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

Multilinear regression is resorted to in situations where the value of a variable Y is believed to depend on one or more fixed variables X, but it has not proved possible to develop a mathematical model based on scientific principles. Usually the variation in Y due to experimental error or sampling variation is considerably greater than the variation in X, and in addition variables X are assumed to be uncorrelated, so that Y can be regarded as a dependent variable, and X as independent variables.

From the main SimFIT menu choose the [A/Z] option, open program **linfit**, choose [multilinear regression] using least squares, then browse the default test file `linfit.tf2` which contains the following data set.

x_1	x_2	x_3	x_4	y	s
7.00	26.0	6.00	60.0	78.50	1
1.00	29.0	15.0	52.0	74.30	1
11.0	56.0	8.00	20.0	104.3	1
11.0	31.0	8.00	47.0	87.60	1
7.00	52.0	6.00	33.0	95.90	1
11.0	55.0	9.00	22.0	109.2	1
3.00	71.0	17.0	6.00	102.7	1
1.00	31.0	22.0	44.0	72.50	1
2.00	54.0	18.0	22.0	93.10	1
21.0	47.0	4.00	26.0	115.9	1
1.00	40.0	23.0	34.0	83.80	1
11.0	66.0	9.00	12.0	113.3	1
10.0	68.0	8.00	12.0	109.4	1

The columns have the following meanings.

1. Column 1: % tricalcium aluminate
2. Column 2: % tricalcium silicate
3. Column 3: % tetracalcium alumino ferrite
4. Column 4: % dicalcium silicate
5. Column 5: Heat evolved in calories per gram of cement
6. Column 5: weighting factor.

Note that the weighting factor s must be supplied as the last column so that SimFIT knows how many variables are present. It is usual to set all the values of s to one as in the above example, but if accurate estimates for the standard deviations of Y are known these could be used so that weighted least squares fitting can be done.

To conclude: if the data set supplied has k columns, then it will be presumed that there are $k-2$ independent variables X in columns $1, 2, \ldots, k-2$, the dependent variable Y is in column $k-1$, and the weighting factors are in column k. If these are all equal to one then unweighted regression will be carried out, but otherwise the values s_i will be assumed to be standard errors for the y_i and weighted regression will be performed using $w_i = 1/s_i^2$. Setting $s = 0$ suppresses corresponding rows of data but this is not recommended.

Analysis of these data then leads to the following tables of results using this model

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{k-2} x_{k-2}$$

where $x_0 = 1$ if β_0 is to be estimated and a constant term is required, or $\beta_0 = 0$ otherwise.

Table 1: Parameter estimates

Number of parameters: 5, Rank: 5, Number of points: 13, Degrees of freedom: 8

Residual-SSQ: 47.864, Mallows' C_p: 5.0, R^2: 0.9824

Parameter	Value	Lower95%cl	Upper95%cl	Std. Error	p	
β_0 (Constant)	62.405	-99.179	223.99	70.071	0.3991	***
β_1	1.5511	-0.16634	3.2685	0.74477	0.0708	*
β_2	0.51017	-1.1589	2.1792	0.72379	0.5009	***
β_3	0.10191	-1.6385	1.8423	0.75471	0.8959	***
β_4	-0.14406	-1.7791	1.4910	0.70905	0.8441	***

The stars shown against the parameter estimates in Table 1 are displayed when the parameter estimates are not significantly different from zero, so this table indicates that none of the five parameters were well determined.

Table 2: Residuals

Number	y-value	Theory	Residual	Leverage	Studentized
1	78.500	78.495	0.0047604	0.55028	0.0029021
2	74.300	72.789	1.5112	0.33324	0.75662
3	104.30	105.97	-1.6709	0.57694	-1.0503
4	87.600	89.327	-1.7271	0.29524	-0.84108
5	95.900	95.649	0.25076	0.35760	0.12791
6	109.20	105.27	3.9254	0.12416	1.7148
7	102.70	104.15	-1.4487	0.36708	-0.74445
8	72.500	75.675	-3.1750	0.40854	-1.6878
9	93.100	91.722	1.3783	0.29431	0.67080
10	115.90	115.62	0.28155	0.70040	0.21029
11	83.800	81.809	1.9910	0.42551	1.0739
12	113.30	112.33	0.97299	0.26298	0.46335
13	109.40	111.69	-2.2943	0.30372	-1.1241

However Table 2 does show a good scatter of residuals about zero with no particular bias or runs indicated.

Table 3: Analysis of Variance

Source	NDOF	SSQ	Mean SSQ	F-value	p
Total	12	2715.8			
Regression	4	2667.9	666.97	111.48	0.0000
Residual	8	47.864	5.9830		

Table 3 is used in much the same way as for simple linear regression and is defined using $\hat{y}_i$ for the best-fit model as follows.

$$SSQ_{total} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSQ_{regression} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$$SSQ_{residual} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The F value is the ratio of mean regression SSQ to mean residual SSQ, and the significance level p is used to test the null hypothesis

$$H_0 : \beta_i = 0 \text{ for all } i$$

against the alternative hypothesis

$$H_A : \beta_i \neq 0 \text{ for one or more } i.$$

Clearly it must be concluded that, although none of the individual parameters were well determined as judged by the t tests on the ratios of estimates to standard errors, there is a significant overall reduction in the sum of squares by some combinations of variables.

At this point it is customary to see if a satisfactory regression could be achieved with fewer parameters, and a variety of techniques are available to perform such subset regression to find the best explanation of the data in terms of the smallest number of variables. When this is done systematically with large data sets it generates an enormous amount of analysis, which is not normally justified because usually the experimentalist would have a good idea which subsets of variables to try. This is fairly easy to do interactively in SIMFIT by suppressing variables until a fit is achieved where all the parameters are significantly different from zero with the two-tail t test, and the fit is justified by the C_p values.

The way to interpret the Mallows' C_p values should be explained. Program **linfit** first fits a full model and the results from this analysis are saved. This fit is assumed to be the best possible for estimating the variance and the effect of suppressing any variable can then be seen by comparing the effect on the C_p value. From fitting the full model the C_p value will be equal to the total number of parameters and subsequent subset regressions can be judged by the ratio of the C_p values to the number of parameters, where values much greater than the number of parameters estimated suggest a deficient model.

The next table summarizes the results from fitting a constant only, followed by fitting subsets of the additional 1, 2, and 3 variables.

Table 4: C_p values

Variables	C_p	Parameters
Constant only	442.9	1
+1,+2,+3,+4	202.5, 142.5, 315.2, 138.7	2
+12, +13, +14	2.7, 198.1, 5.5	3
+23, +24, +34	62.4, 138.2, 22.4	3
+123, +124, +134, +234	3.0, 3.0, 3.5, 7.3	4
+1234	5.0	5

In Table 4 the first column indicates the subscripts of the variables added to the constant term, column 2 holds the corresponding C_p variables, while column 3 contains the total number of parameters varied including the constant term. Values where C_p divided by the number of parameters in the regression are less than or equal to one are highlighted, and it is perfectly clear that the combination of a constant plus variables 1 and 2 seems to be strongly recommended.

Here, for example, are the results from suppressing variables 3 and 4.

Table 1A: parameter estimates with variables 3 and 4 suppressed
Number of parameters: 3, Rank: 3, Number of points: 13, Degrees of freedom: 10
Residual-SSQ: 57.904, Mallows' C_p: 2.6782, R^2: 0.9787

Parameter	Value	Lower95%cl	Upper95%cl	Std. Error	p
β_0 (Constant)	52.577	47.483	57.671	2.2862	0.0000
β_1	1.4683	1.1980	1.7386	0.12130	0.0000
β_2	0.66225	0.56008	0.76442	0.045855	0.0000

Table 2A: Residuals with variables 3 and 4 suppressed

Number	y-value	Theory	Residual	Leverage	Studentized
1	78.500	80.074	-1.5740	0.25119	-0.75590
2	74.300	73.251	1.0491	0.26189	0.50745
3	104.30	105.81	-1.5147	0.11890	-0.67061
4	87.600	89.258	-1.6585	0.24225	-0.79175
5	95.900	97.293	-1.3925	0.83616	-0.60451
6	109.20	105.15	4.0475	0.11512	1.7881
7	102.70	104.00	-1.3021	0.36180	-0.67732
8	72.500	74.575	-2.0754	0.24119	-0.99011
9	93.100	91.275	1.8245	0.17915	0.83687
10	115.90	114.54	1.3625	0.55002	0.84405
11	83.800	80.536	3.2643	0.18402	1.5018
12	113.30	112.44	0.86276	0.19666	0.40002
13	109.40	112.29	-2.8934	0.21420	-1.3564

Table 3A: Analysis of Variance with variables 3 and 4 suppressed

Source	NDOF	SSQ	Mean SSQ	F-value	p
Total	12	2715.8			
Regression	2	2657.9	1328.9	229.50	0.0000
Residual	10	57.904	5.7904		

Comparing the results with all variables present to those with variables 3 and 4 suppressed leads to the following conclusions.

1. **Table 1 compared to Table 1A**
   With all variables present no parameters were well-determined by a two-tail t test, but with variables 3 and 4 suppressed all parameters were well-determined.

2. **Table 2 compared to Table 2A**
   There are no differences to indicate a poorer fit with the simpler model.

3. **Table 3 compared to Table 3A**
   There are no differences to indicate a poorer fit with the simpler model.

Sometimes it is useful to evaluate the best-fit model and, as variables 3 and 4 do not seem to be making an important contribution prediction, this will be demonstrated using the model with variables 3 and 4 suppressed. A vector of default values equal to 1 is supplied and this can be edited interactively to change the variables as follows with variable 2.

Using the best-fit model to predict y given x
x_0 = 1.0, coefficient = 52.577 (the constant term)
x_1 = 1.0, coefficient = 1.4683
x_2 = 1.0, coefficient = 0.66225
$y(x)$ = 54.708

x_0 = 1.0, coefficient = 52.577 (the constant term)
x_1 = 1.0, coefficient = 1.4683
x_2 = 5.0, coefficient = 0.66225
$y(x)$ = 57.357
Of course, users cannot alter the value of x_0 which is always equal to 1, and included or excluded from the regression depending on whether a constant (i.e. intercept) term is included.

Note that, for more advanced analysis of such data sets (including prediction and inverse prediction) the SIMFIT partial least squares procedure should be used.

Theory

Program **linfit** fits a multilinear model in the form

$$y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m,$$

where $x_0 = 1$, but you can choose interactively whether or not to include a constant term β_0, you can decide which variables are to be included, and you can use a weighting scheme if this is required. For each regression sub-set, you can observe the parameter estimates and standard errors, R-squared, Mallows C_p , and $ANOVA$ table, to help you decide which combinations of variables are the most significant. Unlike nonlinear regression, multilinear regression, is based on the assumptions

$$\begin{aligned} Y &= X\beta + \epsilon, \\ \mathrm{E}(\epsilon) &= 0, \\ \mathrm{Var}(\epsilon) &= \sigma^2 I, \end{aligned}$$

where X is the over-determined data matrix (e.g., the 13 rows and first 4 columns of test file `linfit.tf2`), Y is the observation vector (e.g., column 5 of test file `linfit.tf2`), β is the parameter vector and ϵ is the error vector. This allows us to introduce the hat matrix

$$H = X(X^T X)^{-1} X^T,$$

then define the leverages h_{ii}, which can be used to asses influence, and the studentized residuals

$$R_i = \frac{r_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

which may offer some advantages over ordinary residuals r_i for goodness of fit assessment from residuals plots. In the event of weighting being required, Y, X and ϵ above are simply replaced by $W^{\frac{1}{2}}Y$, $W^{\frac{1}{2}}X$, and $W^{\frac{1}{2}}\epsilon$, where W is the diagonal weighting matrix.

Note that examining parameter reliability using the t test as in Tables 1 and 1A and also model discrimination analysis using the F test is applicable for nested linear models as fitted by SimFIT program **linfit**. So several additional options are provide by **linfit** to perform such further investigations. For instance, to perform an F test for excess variance note that, 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. In the **linfit** results tables C_p values refer to the full model being fitted as the reference case.

Finally it should be noted that, for successful analysis of data, the units used to provide values of X should be such that the numerical values are of similar size. If categorical data are mixed with continuous data, or the data set is ill-conditioned, or less than full rank for any reason, a linear model will still be fitted using the singular value decomposition. However, in such cases **linfit** will issue a warning that the estimated parameters are not independent, and the data should be re-scaled, or the number or variables should be reduced until the columns of the X matrix are independent, i.e. the rank is at least as large as the number of parameters estimated.