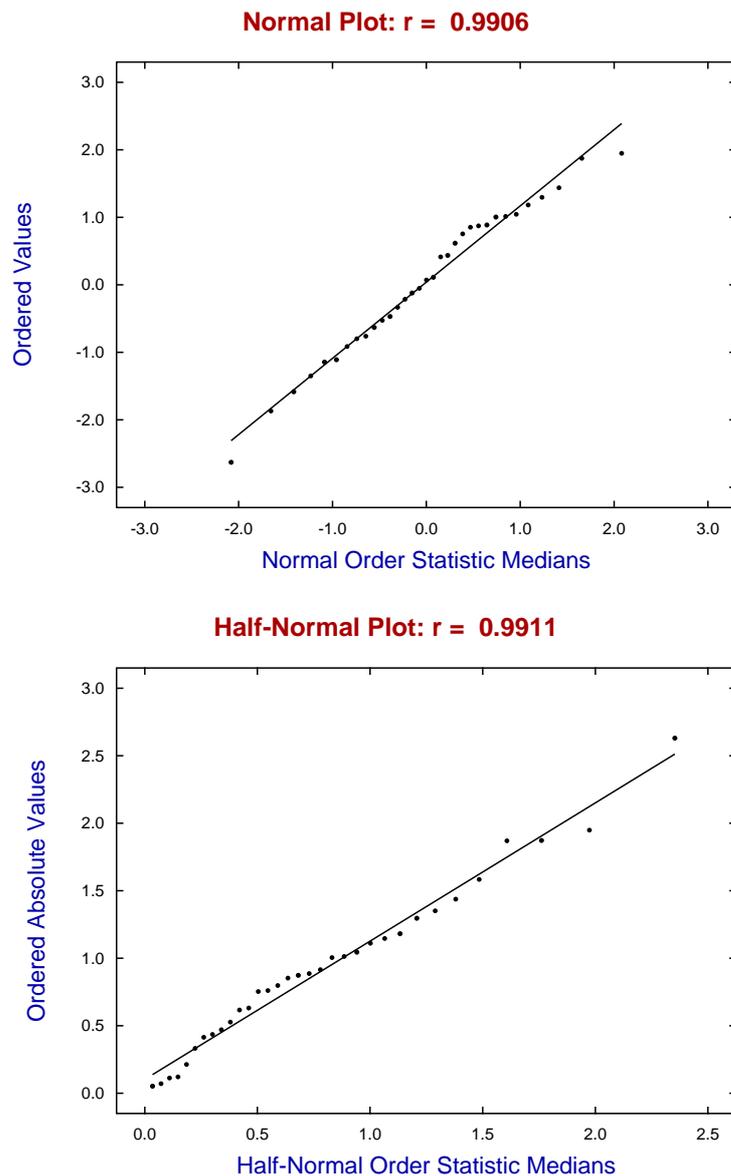
Best-fit lines for the regression of Y on X are also plotted on these graphs along with the Pearson product-moment correlation coefficient r. The significance level p for the r will also be displayed when $p < 0.05$, but this only happens when the residuals show clear departure from linearity in these plots.

Testing for differences between two parameter estimates

This can sometimes be a useful simple procedure when you wish to compare two parameters resulting from a regression, e.g., the final size from fitting a growth curve model, or perhaps two parameters that have been derived from regression parameters e.g., AUC from fitting an exponential model, or LD50 from bioassay.

You input the two parameter estimates θ and ϕ, the standard error estimates s_θ and s_ϕ, the number of experimental observations N_θ and N_ϕ, and the number of parameters estimated from the regression M_θ and M_ϕ. A t test for equality is then performed with the correction for unequal variances by the the Satterthwaite procedure, using a t_c statistic with ν degrees of freedom calculated with the Welch correction for unequal variances given by

$$t_c = \frac{\theta - \phi}{\sqrt{s_\theta^2 + s_\phi^2}}$$

$$\nu = \frac{(s_\theta^2 + s_\phi^2)^2}{s_\theta^4/(N_\theta - M_\theta) + s_\phi^4/(N_\phi - M_\phi)}.$$

Here θ and ϕ refer to the same parameter using the same mathematical model but estimated from two distinct data sets of sizes N_θ and N_ϕ.

Such t tests depend on the asymptotic normality of maximum likelihood parameters, and will only be meaningful if the data set is fairly large and the best fit model adequately represents the data.

Note that t tests on parameter estimates can be especially unreliable because they ignore non-zero covariances in the estimated parameter variance-covariance matrix.

Testing for differences between several parameter estimates

To take some account of the effect of significant off-diagonal terms in the estimated parameter variance-covariance matrix you will need to calculate a Mahalanobis distance between parameter estimates e.g., to test if two or more curve fits using the same model but with different data sets support the presence of significant treatment effects. For instance, after fitting the logistic equation to growth data by nonlinear regression, you may wish to see if the growth rates, final asymptotic size, half-time, etc. have been affected by the treatment.

Note that, after every curve fit, you can select an option to add the current parameters and covariance matrix to your parameter covariance matrix project archive, and also you have the opportunity to select previous fits to compare with the current fit. For instance, you may wish to compare two fits with m parameters, A in the first set with estimated covariance matrix C_A and B in the second set with estimated covariance matrix C_B. The parameter comparison procedure will then perform a t test for each pair of parameters, and also calculate the quadratic form

$$Q = (A - B)^T (C_A + C_B)^{-1}(A - B)$$

which has an approximate chi-square distribution with m degrees of freedom. You should realize that the rule of thumb test using non-overlapping confidence regions is more conservative than the above t test: parameters can still be significantly different despite a small overlap of confidence windows.

This technique must be used with care when the models fitted are themselves sums of k identical sub-functions such as

$$f(\Theta x) = f_1(\Theta, x) + f_2(\Theta, x) + \cdots + f_k(\Theta, x).$$

Examples of where this can occur could be sums of exponentials, Michaelis-Menten terms, High-Low affinity site binding isotherms, Gaussians, trigonometric terms, and so on. This is because the parameters are only unique up to a permutation.

For instance, the terms A_i and k_i are linked in the exponential function

$$f(t) = \sum_{i=1}^{m} A_i \exp(-k_i t)$$

but the order implied by the index i is arbitrary. So, when testing if A_1 from fitting a data set is the same as A_1 from fitting another data set it is imperative to compare the same terms.

The user friendly programs **exfit**, **mmfit**, and **hlfit** attempt to assist this testing procedure by rearranging the results into increasing order of amplitudes A_i but, to be sure, it is best to use **qnfit**, where starting estimates and parameter constraints can be used from a parameter limits file. That way there is a better chance that parameters and covariance matrices saved to project archives for retrospective testing for equality of parameters will be consistent, i.e. the parameters will be compared in the correct order.

The next figure illustrates a common problem, where the same model has been fitted to alternative data sets and it is wished to decide if one or more parameters differ significantly.

Comparing Parameter Estimates for Logistic Models
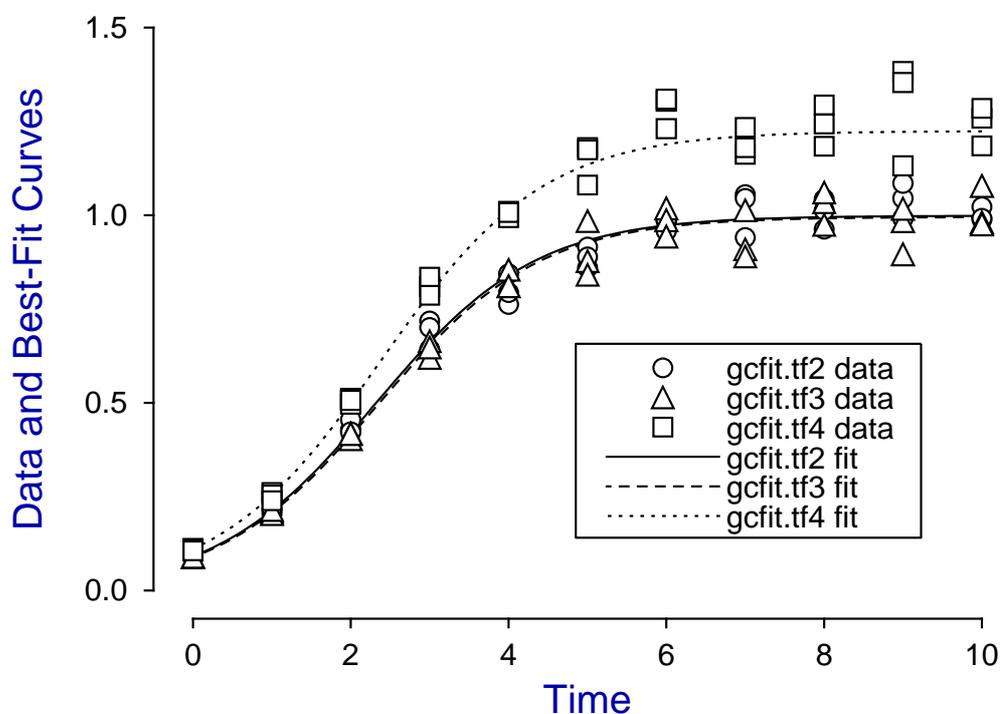
In this case, the logistic model defined as

$$f(t) = \frac{\theta_1}{1 + \theta_2 \exp(-\theta_3 t)}$$

was simulated using **makdat** and **adderr** then fitted by **gcfit**, and the main interest is to decide if the estimated final asymptote i.e. $\hat{\theta}_1$ differs significantly for the test files `gcfit.tf2` and `gcfit.tf3` which actually have identical parameters $\theta_1 = 1$, while `gcfit.tf4` has a slightly larger asymptotic value $\theta_1 = 1.25$, the other parameters being identical $\theta_2 = 10$ and $\theta_3 = 1$.

The next table illustrates how this technique works.

Table of Mahalanobis χ^2, and corrected pairwise t tests for
differences between parameters(A, B) and covariances(Ca, Cb).

Comparison 1: Parameters from `gcfit.tf3` (A) and `gcfit.tf2` (B)
$Q = (A - B)^T (Ca + Cb)^{-1}(A - B) = 2.193E + 00, NDOF = 3$
$P(\chi^2 \geq Q) = 0.5333$

Index	A	B	A - B	t	DOF	p	
1	0.996	0.999	-0.0033	-2.567E-01	53	0.7984	
2	10.15	9.890	0.2600	7.224E-01	40	0.4743	
3	0.985	0.988	-0.0033	-1.164E-02	37	0.9908	

Comparison 2: Parameters from `gcfit.tf4` (A) and `gcffit.tf2` (B)
$Q = (A - B)^T (Ca + Cb)^{-1}(A - B) = 7.492E + 02, NDOF = 3$
$P(\chi_2 \geq Q) = 0.0000$: *Reject H_0 at 1% significance level*

Index	A	B	A - B	t	DOF	p	
1	1.224	0.999	0.2251	19.17	57	0.0000	*****
2	10.04	9.890	0.1500	0.382	50	0.7038	
3	0.969	0.988	-0.0191	-0.063	46	0.9501	

Comparison 3: Parameters from `gcfit.tf4` (A) and `gcfit.tf3` (B)
$Q = (A - B)^T (Ca + Cb)^{-1}(A - B) = 1.064E + 03, NDOF = 3$
$P(\chi^2 \geq Q) = 0.0000$: *Reject H_0 at 1% significance level*

Index	A	B	A - B	t	DOF	p	
1	1.224	0.996	0.2284	16.21	59	0.0000	*****
2	10.04	10.15	-0.1100	-0.443	52	0.6596	
3	0.969	0.985	-0.0158	-0.093	52	0.9265	

The data were fitted using **gcfit** using the option to store parameter estimates and covariance matrices. Then the global tests for different parameter sets, and t tests for individual parameter differences were performed, leading to the results indicated.

Clearly the parameter estimates for test files `gcfit.tf2` and `gcfit.tf3` indicate no significant differences, while `gcfit.tf4` differed significantly from both of these, due to a larger value for the asymptote θ_1 for `gcfit.tf4`.

Graphical deconvolution

There are occasions when a model to be fitted consists of a sum of sub-functions and it is wished to estimate the contribution of the sub-functions to the overall regression. In some instances it may be possible to plot the overall function fitted to the data along with plots for the sub-functions.

This is particularly valuable with models such as the sum of Gaussians, which for three components is

$$f(x) = \frac{A_1}{\sqrt{2}\sigma_1} \exp -\frac{1}{2}\left(\frac{x - \mu_1}{\sigma_1}\right)^2 + \frac{A_2}{\sqrt{2}\sigma_2} \exp -\frac{1}{2}\left(\frac{x - \mu_2}{\sigma_2}\right)^2 + \frac{A_3}{\sqrt{2}\sigma_3} \exp -\frac{1}{2}\left(\frac{x - \mu_3}{\sigma_3}\right)^2.$$

This model is notoriously difficult to fit unless the amplitudes A_i and variances σ_i^2 are of similar size, but the means μ_i are distinct. However it is one of several models where the ability to do such plotting, which is loosely referred to as graphical deconvolution in SimFiT, is provided.

Using SimFiT program **qnfit** to analyze the data in test file `gauss3.tf1` leads to the following table of parameter estimates for parameters defined in terms of Θ as

$$f(x) = \frac{\theta_1}{\sqrt{2}\theta_7}\exp -\frac{1}{2}\left(\frac{x-\theta_4}{\theta_7}\right)^2 + \frac{\theta_2}{\sqrt{2}\theta_8}\exp -\frac{1}{2}\left(\frac{x-\theta_5}{\theta_8}\right)^2 + \frac{\theta_3}{\sqrt{2}\theta_9}\exp -\frac{1}{2}\left(\frac{x-\theta_6}{\theta_9}\right)^2.$$

The columns indicate: the parameter number, the lowest value allowed for the parameter, the highest value allowed for the parameter, the value of the parameter estimate, the standard error of the parameter estimate, the lower 95% confidence limit for the estimate, the upper 95% confidence limit for the estimate, and the significance level for the estimate. The small p values and absence of stars after the last column in the next table of results indicates that all 9 parameters were well determined.

Number	Low-Limit	High-Limit	Value	Std. Error	Lower95%cl	Upper95%cl	p
1	0.000	2.000	0.90754	0.021624	0.8648	0.9503	0.0000
2	0.000	2.000	1.16433	0.042173	1.0810	1.2477	0.0000
3	0.000	2.000	0.92519	0.030130	0.8656	0.9848	0.0000
4	-2.000	2.000	-0.07298	0.015572	-0.1038	-0.0422	0.0000
5	2.000	6.000	3.74510	0.050816	3.6446	3.8456	0.0000
6	8.000	12.00	10.2774	0.096413	10.087	10.468	0.0000
7	0.100	2.000	0.92641	0.014331	0.8981	0.9547	0.0000
8	1.000	3.000	2.34330	0.070567	2.2038	2.4828	0.0000
9	2.000	4.000	2.76906	0.062637	2.6452	2.8929	0.0000

This conclusion is reinforced by the next graphical deconvolution plot showing the data as dots, the best-fit curve as a dotted line, and the components contributing to the best-fit curve as red, green, and blue curves.

Graphical Deconvolution of Three Gaussians