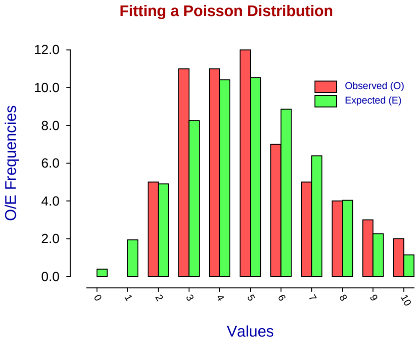
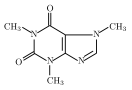
sv_simfit

Simplified version for first time users

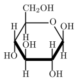
W. G. Bardsley

https://simfit.uk
https://simfit.org.uk
https://simfit.silverfrost.com

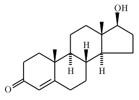




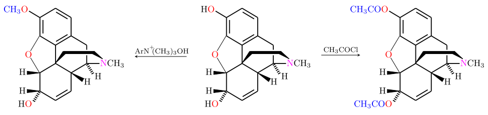
Caffeine



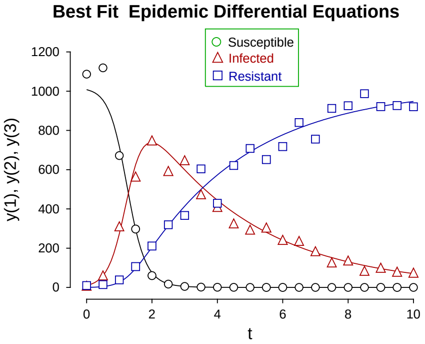
Glucose

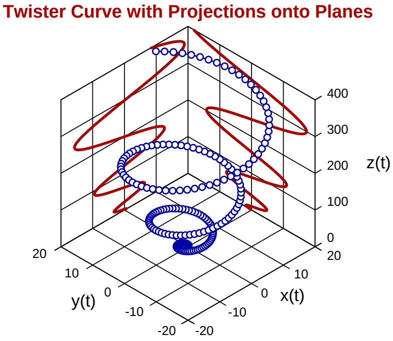


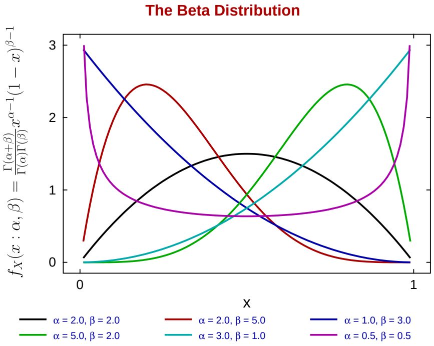
Testosterone

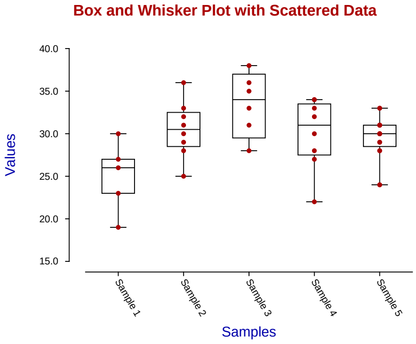


Codeine Morphine Heroin









sv_manual: Version 8.1.4

Contents

1 Introduction 4
- 1.1 The main menu . . . 4
- 1.2 How to use **sv_simfit** . . . 4
- 1.3 File open/close actions and saving results files . . . 5
- 1.4 Viewing and retrieving results . . . 6

2 Make or Edit data files 7
- 2.1 Important facts about making and editing files . . . 7
- 2.2 Make/Edit/Plot a curve-fit/plot file . . . 8
- 2.3 Make a vector/matrix file . . . 8
- 2.4 Edit a vector/matrix file . . . 8
- 2.5 Make a file from *.txt, *.csv, *.html, Excel or any cllipboard data . . . 8

3 Methods to input data 9

4 Linear regression 11
- 4.1 The definition of linear regression . . . 11
- 4.2 What does fitting a model mean . . . 12
- 4.3 Using the test data . . . 12

5 Nonlinear regression 14
- 5.1 Fitting exponentials . . . 14
- 5.2 Fitting growth models . . . 15
  - 5.2.1 Exponential growth . . . 15
  - 5.2.2 Monomolecular growth . . . 15
  - 5.2.3 Logistic growth . . . 16
  - 5.2.4 The Gompertz model . . . 16
  - 5.2.5 Example: Using program **gcfit** . . . 16
- 5.3 Fitting the Michaelis-Menten equation . . . 17

6 Statistics 19
- 6.1 Data exploration . . . 20
  - 6.1.1 Exhaustive analysis of a vector . . . 20
  - 6.1.2 Exhaustive analysis of a matrix . . . 20
- 6.2 Statistical analysis . . . 21
  - 6.2.1 Statistical power and sample size . . . 21
  - 6.2.2 Estimation of parameter confidence limits . . . 23
- 6.3 Numerical analysis . . . 23
  - 6.3.1 Zeros of a polynomial . . . 23
  - 6.3.2 Matrix determinant/eigenvalues/inverses . . . 23
  - 6.3.3 Singular value decomposition . . . 23

7 Calibration 24
- 7.1 Details about calibration . . . 24
- 7.2 Using polynomials . . . 25
- 7.3 Using a straight line . . . 25
- 7.4 Using a quadratic . . . 25
- 7.5 Using cubic splines . . . 28

8 Graph plotting **29**

8.1 Details about using the plotting options 29

8.2 Plotting up to four data sets 30

8.3 A pie chart 30

8.4 A bar chart 31

8.5 A space curve 31

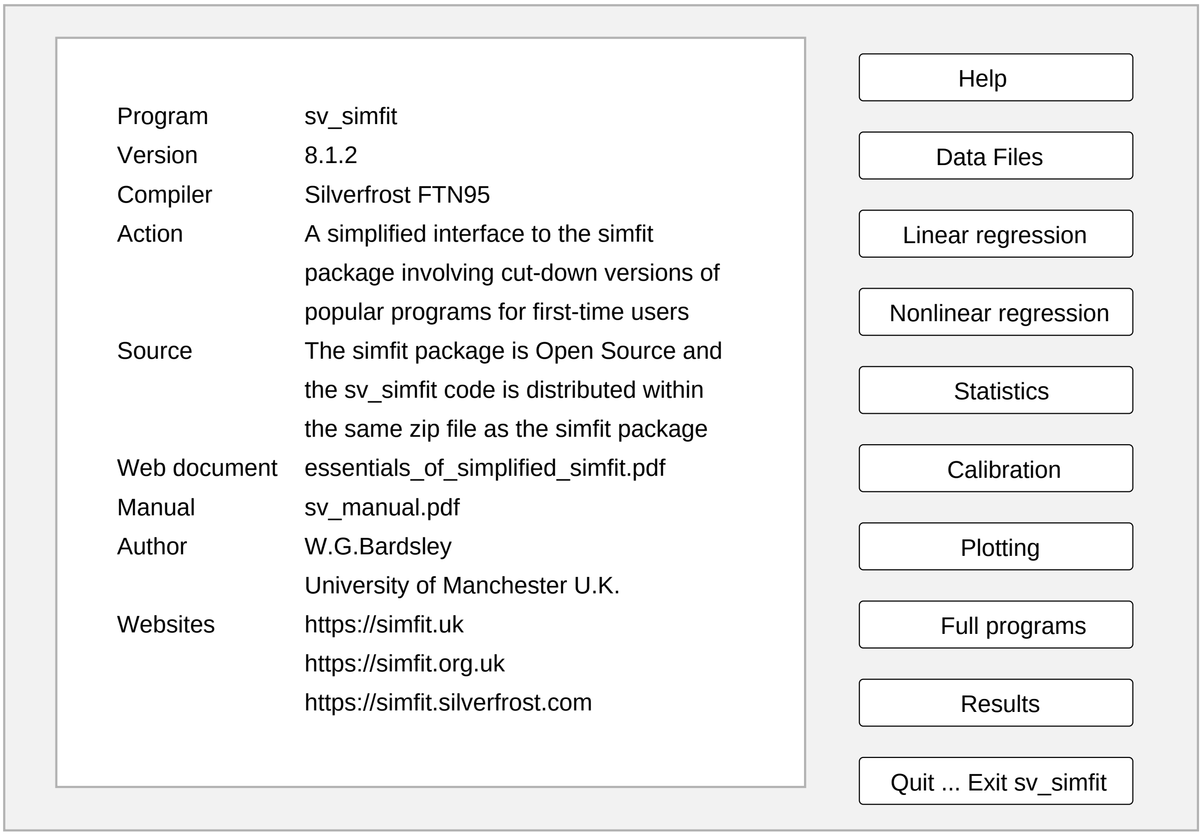
8.6 Surfaces and contours 32

1 Introduction

Note that allowing the mouse to hover over any button on the file–open–save dialogue or on a SimFIT graph displays a message about what the function of the button is.
Also a mouse right–click on any SimFIT graph opens an extensive help document.

1.1 The main menu

On opening program **sv_simfit** the next item appears.



Program	sv_simfit
Version	8.1.2
Compiler	Silverfrost FTN95
Action	A simplified interface to the simfit package involving cut-down versions of popular programs for first-time users
Source	The simfit package is Open Source and the sv_simfit code is distributed within the same zip file as the simfit package
Web document	essentials_of_simplified_simfit.pdf
Manual	sv_manual.pdf
Author	W.G.Bardsley University of Manchester U.K.
Websites	https://simfit.uk https://simfit.org.uk https://simfit.silverfrost.com

1.2 How to use sv_simfit

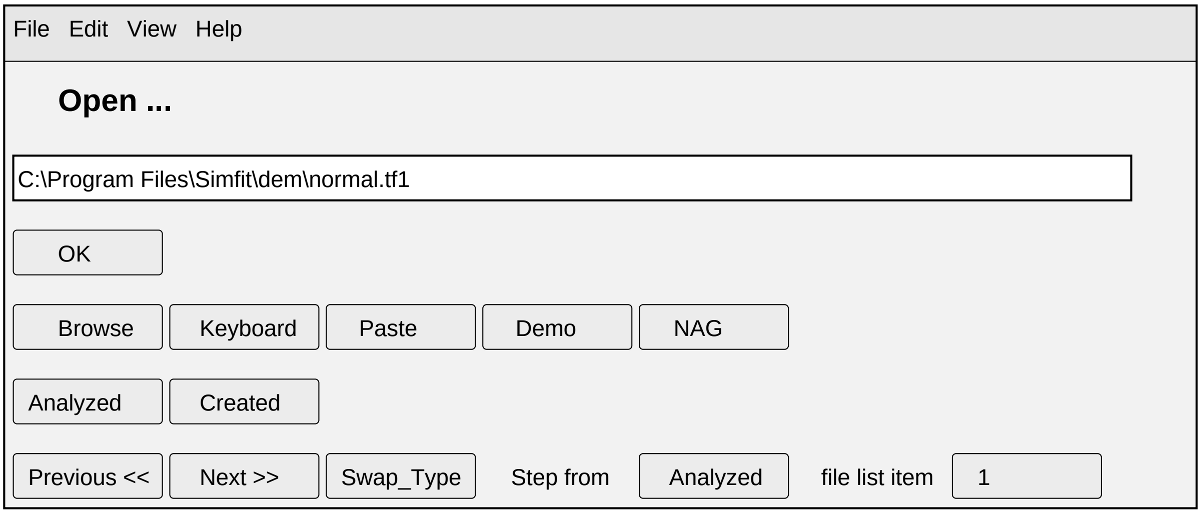
The SimFIT package is very extensive ranging from elementary subjects like linear regression, statistics, and data file creation to advanced subjects like statistical calculations, numerical analysis, simulation or constrained nonlinear optimisation using a built in library of mathematical models, or user–defined models from single equations up to systems of nonlinear differential equations.

As this covers a very large compass where each area is dealt with by individual free–standing programs all driven by the driver program **x64_simfit.exe** it is necessary to provide a simple interface for first–time users. So each time an option is chosen from the simplified program version **sv_simfit.exe** it simply calls one of the full programs with a flag that restricts the number of choices available. For example. The full program **gcfit** fits up to ten growth or survival models plus options for functions such as survival analysis, etc., while **sv_gcfit** just fits up to four growth models.

Clearly, if **sv_simfit** does not seem to provide the options required, users should to choose [Full programs] to try the main parent program. For more information first consult the tutorials then, for the ultimate explanation of all the statistical and mathematical procedures used by SimFIT you have consult the reference manual **w_manual.pdf**.

1.3 File open/close actions and saving results files

Most users will be familiar with the Windows File–Open and File–Save–As dialogs which requires searching for folders which can be tedious. To avoid this SimFIT uses a specialised technique to open or close files as illustrated below.



C:\Program Files\Simfit\dem\normal.tf1

This control helps you to create new files (i.e., in the Save As . . . mode) or analyze existing files (i.e., in the Open . . . mode). The `[File]`, `[Edit]`, `[View]`, and `[Help]` menus allow you to select appropriate test files to use for practise or browse to understand the formatting. Below this is an edit box and a set of buttons. Details of the functions of each button are now described while mouse–hover displays a summary.

`File Name` You can type the name of a file into the edit box but, if you do this, you must type in the full path. If you just type in a file name you will get an error message, since SimFIT will not let you create files in the SimFIT folder, or in the root, to avoid confusion.

`OK` This option indicates that the name in the edit box is the file name required.

`Browse` This option simply transfers you to the Windows control but, when you know how to use the SimFIT file selection control properly, you will almost never use the Windows control.

`Keyboard` This option allows you to type in a data set from the keyboard. It is only useful if you understand the file format required and have very small samples. It creates a temporary file in your usr folder.

`Paste` This option is only activated when SimFIT detects ASCII text data on the clipboard and, if you choose it, then SimFIT will attempt to analyze the clipboard data. If the clipboard data are correctly formatted, SIMFIT will create a temporary file, which you can subsequently save if required. If the data are not properly formatted, however, an error message will be generated. When highlighting data in your spreadsheet to copy to the clipboard, write to a comma delimited ASCII text file, or use a with a macro like `simfit6.xls`, you must be very careful to select the columns for analysis so that they all contain exactly the same number of rows.

`Demo` This option provides you with a set of test files that have been prepared to allow you to see SIMFIT in action with correctly formatted data. Obviously not all the files displayed are consistent with all the possible program functions. With programs like **simstat**, where this can happen, you must use the [Help] option to decide which file to select. When you use a SIMFIT program for the first time, you should use this option before analyzing your own data.

`NAG` This option provides you with a set of test files that have been prepared to allow you to use SIMFIT to see how to use NAG library routines.

`Analyzed` This history option allows you to choose from a list of the last files that SIMFIT has analyzed, but the list does not contain files recently saved.

`Created` This history option allows you to choose from a list of the last files that SIMFIT has created, but the list does not contain files recently analyzed. Of course, many files will first be created then subsequently analyzed, when they would appear in both Analyzed and Created lists.

`Previous` This option allows you to scroll backwards through recent files and edit the filename, if required, before selecting.

`Next` This option allows you to scroll forwards through recent files and edit the filename, if required, before selecting.

`Swap Type` This option toggles between Created and Analyzed file types.

If you name files sensibly, like results.1, results.2, results.3, and so on, and always give your data short meaningful titles describing the data and including the date, you will find the [Back], [Next], [Created] and [Analyzed] buttons far quicker and more versatile than the [Browse] option.

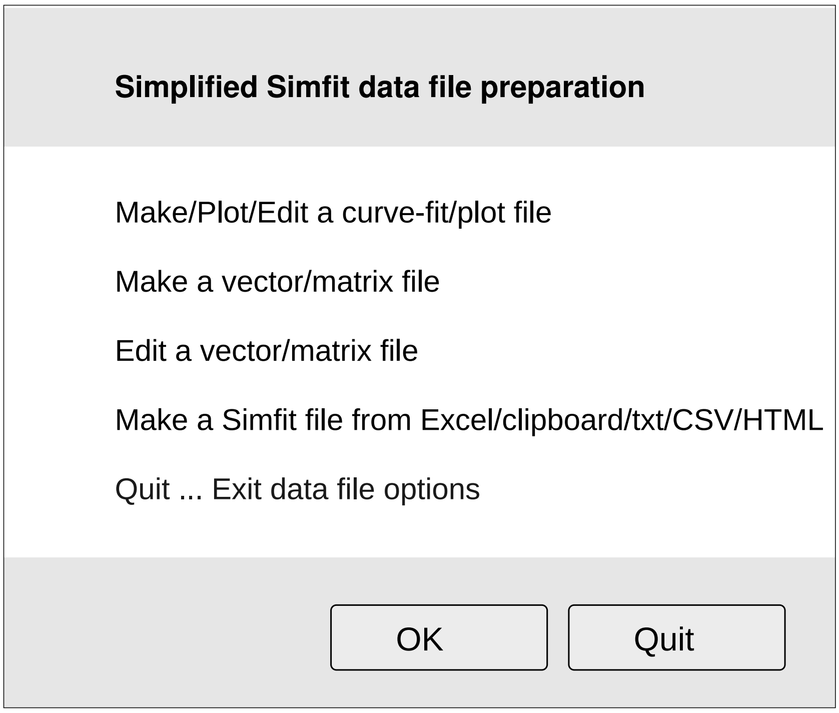
There are many occasions when it is required to input several collected files for plotting, etc. and this is facilitated by the library file technique. This is simply a title followed by a list of files to be used. Filenames must be fully qualified with paths and are most easily prepared using program **maklib.** The method SIMFITprovides for archiving results is by creating tables of parameter estimates and important statistical conclusions. These are copied to a results files e.g., **f$result.001**,etc. for viewing and saving if required later as now explained.

1.4 Viewing and retrieving results

When an analysis is conducted SIMFIT programs save all results to a results file which is added to a list of recently created results files. Each time this happens the results file at the bottom of the list is deleted, the remaining files are renamed, and the new file becomes top of the list. These can be accessed by choosing the [Results] option from the main SIMFIT page which offers the option to view, save, or copy chosen results so that, if some results are likely to be required later, that results file must be saved as it will eventually be deleted.

2 Make or Edit data files

The following options are provided.



Simplified Simfit data file preparation

Make/Plot/Edit a curve-fit/plot file

Make a vector/matrix file

Edit a vector/matrix file

Make a Simfit file from Excel/clipboard/txt/CSV/HTML

Quit ... Exit data file options

2.1 Important facts about making and editing files

Most editors of numerical data tables allow users to enter values anywhere and then Save or Save As then exit. If there are empty cells or further editing is required the files created can be input again. However SIMFIT does not work this way for several reasons.

1. Only relatively small tables can be filled–in using the SIMFIT editors as it is much better with larger tables to copy the rows and columns required from your spreadsheet program to the clipboard and then paste in to the chosen program when required, or else input into program **maksim** which can make SIMFIT files. With Excel the macro **simfit6.xls** is recommended. The advantage of using the SIMFIT editors is that every time a cell is filled–in the numerical value is checked when the enter–key is pressed or the focus is changed so that you are forced to fill cells in a logical order because there may be restrictions such as:

   - columns may have to be in ascending order ,e.g., column1 for curve fitting files, or plotting files or be all non-negative in column 3 if weights are required, or column 1, less than column 2 which is less that column 3 with parameter limits files;
   - sometimes numbers must be integers limited, for instance to 0 or 1 or sometimes -1, 0, 1;

2. Files to be created must be specified before creation or editing is allowed;

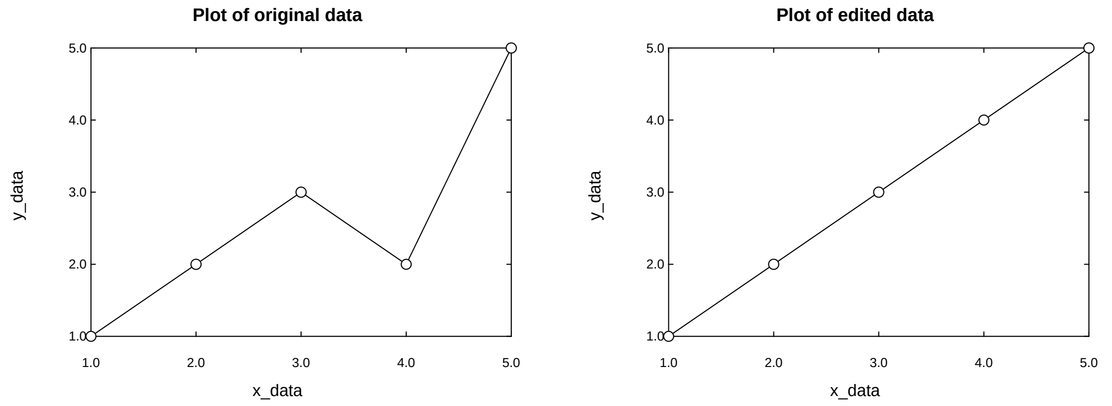
3. The number of rows and columns must be specified before any new files are created; and

4. an exit before all cells have been filled–in causes default values to be inserted into the table which then requires subsequent editing.

2.2 Make/Edit/Plot a curve-fit/plot file

Curve-fitting and straightforward plotting files have two columns for x and y where x would usually be in numerically increasing order. When the values have been input there are subsequent options to view or plot the data entered, and edit if required before saving to the file originally specified. For a simple example, suppose it is required to draw a straight line between $x = 1$ and $x = 5$ for the case $y = x$ but a mistake is made as follows

```
x    y
1    1
2    2
3    3
4    2  <-- ERROR, and below the plots before and after correcting.
5    5
```



2.3 Make a vector/matrix file

The technique is as with making curve-fit files except that there are no restrictions on values entered or the order within or between columns. You still need to use the enter key or a focus change after each cell has been filled in and will be warned about default values being placed in any empty cells on exit. It is very useful for creating files where the first column does not need to be in increasing order, e.g. files for statistical analysis or for plotting data with loops as in drawing an implicit function like a circle.

2.4 Edit a vector/matrix file

Any cell can be edited but, unlike the full version there are no global options and rows or columns cannot be moved, duplicated or deleted.

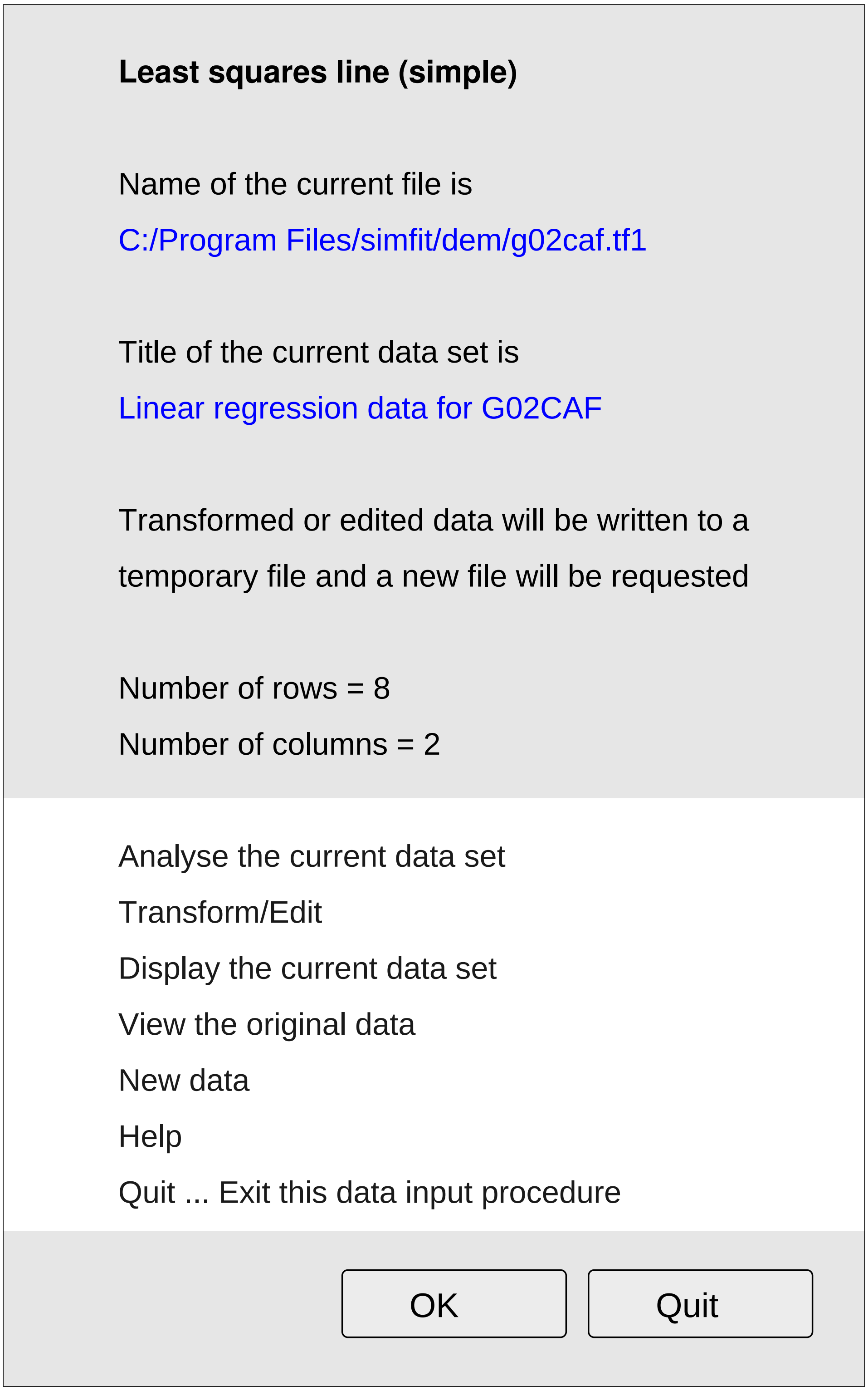
2.5 Make a file from *.txt, *.csv, *.html, Excel or any cllipboard data

This opens a simplified interface to program **maksim**.

Any rectangular numerical table with no missing values can be entered into this program which will then create a SimFIT type file.

3 Methods to input data

Selecting linear regression provides the control shown next which is used everywhere in SimFIT when data input is required.



Least squares line (simple)

Name of the current file is

C:/Program Files/simfit/dem/g02caf.tf1

Title of the current data set is

Linear regression data for G02CAF

Transformed or edited data will be written to a temporary file and a new file will be requested

Number of rows = 8

Number of columns = 2

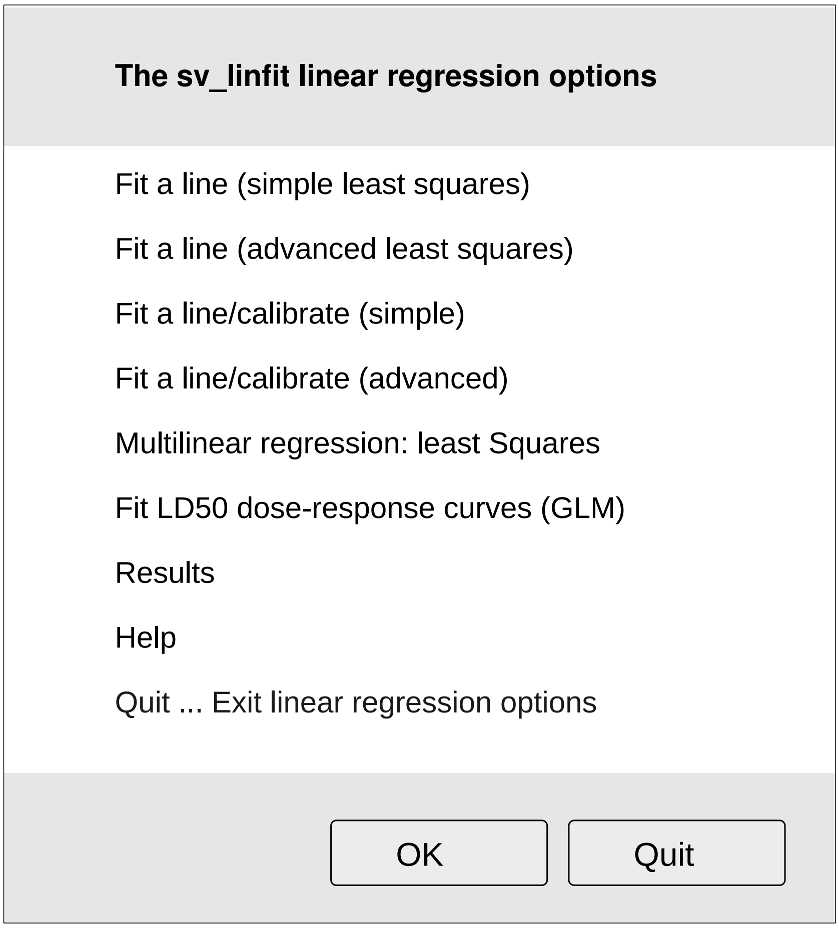
Analyse the current data set
Transform/Edit
Display the current data set
View the original data
New data
Help
Quit ... Exit this data input procedure
OK
Quit

The first time this appears it will refer to an appropriate SimFIT test file for the selected procedure and you can proceed as follows. .

1. Analyse the selected file.
2. Transform or edit the selected file before fitting.
3. Display the numerical data table that has been selected.
4. Display the data file that has been selected with header, data, then trailer sections so you will appreciate how to create a new file containing your own observations.
5. Input a new file from your file store or the clipboard.
6. Read help documents with additional information so you will appreciate the options provided when a file has been opened.

4 Linear regression

This section concerns **sv_linfit** the simplified interface to SIMFIT program **linfit** as illustrated next.


The sv_linfit linear regression options
Fit a line (simple least squares)
Fit a line (advanced least squares)
Fit a line/calibrate (simple)
Fit a line/calibrate (advanced)
Multilinear regression: least Squares
Fit LD50 dose-response curves (GLM)
Results
Help
Quit ... Exit linear regression options
OK
Quit

4.1 The definition of linear regression

First of all it is important to be clear about the definition of linear regression.

Linear regression concerns fitting a linear model to weighted or unweighted data where a linear model $f(\theta, x)$ is defined as

$$f(\theta, x) = \theta(1)g_1(x) + \theta(2)g_2(x) + \cdots + \theta(m)g_m(x)s$$

where the model is a linear sum of m components as follows

1. there are m parameters $\theta(1), \theta(2), \ldots \theta(m)$ to be estimated,
2. there are m functions $g_1, g_2, \ldots, g_m$ to be evaluated of variables x, where
3. x could be a multidimensional vector of variables that may be weighted.

Because the model only involves the parameters as linear coefficients then fitting is easy and, in well–defined examples, should yield the same unique parameter estimates.

Some simple examples could be

- A simple straight line $y = ax + b$, where $\theta(1) = a$ and $\theta(2) = b$
- A quadratic $y = ax^2 + bx + c$
- A polynomial of degree n such as $y = p(0) + p(1)x + p(2)x^2 + \cdots + p(n)x^n$
- A multilinear model $f(A, x) = A(0)x_1 + A_2x_2 + \cdots + A(n)x_n$

There is also the case of Generalized Linear Models (GLM) which is more advanced as the models are defined implicitly not explicitly. However, as this technique is widely used for analysis, e.g., in logistic regression a simple example is included in the linear regression section.

4.2 What does fitting a model mean

Given n observations y_i with unknown experimental error but with known or estimated standard deviations s_i at n values of the variable x_i presumed to be without measurement error then the parameters are calculated by minimising the weighted sum of squared residuals $WSSQ$ where the residuals, using the difference between the data and theoretical prediction to estimate the experimental error, are

$$y_i - f(\theta(i), x_i)$$

and $WSSQ$ is defined as

$$WSSQ = \sum_{i=1}^{n} \left(\frac{y_i - f(\theta(i), x_i)}{s_i} \right)^2$$

.

4.3 Using the test data

This is the data contained in test file g02caf.tf1

x	y
0.0	10.0
0.0	15.5
1.0	20.0
2.5	24.5
4.0	28.3
5.0	31.2
7.5	45.0
10.0	99.0

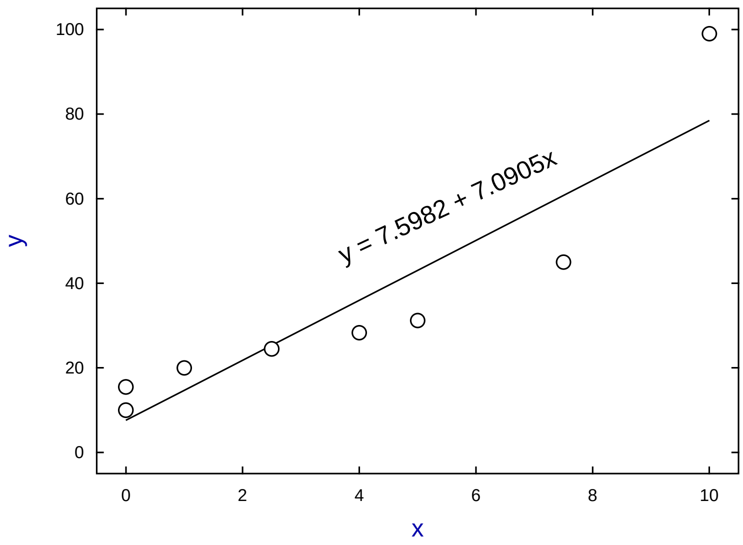
Analysis yields the following results table and plot for the least squares best-fit straight line.

Parameter	Value	Std. Error	Lower95%cl	Upper95%cl	p	
constant (c)	7.5982	6.6858	-8.7613	23.958	0.2991	**
slope (m)	7.0905	1.3224	3.8548	10.326	0.0017	

($r^2 = 0.8273$, $r = 0.9096$, $p = 0.0017$)

This table displays the parameter estimates together with the estimated standard error and 95% confidence limits and, after fitting, the graph of best–fit line and data is displayed as shown next.

Least Squares Linear Regression for G02CAF.TF1

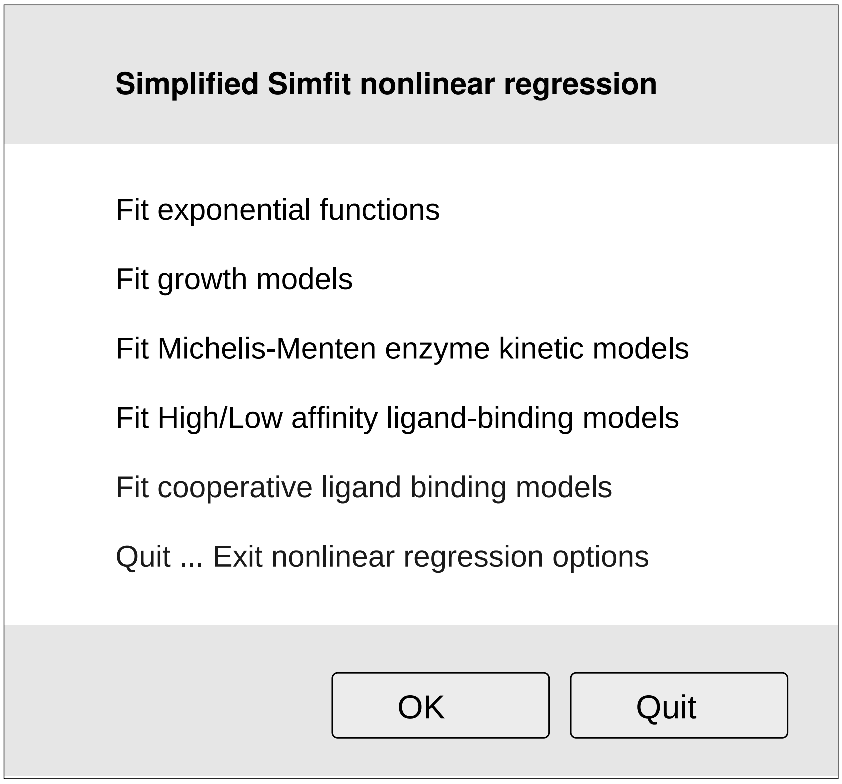


5 Nonlinear regression

Linear regression is extremely easy to perform but unfortunately in real life few if any phenomena are linear so fitting linear models to nonlinear data always produces biased results. Unfortunately nonlinear regression is much more difficult because of several reasons as follows.

1. The model chosen must be correct
2. The data must be extensive and accurate
3. Starting estimates must be close to the correct parameters being estimated
4. The regression requires iterative procedures and may lead into a local rather than global minimum
5. The parameters must be constrained which requires constrained nonlinear optimisation
6. To avoid the solution being influenced by large observed values weights are often required

sv_simfit provides several programs as in the next menu where starting estimates are estimated from the data and analysis is provided to check the validity of the solution.



Simplified Simfit nonlinear regression

Fit exponential functions

Fit growth models

Fit Michelis-Menten enzyme kinetic models

Fit High/Low affinity ligand-binding models

Fit cooperative ligand binding models

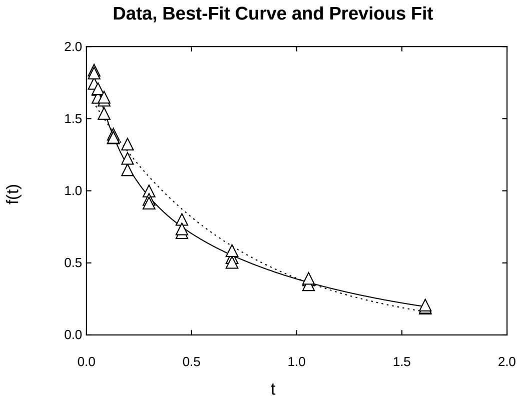
Quit ... Exit nonlinear regression options

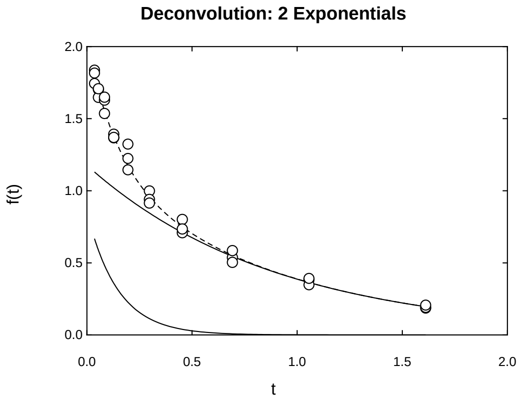
5.1 Fitting exponentials

This is extensively used in chemical kinetics and particularly compartmental analysis in pharmacokinetics where the number of exponentials is equal to the number of compartments. As an example we consider a simplified version of SimFIT program **exfit** used to fit a sum of two exponentials, i.e., the model

$$f(t) = A(1)\exp -k_1 t + A(2)\exp -k_2 t$$

to the test file `exfit.tf4`. Next a plot of the fit to one then two exponentials followed by the graphical deconvolution to observe the contribution of the two exponentials to the overall fit.





5.2 Fitting growth models

Whereas the treatment of exponentials and the Michaelis-Menten case involves models that are a linear combination of sub–models there are many examples of models that can fit a data set where a decision as to the best–fit model is not so straightforward. For instance growth models are required for many different situations ranging from the size of bacterial cultures to the development of human infants and there are also situations where the growth models can be used to model decay or survival. Some simple models for size as a function of time follow and it is usual to choose a subset from the full set of models and observe a summary of the statistics provided to aid selection of the best fit model.

5.2.1 Exponential growth

The exponential model

$$\begin{aligned} dS/dt &= kS \\ S(t) &= A\exp(kt) \\ S(0) &= A \end{aligned}$$

is only used for the initial phase of growth of a bacterial culture before the food supply or other factors slow down the growth process but is also used for the early stages of many growth processes.

5.2.2 Monomolecular growth

The monomolecular model defined by

$$\begin{aligned} ds/dt &= k(A - S) \\ S(t) &= A[1 - B\exp(-kt)] \\ S(0) &= A(1 - B) \\ S(\infty) &= A \end{aligned}$$

is only used when the data are very limited and it is only needed to estimate the value at $t = 0$ and the asymptote as $t \to \infty$.

5.2.3 Logistic growth

The logistic model defined by

$$
\begin{aligned}
dS/dt =& kS(A - S)/A \\
S(t) =& \frac{A}{1 + B\exp(-kt)} \\
S(0) =& \frac{A}{1 + B} \\
S(\infty) =& A
\end{aligned}
$$

is widely used as it starts like exponential growth and finishes like monomolecular growth. If data are provided in survival i.e., decreasing form the program recognises this and fits $A - f(t)$.

5.2.4 The Gompertz model

The Gompertz model defined by

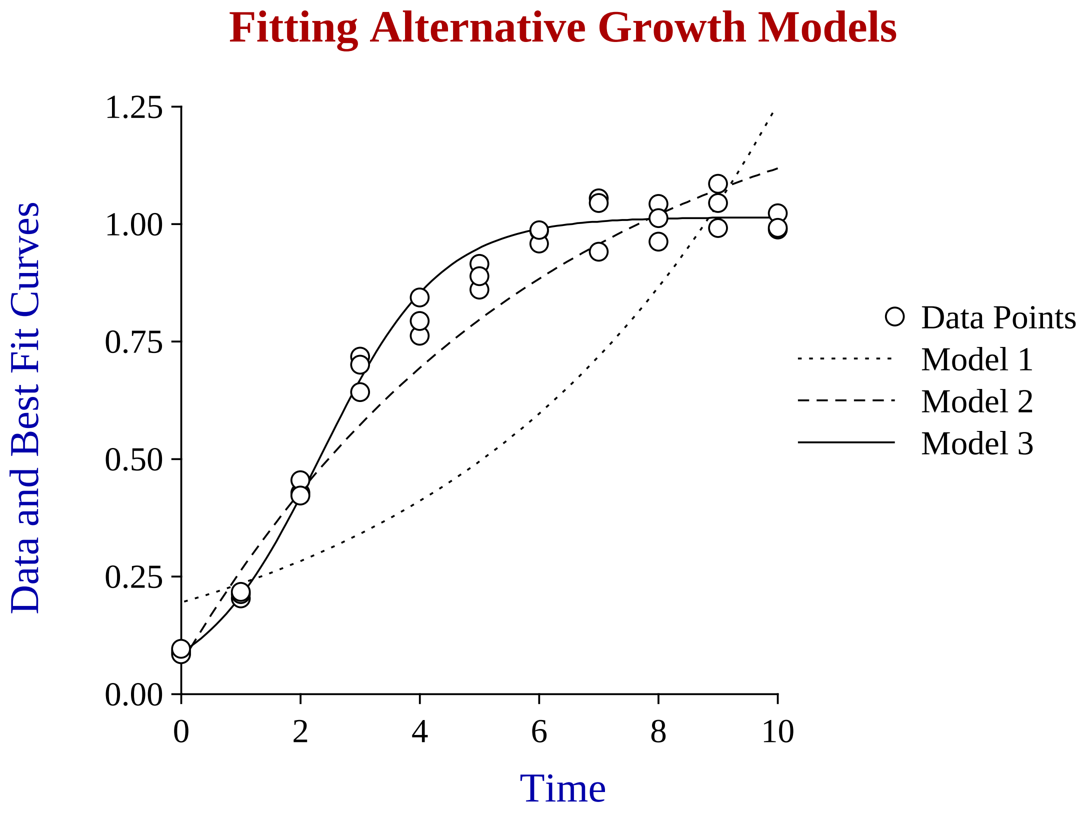
$$
\begin{aligned}
dS/dt =& kS[\log(A) - \log(S)] \\
S(t) =& A\exp[-B\exp(-kt)] \\
S(0) =& A\exp(-B) \\
S(\infty) =& A
\end{aligned}
$$

often gives a better fit than the logistic model as it has a more flexible form.

5.2.5 Example: Using program gcfit

Fitting the test data gcfit2 using models 1(exponential), 2(monomolecular), and 3(logistic) we can compare the graphs of best-fit curves to data shown next.

The exponential model gives a very poor fit, the monomolecular model leads to an improved fit, but the logistic model is much better.



The program also provides information about the best-fit parameters and a table comparing numerous statistics to confirm what is evident from the graphs as follows:

Model	$WSSQ/NDOF$	$P(C \geq W)$	$P(\text{Runs} \leq r)$	$N > 10\%$	$N > 40\%$	$Av.r\%$	Verdict
1	152.0	0.000	0.000	29	17	40.03	Very bad
2	18.1	0.000	0.075	20	0	12.05	Very poor
3	1.32	0.113	0.500	0	0	3.83	Incredible

5.3 Fitting the Michaelis-Menten equation

A simplified version of SimFIT program **mmfit** is provided to fit just one equation or the case of a mixture of several isoenymes with different parameters, or an enzyme with two or more independent active sites.

Before the availability of computers experimentalists had to analyse Michaelis-Menten kinetic by transforming the data into double reciprocal form and extrapolating plots to estimate parameters as follows.

$$v = \frac{V_{max}S}{K_m + S} \quad \text{or, in double reciprocal form} \quad \frac{1}{v} = \frac{K_m}{V_{max}}\left(\frac{1}{S}\right) + \frac{1}{V_{max}}$$

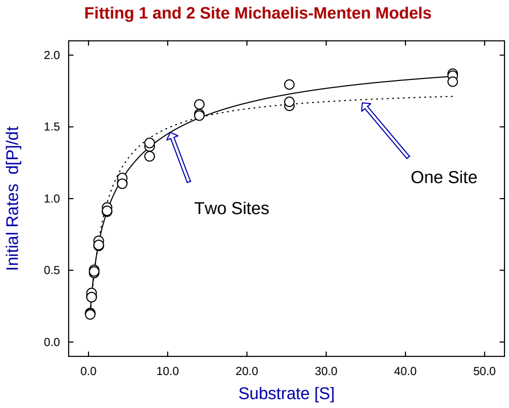
This has the disadvantage that it gives biased parameter estimates and does not allow deviations from Michaelsi-Menten Menten due to a mixture enzymes or several active site. So today constrained nonlinear optimisatiom is used but this requires parameter starting estimates and statistical techniques to confirm conclusions about the need for multiple active sites.

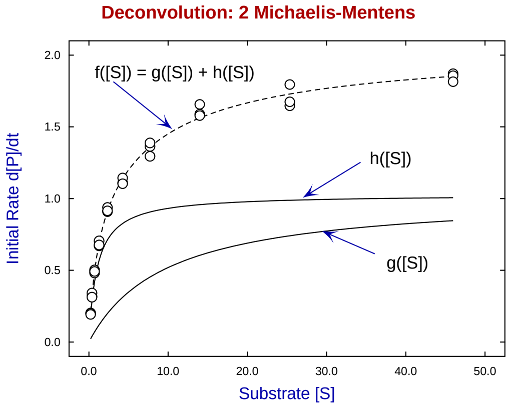
The test file provided is `mmfit.tf4` and the program will scale the data and estimate parameter starting estimates by calculation and also by performing a random search using the next equations

$$f_1(S) = \frac{V_{max}S}{K_m + S}$$

$$f_2(S) = \frac{V_{max_1}S}{K_{m_1} + S} + \frac{V_{max_2}S}{K_{m_2} + S}$$

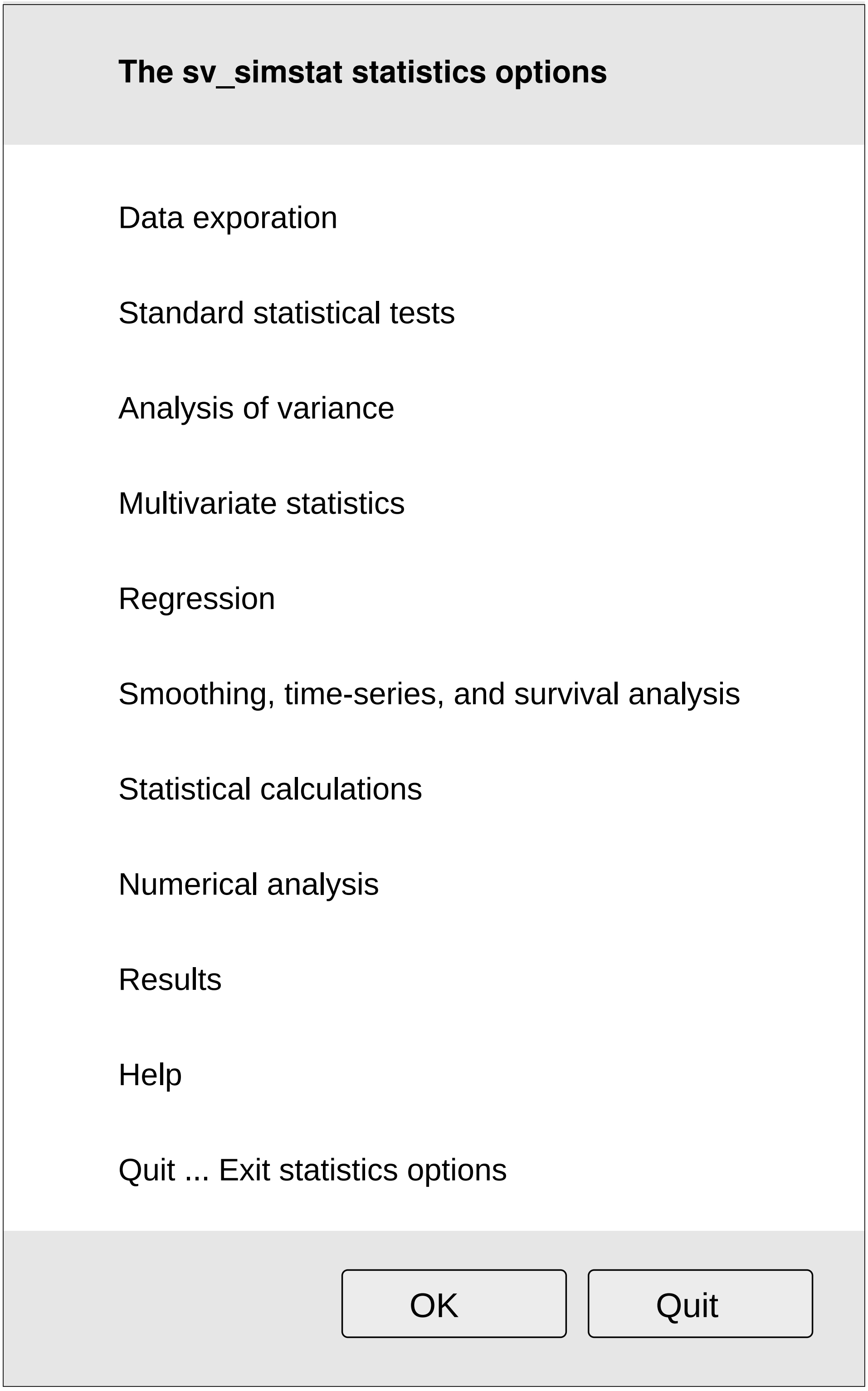
which leads to the following plots


Fitting 1 and 2 Site Michaelis-Menten Models


Deconvolution: 2 Michaelis-Mentens

6 Statistics

The SimFIT statistical analysis options are very extensive and the simplified version of **simstat** only provides a limited number of these so, if the tests you require are not included here, you will have to use the main program. The options available in the simplified version are shown next.



The sv_simstat statistics options
Data exporation
Standard statistical tests
Analysis of variance
Multivariate statistics
Regression
Smoothing, time-series, and survival analysis
Statistical calculations
Numerical analysis
Results
Help
Quit ... Exit statistics options
OK
Quit

Most of these options such as t-tests or analysis of variance will be obvious but some may not be familiar and should be explained.

6.1 Data exploration

This provides many techniques for analysing arbitrary samples and two examples will be given.

6.1.1 Exhaustive analysis of a vector

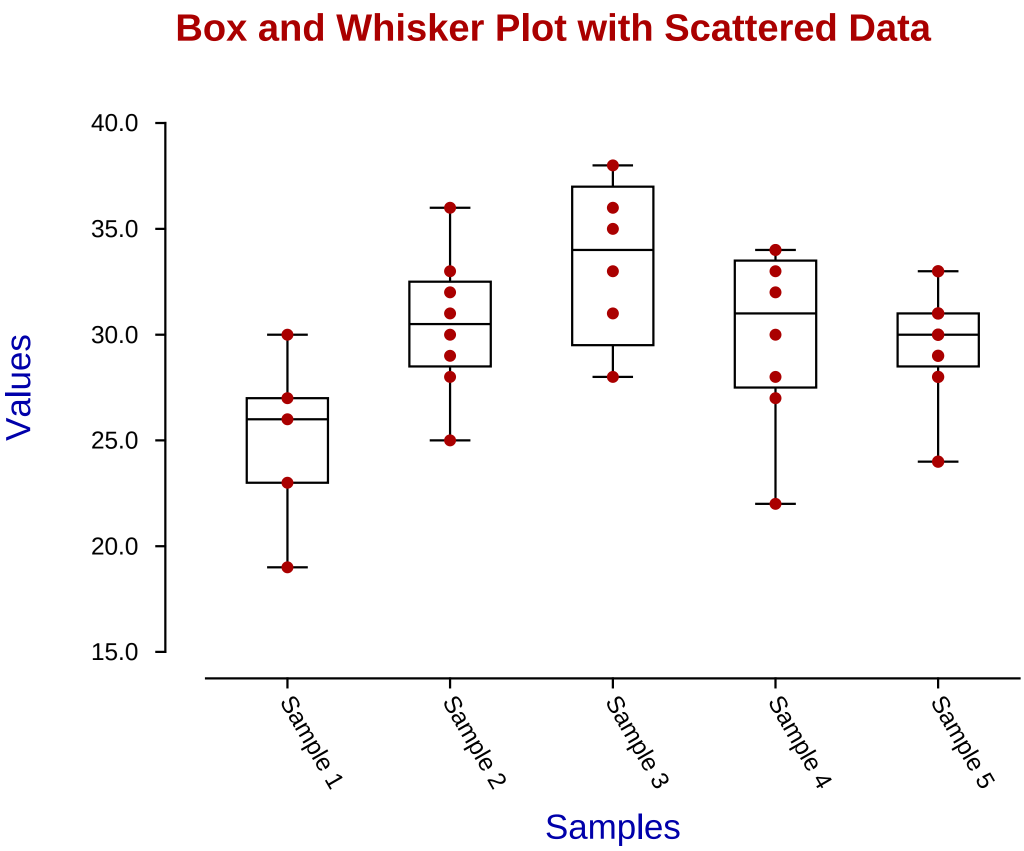
This option makes a very thorough investigation of a single sample of observations arranged as a vertical column, i.e., a vector, including calculation of sample moments and testing for a normal distribution as follows.

Exhaustive analysis of a vector
Data: Test file normal.tf1: 50 random numbers

Sample size	50
Minimum, Maximum values	-2.20820, 1.61750
Lower and Upper Hinges	-0.85502, 0.78597
Coefficient of skewness	-0.01669
Coefficient of kurtosis	-0.76840
Median value	-0.09736
Sample mean	-0.02579
Sample standard deviation	1.00553: CV% = 3899%
Standard error of the mean	0.14220
Upper 2.5% t-value	2.00958
Lower 95% confidence limit for mean	-0.31156
Upper 95% confidence limit for mean	0.25998
Variance of the sample	1.01109
Lower 95% confidence limit for variance	0.70552
Upper 95% con limit for variance	1.57006
Shapiro-Wilks W statistic	0.96270
Significance level for W	0.1153 *Tentatively accept normality*

6.1.2 Exhaustive analysis of a matrix

This analyses an arbitrary data sample arranged as a rectangular table,i.e., a matrix, using numerical and graphical techniques. For instance scatter/error bar plots can be created as well as box and whisker plots as in the following figure. Note that the data sample can be incomplete for this particular analysis , i.e. can have missing values.



6.2 Statistical analysis

SimFIT provides many options to analyse data sets to analyse the evidence to support or question hypothesis tests and two examples are given.

6.2.1 Statistical power and sample size

For instance a power calculations for 2 supposed normal samples in an unpaired t-test. This calculation is based upon the pooled variance s_p^2 which should be input as the estimate for the common variance σ^2, i.e.,n

$$s_p^2 = \frac{\sum_{i=1}^{n_x}(x_i - \bar{x})^2 + \sum_{j=1}^{n_y}(y_j - \bar{y})^2}{n_x + n_y - 2}$$

where X has sample size n_x and Y has sample size n_y. The following options are available:

- To calculate the sample size necessary to estimate the difference between the two population means within a half width h

$$n = \frac{2s_p^2 t_{\alpha/2,2n-2}^2}{h^2};$$

- To calculate the sample size necessary to detect an absolute difference δ between population means

$$n = \frac{2s_p^2}{\delta^2}(t_{\alpha/2,2n-2} + t_{\beta,2n-2})^2; \text{ or}$$

- To estimate the power

$$t_{\beta,2n-2} = \frac{\delta}{\sqrt{2s_p^2/n}} - t_{\alpha/2,2n-2}.$$

The *t* test has maximum power when $n_x = n_y$ but, if the two sample sizes are unequal, calculations based on the the harmonic mean n_h should be used, i.e.,

$$\begin{aligned} n_h &= \frac{2n_x n_y}{n_x + n_y}, \\ \text{so that } n_y &= \frac{n_h n_x}{2n_x - n_h}. \end{aligned}$$

Example

In order to perform calculations it is necessary to assume that both samples are from normal distributions with the same variance and then input those of the following parameters as required.

- The significance level α
- An accurate estimate for the common variance s^2
- Choice of a 2-tail test or 1-tail test
- The half width h to determine a 95% confidence range $2h$ for the difference between the sample means
- The minimum absolute difference δ between the sample means that can be detected
- The power $100(1 - \beta)\%$
- The sample size n

The following table resulted.

Power analysis for the *t* test

$n(h)$	$h = 1$	$\alpha = 0.05$		$s^2 = 1$	$n = 9$
$n(\delta)$	$\delta = 1$	$\alpha = 0.05$	$\beta = 0.2$	$s^2 = 1$	$n = 17$
$\delta(n)$	$n = 17$	$\alpha = 0.05$	$\beta = 0.2$	$s^2 = 1$	$\delta = 0.9912$
$\beta(n)$	$n = 17$	$\delta = 1$	$\alpha = 0.05$	$s^2 = 1$	$\beta = 0.1931$
$n(h)$	$h = 0.5$	$\alpha = 0.05$		$s^2 = 0.5193$	$n = 18$
$n(\delta)$	$\delta = 0.5$	$\alpha = 0.05$	$\beta = 0.1$	$s^2 = 0.5193$	$n = 45$
$\beta(n)$	$n = 15$	$\delta = 1$	$\alpha = 0.05$	$s^2 = 0.5193$	$\beta = 0.0454$

The last three entries in the above table would be typical. They are for two samples of size $n = 15$ with pooled variance $s^2 = 0.5193$, and the results would be interpreted as follows.

- $n(h)$ shows that a sample size of $n = 18$ would be required to have a 95% confidence interval for the difference between the true means no larger than 1, that is with $h = 0.5$.

- $n(\delta)$ illustrates that a sample size of $n = 45$ is necessary in order for a 90% chance of detecting a difference δ between the true means as small as 0.5.
- $\beta(n)$ demonstrate that the power for detecting a difference of $\delta = 1$ between the true means has a power of 95.46%.

6.2.2 Estimation of parameter confidence limits

Given a sample and supposed distribution then parameters with confidence limits can be calculated. For example confidence limits for a normal mean and variance.
If the sample mean is $\bar{x}$, and the sample variance is s^2, with a sample of size n from a normal distribution having mean μ and variance σ^2, the confidence limits are defined by

$$P(\bar{x} - t_{\alpha/2,n-1}s/\sqrt{n} \leq \mu \leq \bar{x} + t_{\alpha/2,n-1}s/\sqrt{n}) = 1 - \alpha,$$
$$\text{and } P((n-1)s^2/\chi^2_{\alpha/2,n-1} \leq \sigma^2 \leq (n-1)s^2/\chi_{1-\alpha/2,n-1}) = 1 - \alpha$$

where the upper tail probabilities of the t and chi-square distribution are used.

Example

The body temperature of 25 intertidal crabs was recorded in °C as follows: 24.3, 25.8, 24.6, 26.1, 22.9, 25.1, 27.3, 24.0, 24.5, 23.9, 26.2, 24.3, 24.6, 23.3, 25.5, 28.1, 24.8, 23.5, 26.3, 25.4, 25.5, 23.9, 27.0, 24.8, 22.9, 25.4. The sample mean, variance and standard deviation were $\bar{x} = 25.03$, $s^2 = 1.8$, and $s = 1.3416408$ leading to the following central confidence intervals for the mean and unsymmetrical confidence limits for the variance.

Sample size	Level	Parameter	Estimate	Interval
25	95%	Mean	25.03	$24.4762 \leq \mu \leq 25.5838$
25	99%	Mean	25.03	$24.2795 \leq \mu \leq 25.7805$
25	95%	Variance	1.8	$1.09745 \leq \sigma^2 \leq 3.48355$
25	99%	Variance	1.8	$0.948231 \leq \sigma^2 \leq 4.36971$

6.3 Numerical analysis

Most of the standard vector and matrix analysis techniques that are useful in data analysis are available.

6.3.1 Zeros of a polynomial

There are many situations when this technique is useful including estimating the zeros of binding polynomials with cooperative ligand binding studies.

6.3.2 Matrix determinant/eigenvalues/inverses

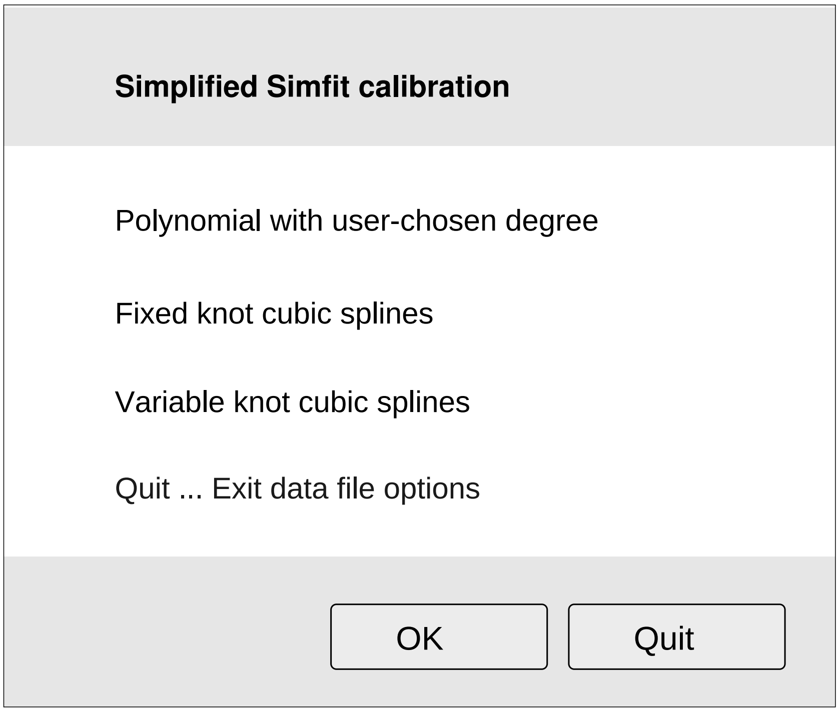
Many problems in data analysis result from data matrices that cannot be inverted, e.g., linear regression.

6.3.3 Singular value decomposition

This is very useful in multivariate analysis to determine the rank of a matrix accurately. Many statistical tests depend on the rank off the data matrix and, in instances where multivariate or regression analysis fails, the data matrix is not of full rank so that analysis cannot proceed. This deficiency in information content can be exposed by viewing the results of a singular value decomposition.

7 Calibration

The SimFIT calibration options are very extensive and the simplified version offers calibration using linear models in the linear regression options while the nonlinear regression options contain a number of specialised calibration options. This section describes programs that specifically deal with fitting flexible curves to data sets that are clearly nonlinear as shown next.


Simplified Simfit calibration
Polynomial with user-chosen degree
Fixed knot cubic splines
Variable knot cubic splines
Quit ... Exit data file options
OK
Quit

7.1 Details about calibration

Given n observations y_i at experimental values x_i the idea is to construct a reference curve $y = f(x)$ so that when further observations Y are available then corresponding estimates for X can be made made by calculating $X = f^{-1}(Y)$, i.e., by using the inverse function $f^{-1}(.)$. If it is required to avoid bias and/or estimate confidence limits on these predictions then several assumptions must be made.

1. The model $f(x)$ must be a good fit to the data
2. It must be possible to evaluate the inverse function f^{-1} explicitly or by iterative calculation.
3. If the variance of y is constant then weighting is not required
4. In general the variance of y increases as the absolute value of y increases and the fitting will be dominated by the larger values so that weighting is advised
5. Ideally the variance of y used to generate the standard curve should be determined independently otherwise replicates should added to the third column of the data file
6. There is no such thing as unweighted curve fitting as absence of weights indicates that the variance of y is constant, which is unlikely
7. If weights are calculated from the replicates this could be worse that no weighting unless the number of replicates at x_i is greater than or equal to 5

7.2 Using polynomials

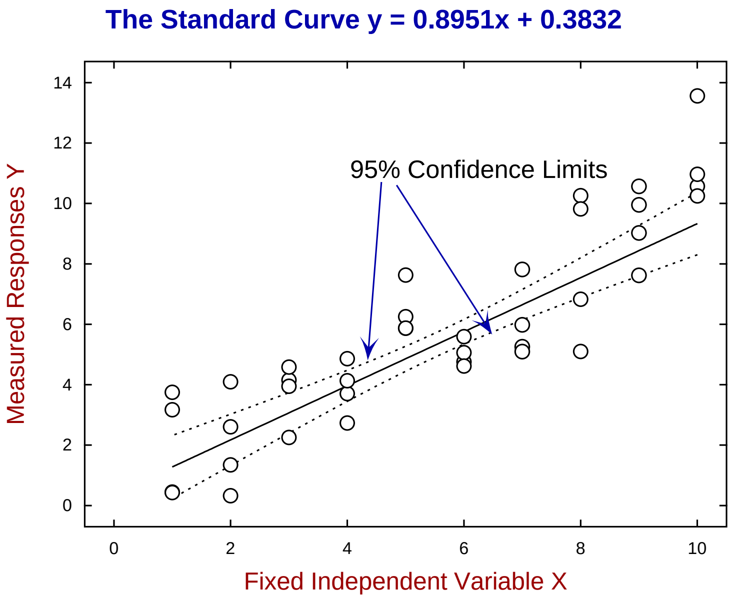
One of the most useful models for calibration is the polynomial of degree n defined as

$$y = p_0 + p_1 x + p_2 x^2 + \cdots + p_n x^n$$

as it deals with the special cases such as $p_i = 0$ for $i > 0$ (a constant), $p_i = 0$ for $i > 1$ (a straight line) , or $i = 0$ for $i > 2$ (a quadratic), etc.

7.3 Using a straight line

This can be done using SIMFIT program **polnom** but is most easily done using the linear regression options and lead to standard curves like the following.

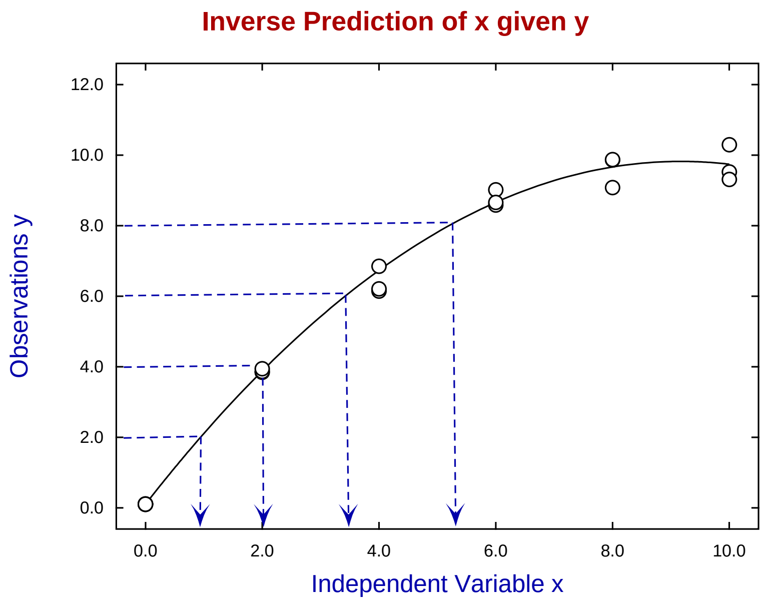


In the case of such a straight line with constant variance then 95% confidence limits on the best–fit line can be calculated as shown and, if it is also assumed that the value of Y input is free from error, then confidence limits can also be calculated for the values of X predicted. In all other situations biased predictions and error estimates will be calculated.

7.4 Using a quadratic

When using SIMFIT program **polnom** polynomials up to a maximum degree of 6 will be fitted, and for each degree then statistics are provided to choose the best–fit degree to use for calibration. If it is only wished to fit an arbitrary degree for smoothing this recommendation can be overridden but it should be remembered that a polynomial of degree n can have up to $n - 1$ turning points and cubic splines would be better.

However, as most calibration curves only have modest curvature the quadratic is often the best choice as it is often impossible to predict X given Y when there are multiple solutions for $n > 1$. Even a quadratic can have a turning point in the range of X provided and then the option is given to search upwards from the start of the data or backwards from the end of the data to obtain a unique solution. The case of a quadratic is shown next.



As an example consider the results when fitting a polynomial to the test file **polnom.tf1**.
The idea of this systematic procedure is to determine if there is statistical evidence to justify a trend line or progressive curvature in noisy data, or to select a model equation to use as a calibration curve for inverse prediction. To appreciate this aspect consider the following results tables when the data are analyzed.

Table 1: Degree fitted and Chebyshev coefficients

m	A_0	A_1	A_2	A_3	A_4	A_5
0	0.31113					
1	16.034	7.9080				
2	12.737	4.8194	-1.4456			
3	12.735	4.8132	-1.4591	-0.0083774		
4	12.762	4.8342	-1.4387	-0.055083	-0.059600	
5	12.654	4.6602	-1.3858	-0.087456	-0.035275	0.22979

Another table of statistics required to determine the degree of the polynomial required is also displayed as follows.

Table 2: Statistics to determine degree of the fitted polynomial

m	σ	%change	$WSSQ$	%change	$P(\chi^2 \geq WSSQ)$	5%	FV	$P(F \geq FV)$	5%
0	36.703		22901		0.0000	no			
1	8.0833	77.98	1045.4	95.44	0.0000	no	334.50	0.0000	yes
2	0.9914	87.73	14.744	98.59	0.4700	yes	1048.6	0.0000	yes
3	1.0253	3.42	14.718	0.18	0.3977	yes	0.0249	0.8769	no
4	1.0511	2.52	14.363	2.41	0.3488	yes	0.3213	0.5805	no
5	1.0000	4.87	11.999	16.46	0.4457	yes	2.3639	0.1501	no

Here m is the degree fitted, $\sigma = \sqrt{WSSQ/NDOF}$, and FV is the F value for assessing the significance of variance reduction by adding higher degree terms.

There are many results displayed in Tables 1 and 2 in order to suggest the highest degree that can be justified statistically. The qualitative conclusions do not use a Bonferroni correction, but the actual significance levels are also provided for purists. At this point SimFIT program **polnom** outputs the next table to aid decision.

Table 3: information to help you select a best-fit polynomial

Lowest degree where < 10% change in σ	2
Lowest degree where < 10% change in $WSSQ$	2
Lowest degree by chi-sq. at 5% significance level	2
Lowest degree by chi-sq. at 1% significance level	2
Lowest degree by F test at 5% significance level	2
Lowest degree by F test at 1% significance level	2

Accepting the recommendations of Table 3 leads to Table 4 for the best-fit quadratic.

Table 4: Results for weighted fitting $(w = 1/s^2)$

Parameter	Value	Std. error	Lower95%cl	Upper95%cl	p
θ_0	0.10347	0.0032091	0.096630	0.11031	0.0000
θ_1	2.1203	0.019731	2.0783	2.1624	0.0000
θ_2	-0.11565	0.0035714	-0.12326	-0.10803	0.0000

Correlation matrix

1		
-0.0960	1	
0.0516	-0.8432	1

If you selected to predict x from y the following warning is issued.

```
You must be very careful if you wish to use this best-fit
curve as a calibration curve for predicting x given y since
there are turning points for X_min ≤ x ≤ X_max as follows:
x-value                        y-value
9.1673                         9.8224
```

This is because the quadratic has a turning point within the range of the data, and so predicting x from y could be misleading if a horizontal line for $y = y_0$ for some y_0 intersected the best fit curve twice. So you have to choose whether to search upwards or downwards along the x axis for the prediction required. If a spurious prediction results you have to change the search order. For degrees greater than two there may be multiple turning points, so using degrees greater than two is not normally recommended for inverse prediction. Table 5 results from choosing to predict x from y along with 95% confidence ranges using the data supplied in test files `polnom.tf2` and `polnom.tf3` or typed in interactively.

Table 5: using a best-fit polynomial to predict x given y

Inverse prediction data for program **polnom** : y = 2, 4, 6, 8

y-measured	x-predicted	95% confidence limits
2.0	0.9429	0.9253, 0.9612
4.0	2.0718	2.0347, 2.1100
6.0	3.4182	3.3566, 3.4819
8.0	5.1976	5.0739, 5.3342

7.5 Using cubic splines

When fitting a line or polynomial fails there are four approaches that can be used.

1. Data transforming can be attempted.
   The most popular method is using logarithms then fitting a line or polynomial.

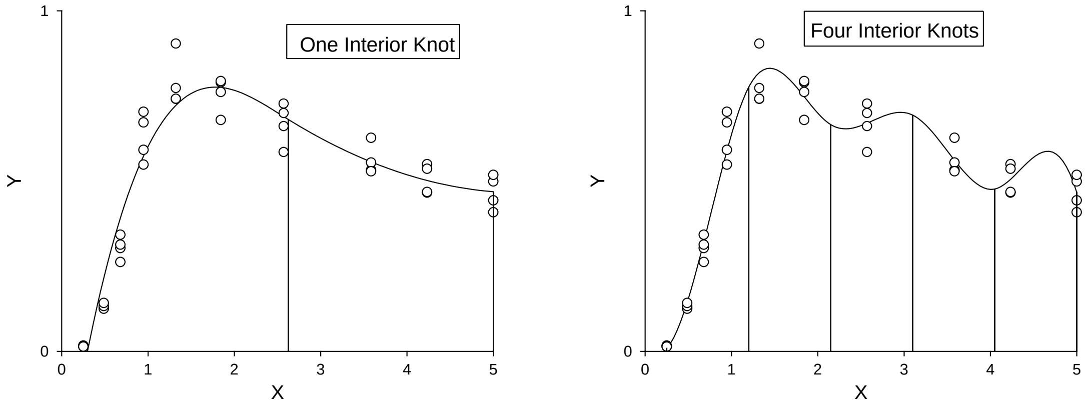
2. A chosen model can be used.
   This can be done by program **qnfit** using a model from the library or a user–defined model but requires considerable expertise in nonlinear optimisation.

3. Generalised Linear Modeling (GLM) can be used.
   This procedure is provided by SIMFIT but is seldom used except for estimating LD50 valuers where a simple interface is provided.

4. Using cubic splines.
   These can be used to fit a best–fit smooth curve to any data set but some care is required. SIMFIT provides program **calcurve** which is very successful and easy to use but it requires users to specify the number of knots required, that is, the points where separate cubic curves join up to create the spline. This program is very useful for titration curves in immunology and can perform an automatic logarithmic transform.

   There is another program **spline** that can automatically choose the number and spacing of points but this can also prove misleading unless the data and standard curve are chosen carefully.

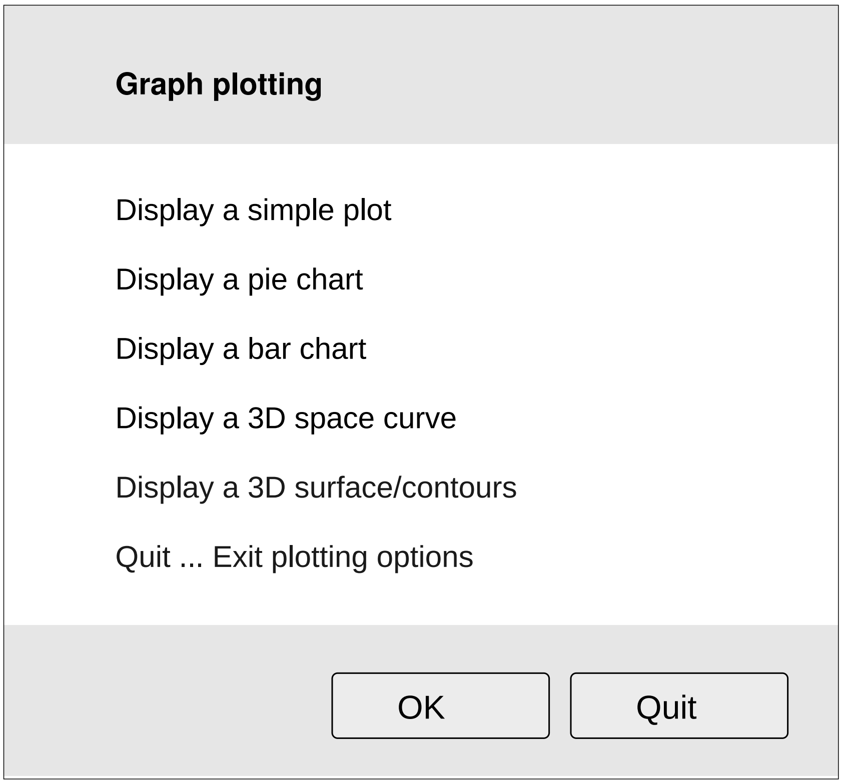
To summarise: spline curves are very flexible but the fit created depends on the method used and the number of interior knots and interior positions chosen where the cubic sections connect. Few knots give a smooth curve for calibration while too many knots can lead to over–fitting as in the next plots.



8 Graph plotting

8.1 Details about using the plotting options

Opening this section shows a restricted set of options from the SIMFIT program **simplot** as follows.



Each of these works in the same way as follows

- Plot the test file selected
- View the suggested test file and observe the format
- Read the trailer section of the test file to help you understand the format
- Note that some test files have `begin{labels} ... end{labels}`to label the plot
- Create your own file using the Data File option for file creation in the sv_simfit package
- Read in your own data file formatted as for the test data file
- Check if it displays properly

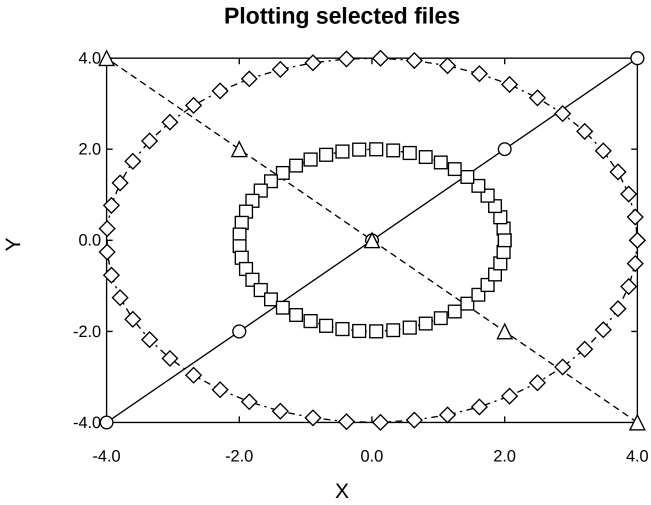
It should be noted that there two types of 2D graphics namely

1. Simple graphics only offers trivial editing but contain an option to proceed to Advanced Graphics
2. Advanced graphics provides a very comprehensive set of options for graph editing for advanced users

A right mouse click on any SIMFIT plot provides an explanation about Simple and Advanced Graphics. Note that advanced graphics allows the use of library files which allow multiple data files to be plotted together.

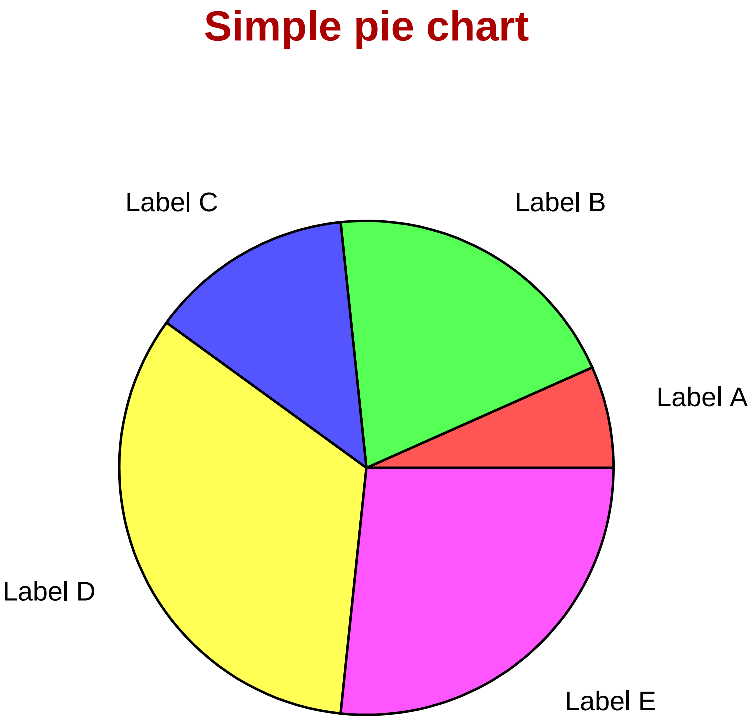
8.2 Plotting up to four data sets

After loading test files the following plot is created and the data can be selected, viewed or replaced as required.



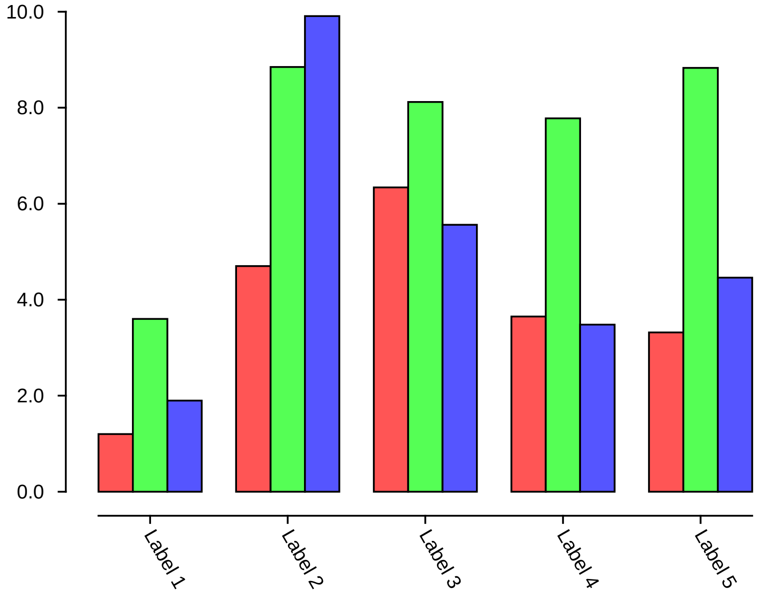
8.3 A pie chart

The test file **sv_plot.tf2** is a vector of non-negative numbers and the size of the segments is in proportion to these. The title is the title of the test file while labels are added as a `begin{labels} ... \end{labels}` section. If this is missing default labels will be added.



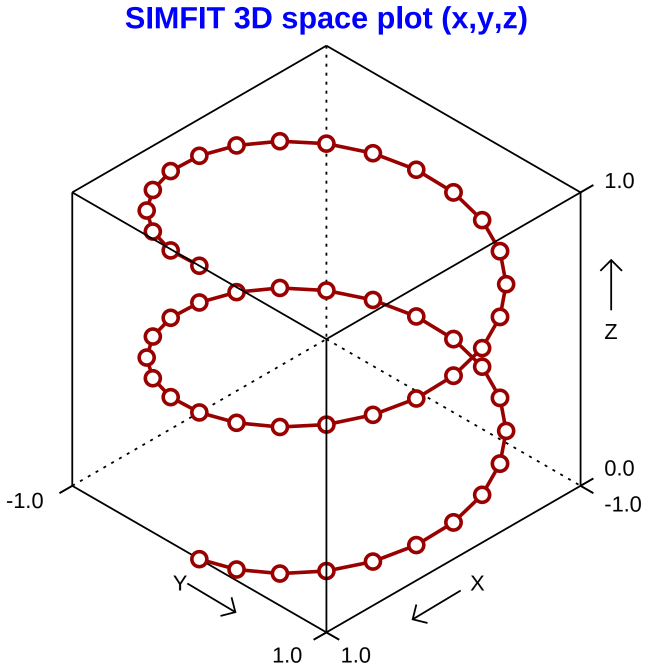
8.4 A bar chart

Test file **sv_plot.tf3** is just a rectangular matrix where each column contains the bar values for that group and labels for the groups are added as `begin{labels}... end{labels}`.



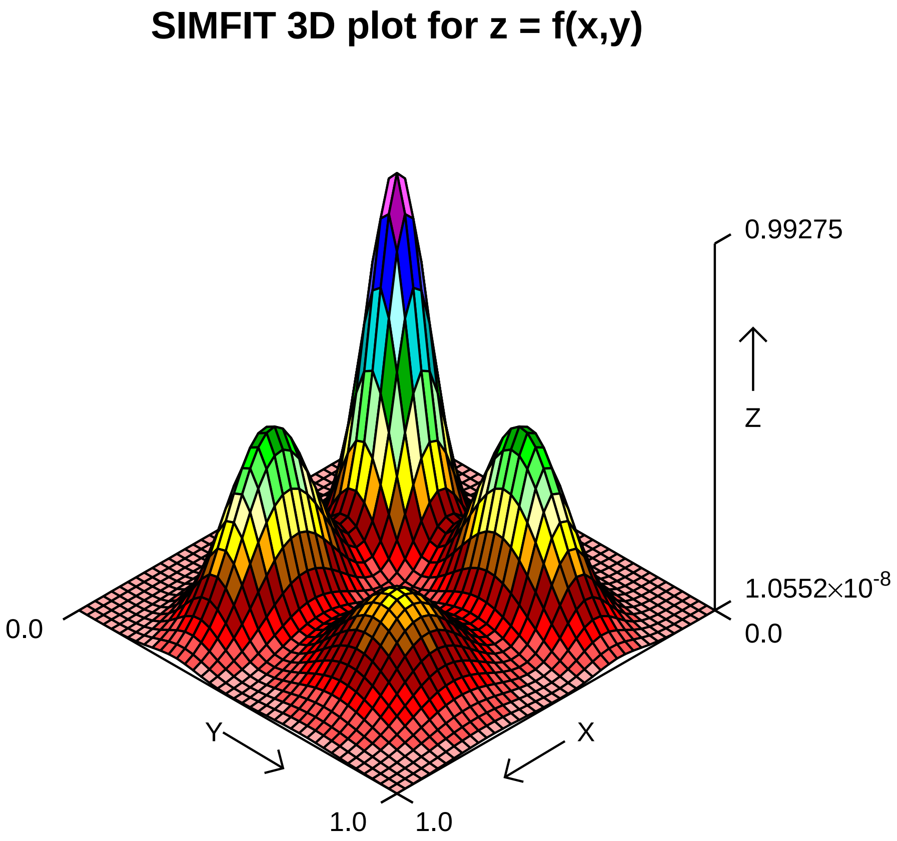
8.5 A space curve

Test file **sv_plot.tf4** is just a three column matrix where columns contains coordinates for x, y, z and colours, line widths, symbol types, etc. can be edited interactively.

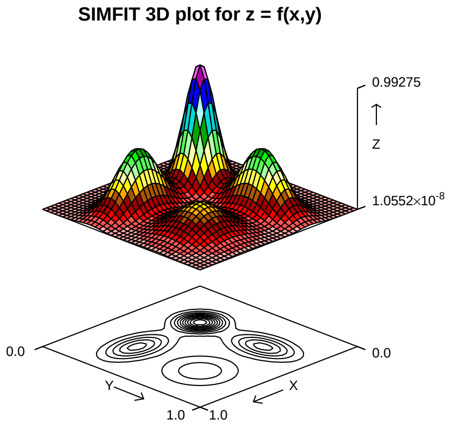


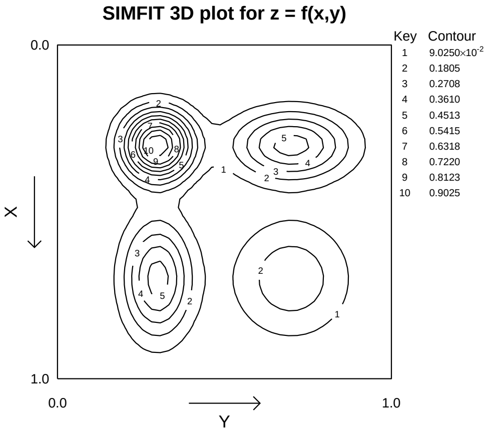
8.6 Surfaces and contours

Test file **sv_plot.tf5** displays this three dimensional surface.



The routine creating this surface also provides many options to display the surface in various ways and orientations. For instance, by simply choosing from menus the next figures show the surface with contours projected onto a two dimensional x, y subspace while the adjacent figure just just plots the contours in standard x, y space along with a table of $f(x, y)$ values corresponding to the contour values.


SIMFIT 3D plot for z = f(x,y)
Z
1.0552×10^-8
Y
X


SIMFIT 3D plot for z = f(x,y)
Key Contour
1 9.0250×10^-2
X
Y