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

It has been explained that all SIMFIT requires to perform analysis is a rectangular table of numbers, with or without row and column labels. So why is there a need for data files formatted according to the SIMFIT convention? There are two answers to this question.

1. It is very easy to create data files according to the SIMFIT convention using any text editor, the programs supplied by SIMFIT, or the Excel macro **simfit6.xls**. Such files facilitate archiving and repeated analysis.
2. Many procedures used by SIMFIT require additional information such as starting values and ranges of parameter values for curve fitting, initial conditions from which to advance the solution of differential equations, flags to indicate which variables are to be included in multivariate analysis, etc.

Example 1

Consider a very simple example, namely the SIMFIT default ANOVA test file `anova1.tf1` shown below.

```
1-way ANOVA data from Zar: Biostatistics 3rd. edn. p-213
     6          5
  28.2       39.6       46.3       41.0       56.3
  33.2       40.8       42.1       44.1       54.1
  36.4       37.9       43.5       46.4       59.4
  34.6       37.1       48.8       40.2       62.7
  29.1       43.6       43.7       38.6       60.0
  31.0       42.4       40.1       36.3       57.3
     5
line 1: title for this data set
line 2: number of rows then number of columns
line 3: first row of data values
line 8: last row of data values
line 9: number of additional comment lines in the file
```

This data file has three sections as follows.

1. **The Header**
   The first line is the title and this is very useful as many SIMFIT procedures output results tables with the titles to identify the data set. It is also very convenient to scan the first line of a data file to quickly remind you about the contents. The second line is the size in the form of the number of rows (6 in this case) and the number of columns (5 in this case). There are some SIMFIT functions that must have these two dimensions in order to make decisions about the type of data. For instance, some graphics and curve fitting procedures.
   Note that, although a header section is not always required, it is very useful to supply one.
2. **The Data**
   This is just the rectangular table of data values with no missing values.
3. **The Trailer**
   In this example the first line of the trailer has the number of extra lines appended to the data. This value is not always necessary but is useful for some SIMFIT programs that edit data files. Also note that, in this case, the only material contained in the trailer section is advisory information. However this is not always the case. Although a trailer section is never vital and can always be omitted, nevertheless there are many circumstances when extremely important information required by SIMFIT can be conveniently added to the trailer which greatly simplifies analysis. This will be clear after analyzing another test file.

Example 2

Now consider another typical SimFIT data file, namely `kmeans.tf1`, the default file to illustrate K-means clustering. Note that line numbers have been included in the first column of the following table for reference only and are not part of the actual data file.

Line 1	Data for 5 variables on 20 soils ...				
Line 2	20	5			
Line 3	77.3	13.0	9.7	1.5	6.4
Line 4	82.5	10.0	7.5	1.5	6.5
Line 5	66.9	20.6	12.5	2.3	7.0
Line 6	47.2	33.8	19.0	2.8	5.8
Line 7	65.3	20.5	14.2	1.9	6.9
Line 8	83.3	10.0	6.7	2.2	7.0
Line 9	81.6	12.7	5.7	2.9	6.7
Line 10	47.8	36.5	15.7	2.3	7.2
Line 11	48.6	37.1	14.3	2.1	7.2
Line 12	61.6	25.5	12.9	1.9	7.3
Line 13	58.6	26.5	14.9	2.4	6.7
Line 14	69.3	22.3	8.4	4.0	7.0
Line 15	61.8	30.8	7.4	2.7	6.4
Line 16	67.7	25.3	7.0	4.8	7.3
Line 17	57.2	31.2	11.6	2.4	6.5
Line 18	67.2	22.7	10.1	3.3	6.2
Line 19	59.2	31.2	9.6	2.4	6.0
Line 20	80.2	13.2	6.6	2.0	5.8
Line 21	82.2	11.1	6.7	2.2	7.2
Line 22	69.7	20.7	9.6	3.1	5.9

Lines numbered 1 to 2 are the optional header while lines 3 to 22 contain the data table as follows.

Line 1 This is the title of the data set

Line 2 This contains the size, i.e. the number of rows and columns

Line 3 to **Line 22** contain the 20 by 5 data table

Line 23	44				
Line 24	Usage:				
Line 25	Select statistics, then run program simstat, choose				
Line 26	multivariate statistics, then go to K-means clustering				
Line 27					
Line 28	The next line defines the starting clusters for k = 3				
Line 29	begin{values} <– token to flag start of appended values				
Line 30	82.5	10.0	7.5	1.5	6.5
Line 31	47.8	36.5	15.7	2.3	7.2
Line 32	67.2	22.7	10.1	3.3	6.2
Line 33	end{values}				

Lines 23 to 28 are simply advisory but lines 29 to 33 illustrates the technique to set default starting estimates for the K-means clusters centroids. Note here how SimFIT data files and user-supplied model files define various environments (in this case the environment is values) using flags as in

begin{values} ... end{values}

The final part of the trailer section contains lines 35 to 67 as follows.

Line 35	The next line defines the variables as 1=include, 0=suppress
Line 36	begin{indicators} <– token to flag start of indicators
Line 37	1 1 1 1 1
Line 38	end{indicators}
Line 39	
Line 40	The next line defines the row labels for plotting
Line 41	begin{labels} <– token to flag start of row and column labels
Line 42	A
line 43	B
Line 44	C
Line 45	D
Line 46	E
Line 47	F
Line 48	G
Line 49	H
Line 50	I
Line 51	J
Line 52	K
Line 53	L
Line 54	M
Line 55	N
Line 56	O
Line 57	P
Line 58	Q
Line 59	R
Line 60	S
Line 61	T
Line 62	V1
Line 63	V2
Line 64	V3
Line 65	V4
Line 66	V5
Line 67	end{labels}

Here we see that two further environments are defined.

1. **indicators**
   Lines 36 to 38 define the indicators, i.e. which variables are to be included in the analysis using the scheme that a 1 indicates a variable to be included while a 0 indicates a variable to be suppressed. So here the default position is to include all variables.

2. **labels**
   Lines 41 to 67 define the labels. First the row labels in lines 42 to 61 then the column labels in lines 62 to 66. These labels can then be used to identify rows and/or columns when graphs are plotted. It is recommended to use very short labels, as done here, to avoid confusion resulting from long labels.

Note that environments defining parameters such as values, indicators, and labels as illustrated in this test file can be placed anywhere in the trailer section. SimFIT simply scans the trailer section of data files for appropriate environments and, if none are found, it uses default settings which can be edited retrospectively as required. However it should be pointed out that with many advanced SimFIT techniques, such as constrained nonlinear regression or simulating and fitting differential equations, supplying starting estimates or initial conditions is very much easier if these are appended to the individual data sets.