Partial Least Squares

Summary

The **Partial Least Squares (PLS)** procedure is designed to construct a statistical model relating multiple independent variables X to multiple dependent variables Y. The procedure is most helpful when there are many factors and the primary goal is prediction of the response variables. PLS is widely used by chemical engineers and chemometricians for spectrometric calibration.

Sample StatFolio: *pls.sgp*

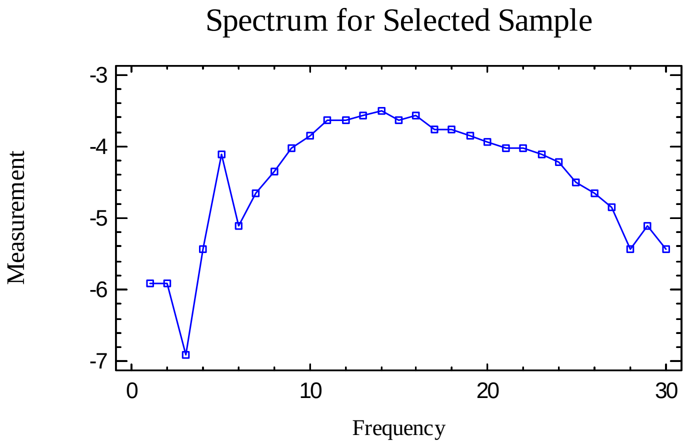
Sample Data:

The file *spectra.sgd* contains the observed spectra for *n* = 33 samples containing known concentrations of two amino acids, tyrosine and trytophan. The spectra are measured at *k* = 30 frequencies. A portion of the data, from McAvoy et al. (1989), is shown below:

Sample	*Tryptophan*	*Tyrosine*	*f1*	*f2*	*f3*	*f4*	*f5*
17mix35	0.00003000	0.00000001	-6.215	-5.809	-5.114	-3.963	-2.897
19mix35	0.00002970	0.00000030	-5.516	-5.294	-4.823	-3.858	-2.827
21mix35	0.00002925	0.00000075	-5.519	-5.294	-4.501	-3.863	-2.827
23mix35	0.00002850	0.00000150	-5.294	-4.705	-4.262	-3.605	-2.726
25mix35	0.00002700	0.00000300	-4.600	-4.069	-3.764	-3.262	-2.598
27mix35	0.00002250	0.00000750	-3.812	-3.376	-3.026	-2.726	-2.249
29mix35	0.00001500	0.00001500	-3.053	-2.641	-2.382	-2.194	-1.977
28mix35	0.00000750	0.00002250	-2.626	-2.248	-2.004	-1.839	-1.742
26mix35	0.00000300	0.00002700	-2.370	-1.990	-1.754	-1.624	-1.560
24mix35	0.00000150	0.00002850	-2.326	-1.952	-1.702	-1.583	-1.507

The leftmost column identifies each sample. The next 2 columns are known concentrations of the amino acids. The remaining 30 columns contain the measured spectra. Note: concentrations originally equal to 0 have been set to 1.0E-8 so that logarithmic transformations may be taken.

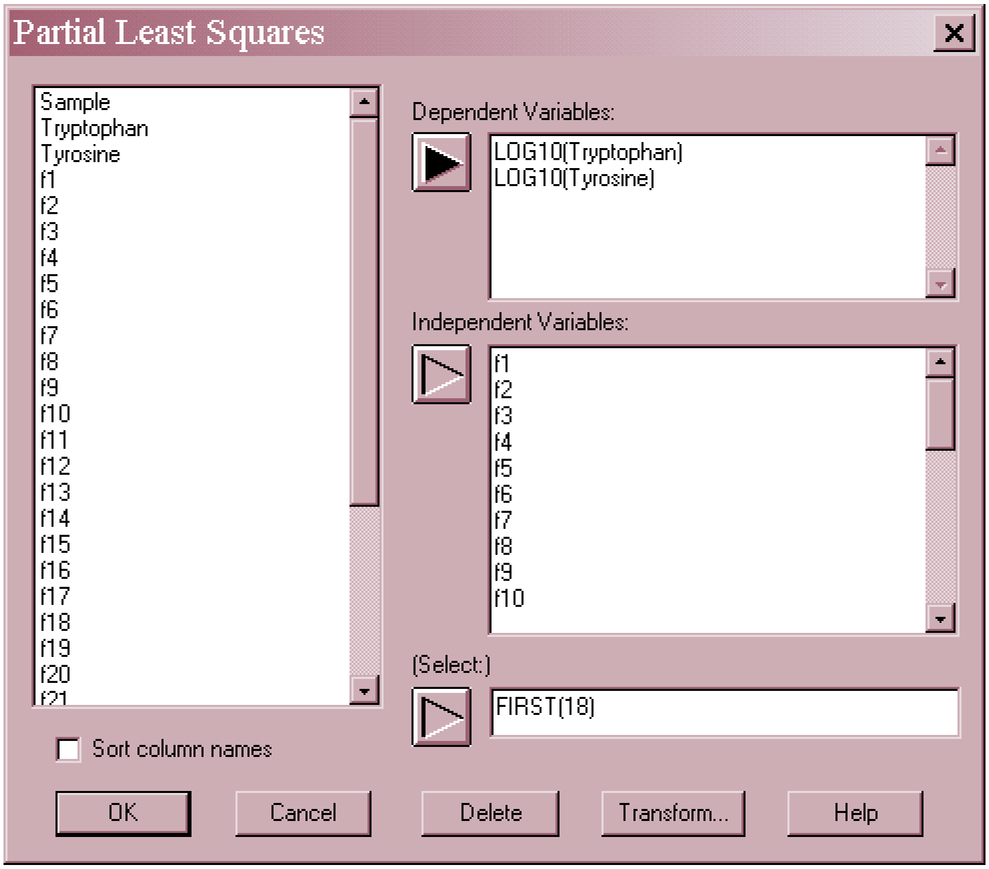
The observed spectrum for a typical sample is shown below:



The first 18 samples will be used as a training set to estimate a predictive model. The model will then be tested on the remaining 15 samples.

Data Input

The data input dialog box requests the names of the columns containing the dependent variables *Y* and the independent variables *X*:



Dependent Variables:
LOG10(Tryptophan)
LOG10(Tyrosine)

Independent Variables:
f1
f2
f3
f4
f5
f6
f7
f8
f9
f10

(Select:)
FIRST(18)

- **Y:** one or more numeric columns containing the *n* observations for the dependent variables *Y*. Either column names or STATGRAPHICS expressions may be entered.

- **X:** one or numeric columns containing the *n* values for the independent variables *X*.

- **Select:** subset selection. Rows selected will be used as the training set. Rows not selected may be used as a test set to validate the fitted model.

In the example, base 10 logarithms of the concentrations have been used to create 2 dependent variables. All 30 frequencies have been entered in the *Independent Variables* field. The entry in the *Select* field will cause the first 18 rows to be used as the training set.

Statistical Model

As in multiple regression, the goal of PLS is to construct a linear model of the form

$$Y = X\beta + E \qquad (1)$$

where *Y* is an *n* by *m* matrix containing the *n* standardized values of the *m* dependent variables, X is an *n* by *p* matrix containing the standardized values of the *p* predictor variables, β is a *p* by *m* matrix of model parameters, and *E* is an *n* by *m* matrix of errors. Unlike multiple regression, the number of observations *n* may be less than the number of independent variables *p*.

Rather than estimating β directly, however, c components are first extracted. The coefficients are then calculated from the product of two matrices:

$$\beta = WQ \qquad (2)$$

where *W* is a *p* by *c* matrix of weights that transform *X* into a matrix of factor scores *T* according to

$$T = XW \qquad (3)$$

and *Q* is a matrix of regression coefficients (loadings) that express the dependence between *Y* and the factor scores:

$$Y = TQ + E \qquad (4)$$

The matrix of independent variables can also be represented in terms of a *c* by *p* factor loading matrix P as

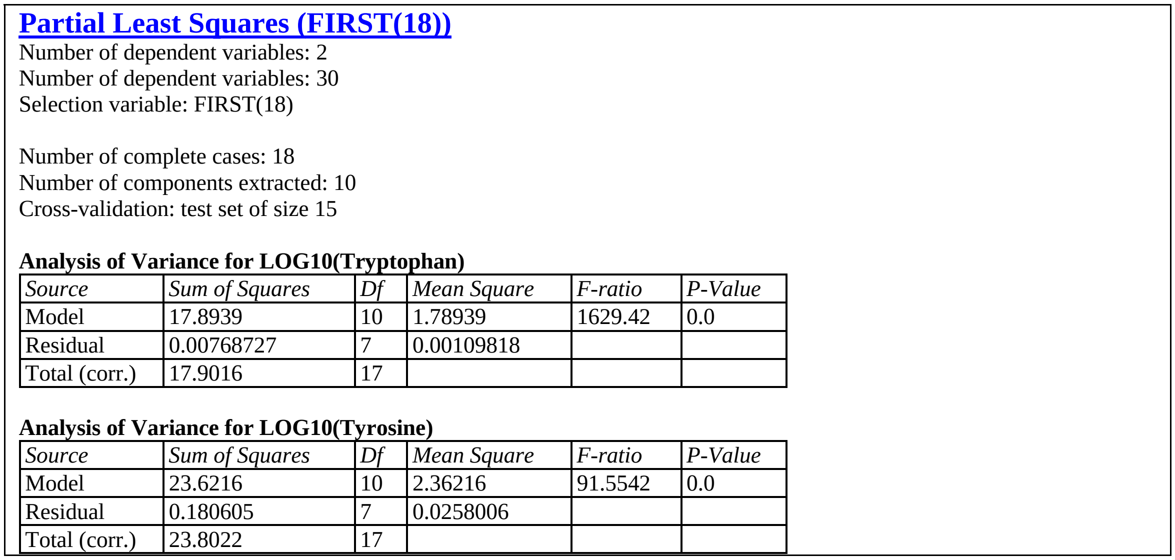
$$X = TP + F \qquad (5)$$

where *F* is an *n* by *p* matrix of deviations.

Part of the task in performing a PLS analysis is determining the proper number of components *c*. If *c* is set too low or too high, the model may not give good predictions for future observations.

Analysis Summary

The *Analysis Summary* shows information about the fitted model. The top section of the output summarizes the input data and displays an analysis of variance for each dependent variable.



Partial Least Squares (FIRST(18))

Number of dependent variables: 2
Number of dependent variables: 30
Selection variable: FIRST(18)

Number of complete cases: 18
Number of components extracted: 10
Cross-validation: test set of size 15

Analysis of Variance for LOG10(Tryptophan)

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-ratio*	*P-Value*
Model	17.8939	10	1.78939	1629.42	0.0
Residual	0.00768727	7	0.00109818		
Total (corr.)	17.9016	17			

Analysis of Variance for LOG10(Tyrosine)

Source	*Sum of Squares*	*Df*	*Mean Square*	*F-ratio*	*P-Value*
Model	23.6216	10	2.36216	91.5542	0.0
Residual	0.180605	7	0.0258006		
Total (corr.)	23.8022	17			

Included in the output are:

- **Variable Summary:** an indication of the number of X variables (p) and the number of Y variables (m).

- **Number of Complete Cases:** the number of observations n in the training set.

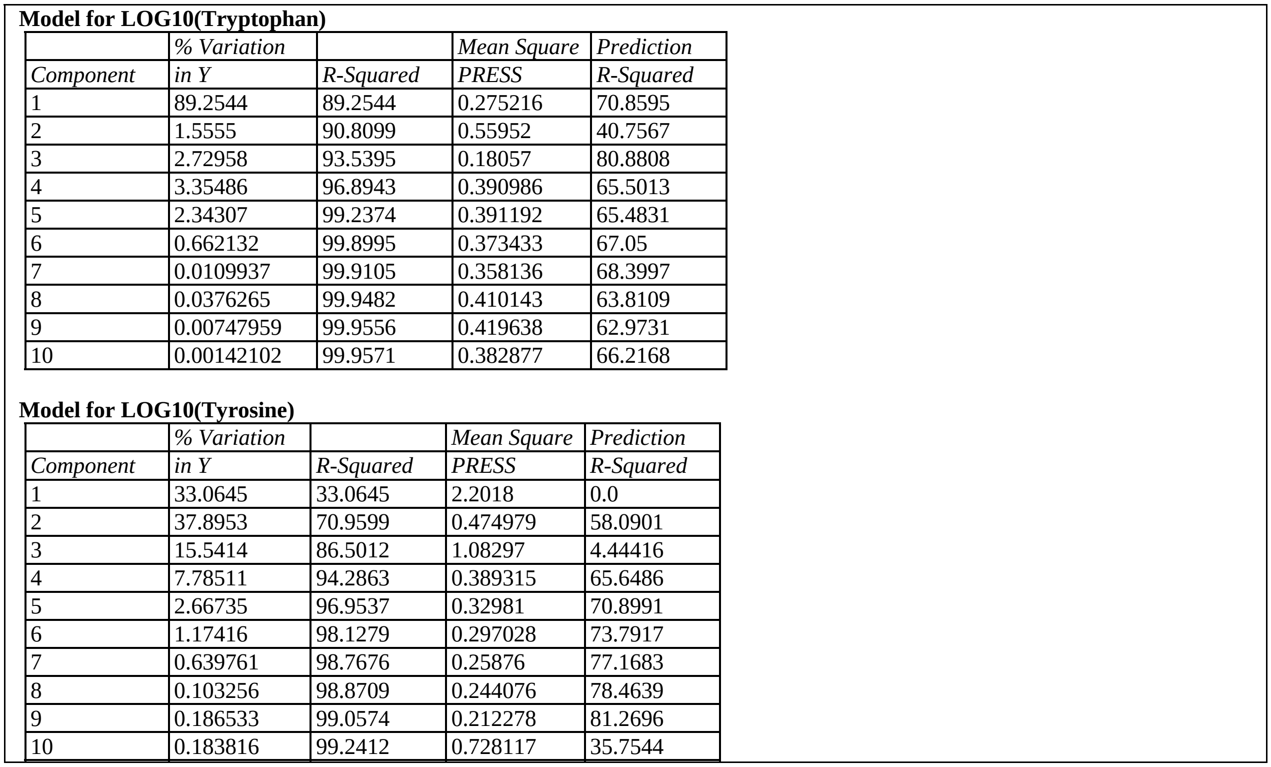
- **Number of Components Extracted**: the number of components c used to fit the model. c may not be greater than the smaller of p and (n – 1).

- **Cross-validation**: the method used to validate the predictive model. Depending on *Analysis Options*, an internal or external test set may be used to help select the number of components.

- **Analysis of Variance**: an ANOVA table for each of the dependent variables. Small P-values (less than 0.05 if operating at the 5% significance level) indicate that the model is statistically significant.

In the example above, 10 components have been extracted. The resulting models are significant predictors for the concentration of both amino acids, since both P-values are extremely small.

The second part of the output illustrates the usefulness models with different numbers of components:



Model for LOG10(Tryptophan)

	% Variation		*Mean Square*	*Prediction*
Component	*in Y*	*R-Squared*	*PRESS*	*R-Squared*
1	89.2544	89.2544	0.275216	70.8595
2	1.5555	90.8099	0.55952	40.7567
3	2.72958	93.5395	0.18057	80.8808
4	3.35486	96.8943	0.390986	65.5013
5	2.34307	99.2374	0.391192	65.4831
6	0.662132	99.8995	0.373433	67.05
7	0.0109937	99.9105	0.358136	68.3997
8	0.0376265	99.9482	0.410143	63.8109
9	0.00747959	99.9556	0.419638	62.9731
10	0.00142102	99.9571	0.382877	66.2168

Model for LOG10(Tyrosine)

	% Variation		*Mean Square*	*Prediction*
Component	*in Y*	*R-Squared*	*PRESS*	*R-Squared*
1	33.0645	33.0645	2.2018	0.0
2	37.8953	70.9599	0.474979	58.0901
3	15.5414	86.5012	1.08297	4.44416
4	7.78511	94.2863	0.389315	65.6486
5	2.66735	96.9537	0.32981	70.8991
6	1.17416	98.1279	0.297028	73.7917
7	0.639761	98.7676	0.25876	77.1683
8	0.103256	98.8709	0.244076	78.4639
9	0.186533	99.0574	0.212278	81.2696
10	0.183816	99.2412	0.728117	35.7544

For each dependent variable, the tables show:

- **% Variation in Y**: the percentage of the total corrected sum of squares for the training set explained by each component as it is added to the fit.

- **R-Squared:** the cumulative percentage of the total variation explained by models with the indicated number of components, on a scale of 0% to 100%.

- **Mean Square PRESS**: the average prediction sum of squares, calculated from the cross-validation test set. This statistic is comparable to the residual mean square in the ANOVA table, except that it is calculated from predictions for observations when they are not used to fit the model. When selecting the number of components to extract, you should look for a model with a small mean square PRESS.

- **Prediction R-Squared**: ratio of the *Mean Square PRESS* for the indicated number of components to the value when a model is fit with only a constant term. High values indicate good models.

The *Prediction R-Squared* peaks for *LOG10(Tryptophan)* at 3 components, and for *LOG10(Tyrosine)* at 9 components.

The last section of the output displays a similar table for the percentage of total variation in the X and Y variables explained as the number of components is increased.

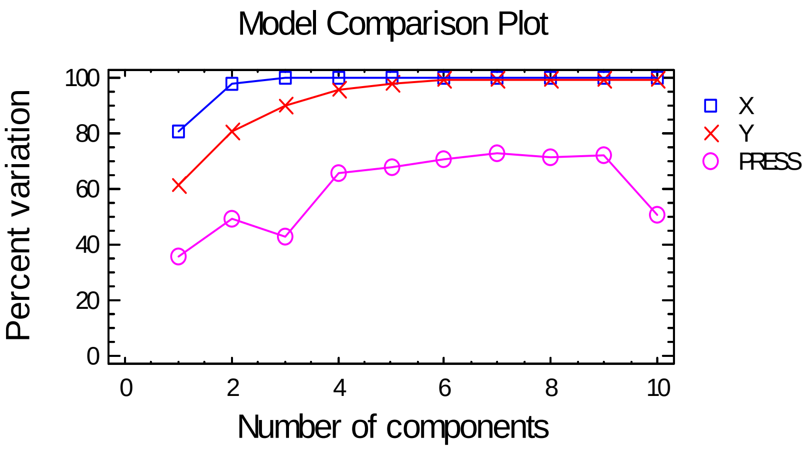
Independent and Dependent Variables

	% Variation	*Cumulative %*	*% Variation*	*Cumulative %*	*Average Prediction*
Component	*in X*	*of X*	*in Y*	*of Y*	*R-Squared*
1	81.0322	81.0322	61.1595	61.1595	35.4297
2	16.8495	97.8816	19.7254	80.8849	49.4234
3	1.85606	99.7377	9.13549	90.0204	42.6625
4	0.197979	99.9357	5.56999	95.5903	65.5749
5	0.0276934	99.9634	2.50521	98.0956	68.1911
6	0.0128011	99.9762	0.918146	99.0137	70.4208
7	0.00539246	99.9816	0.325377	99.3391	72.784
8	0.00581347	99.9874	0.0704414	99.4095	71.1374
9	0.00468166	99.9921	0.0970064	99.5065	72.1213
10	0.00405589	99.9961	0.0926184	99.5991	50.9856

The last column shows the average *Prediction R-Squared* across all dependent variables. The average peaks at 7 components, suggesting that a model with seven components would be a good choice.

Model Comparison Plot

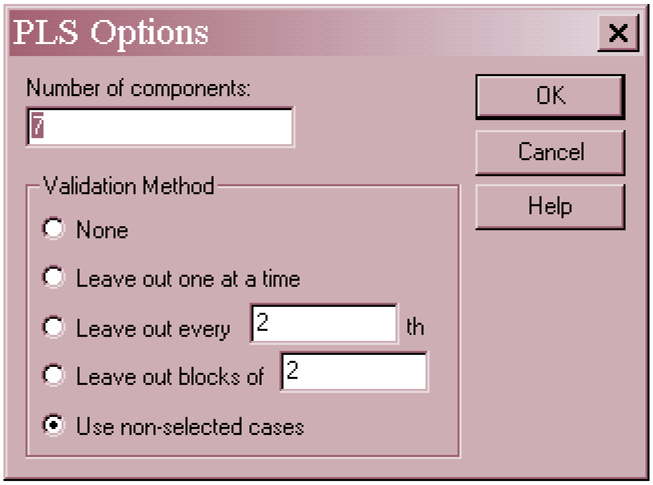
The cumulative percent variation in X and Y and the average prediction R-squared displayed in the table above are plotted by the *Model Comparison Plot*.



This plot is helpful in visualizing how many components need to be extracted. Note that the percent variation for PRESS increases through 7 components.

Note: In the rest of this document, results will be shown for a model with 7 components.

Analysis Options



PLS Options

Number of components: 7

Validation Method

○ None

○ Leave out one at a time

○ Leave out every 2 th

○ Leave out blocks of 2

● Use non-selected cases

- **Number of components**: the number of components to include in the fitted model. This number cannot exceed the smaller of the number of independent variables or *n* – 1.

- **Validation Method**: the method used to cross-validate the model. This consists of exercising the model to predict observations excluded from the model fit. The following methods may be used:

  1. *None* – no cross-validation is performed.

  2. *Leave out one at a time* – the modeling is refit n times, each time leaving out 1 of the observations and refitting the model using the other *n* – 1. The omitted observation is then predicted with the model from which it was excluded.

  3. *Leave out every k-th* - this is similar to method #2, except that only every *k-th* observation is omitted and then predicted. This shortens the process on large data sets.

  4. *Leave out blocks of k* – observations are removed in groups of *k*, the model refit, and the *k* observations predicted.

  5. *Use non-selected cases* – if an entry was made in the *Select* field on the data input dialog box, the cases excluded by that entry are used as test cases.

In the example, the *Select* field chose the first 18 rows to use as a training set for the model, with the remaining 15 rows making up a test set.

Regression Coefficients

The *Regression Coefficients* table shows the estimated coefficients of the fitted models. Both standardized and unstandardized coefficients are displayed. A small section of the output is shown below:

Regression Coefficients

Standardized Coefficients

	LOG10(Tryptophan)	*LOG10(Tyrosine)*
Constant	0.0	0.0
f1	-0.160437	1.27641
f2	0.1732	0.767133
f3	-0.170751	2.07999
f4	0.422583	-3.19308
…	…	…

Unstandardized Coefficients

	LOG10(Tryptophan)	*LOG10(Tyrosine)*
Constant	-4.85093	-0.374954
f1	-0.104881	0.962157
f2	0.113427	0.579294
f3	-0.126316	1.77426
f4	0.406053	-3.53788
…	…	…

The unstandardized model shows the fitted equation in the metric of the original measurements. For example, the model for the first dependent variable is

$$\log(\mathit{Tryptopan}) = -4.851 - 0.105f_1 + 0.113f_2 - 0.126f_3 + 0.406f_4 + \ldots \qquad (6)$$

The standardized model reexpresses each of the variables in a standardized form by subtracting its sample mean and dividing by its sample standard deviation. Expressing the new variables as Y, X_1, X_2, and so on, the standardized model for the sample data is

$$Y = -0.160X_1 + 0.173X_2 - 0.171X_3 + 0.423X_4 + \ldots \qquad (7)$$

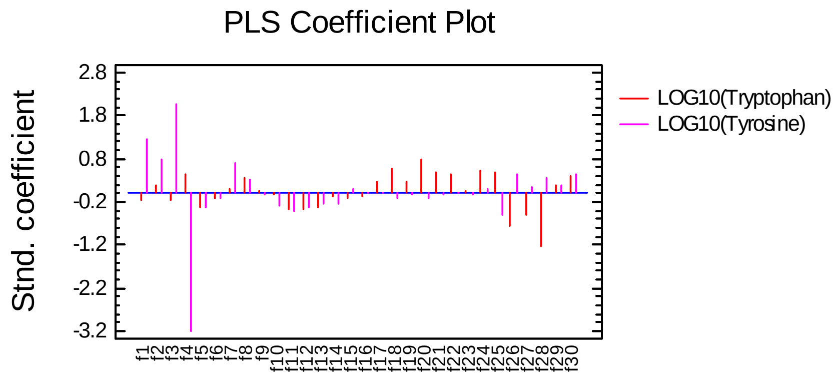
While the unstandardized model is useful for making predictions for new samples, the coefficients of the standardized model are more easily compared with each other when the predictor variables have different units.

Coefficient Plot

The *Coefficient Plot* displays either of two quantities:

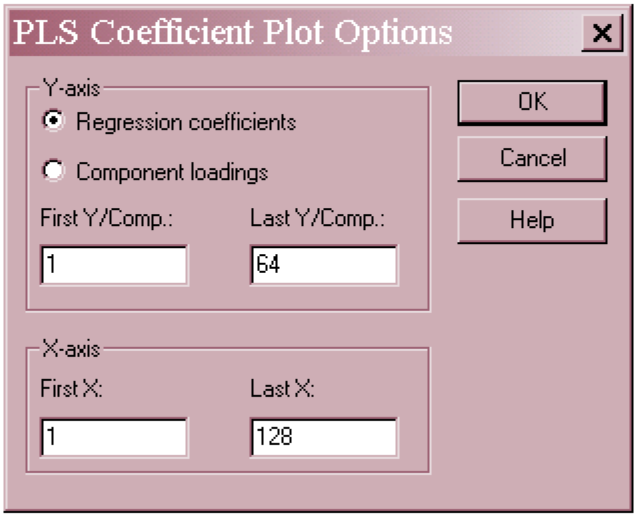
1. The standardized regression coefficients β for each dependent variable.
2. The component loadings *Q* for each dependent variable.

The example below plots the β's:



The coefficients provide a type of signature for each dependent variable. Note the large negative coefficients for *f4* when predicting *LOG10(Tyrosine).*

Pane Options



PLS Coefficient Plot Options

Y-axis

● Regression coefficients

○ Component loadings

First Y/Comp.: 1

Last Y/Comp.: 64

X-axis

First X: 1

Last X: 128

- **Y-axis**: the quantity and value to plot on the vertical axis.

- **First Y/Comp**: the index of the first variable or component to include in the plot.

- **Last Y/Comp**: the index of the last variable or component to include in the plot.

- **First X**: the index of the first independent variable to include in the plot.

- **Last X**: the index of the last independent variable to include in the plot.

Component Weights and Loadings

The *Component Weights and Loadings* table identifies each of the components that was extracted from the data. A portion of the table is shown below:

Component Weights and Loadings

Dependent Variables

	1	*2*	*3*	*4*	*5*	*6*	*7*
LOG10(Tryptophan)	0.192348	0.0570662	0.229545	-0.764634	1.69537	-1.39671	-0.294154
LOG10(Tyrosine)	-0.117072	0.281668	0.547727	1.16479	1.80889	1.85993	2.24394

Independent Variables

	1	*2*	*3*	*4*	*5*	*6*	*7*
f1	-0.172149	0.403733	0.391608	0.298026	0.232811	0.334922	-0.206403
f2	-0.168901	0.414018	0.399816	0.327923	0.181725	0.00542026	-0.137708
f3	-0.163081	0.403805	0.290047	0.156741	0.045198	-0.0121344	0.689201
f4	-0.151243	0.372398	0.0731797	-0.205695	-0.447595	-0.61829	-0.515405
…	…	…	…	…	…	…	…

Included in the table are:

1. Q, the c by m matrix of **loadings** (regression coefficients) relating the factor score matrix T to the dependent variable Y:

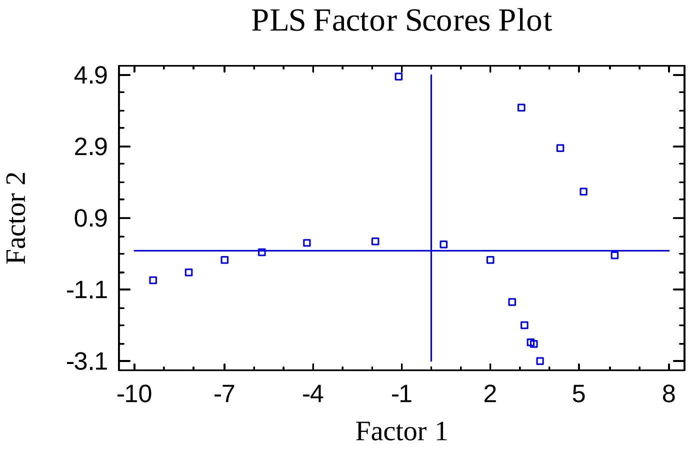
$$Y = TQ + E \tag{8}$$

2. W, the p by c matrix of factor weights, which create the factor scores from the standardized values of the independent variables according to

$$T = XW \tag{9}$$

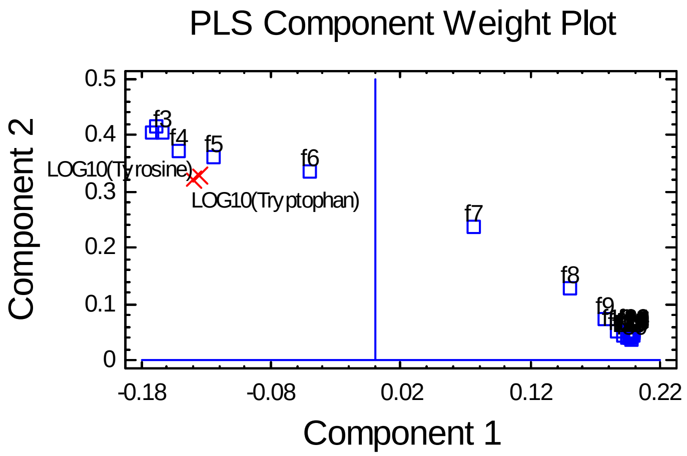
2D Component Plots

The *2D Component Plots* option will display either the factor score matrix T or the component weight matrices W and P. In the case of the factor score matrix, the plot takes the following form:



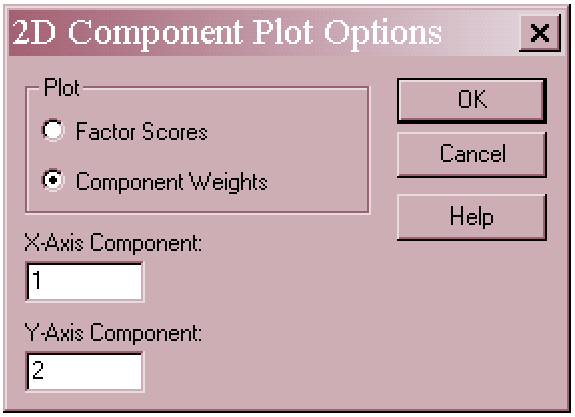
Two factors are selected, one for each axis, and n points are plotted representing the n rows in the corresponding columns of T. In situations where the factors are interpretable, this plot shows each sample's value for those factors.

If the component weights are selected, the plot has the following form:



Two components are selected, one for each axis, and *p* + *m* points are plotted representing the *p* independent variables and the *m* dependent variables. From this plot, it may be seen how each of the original variables affects the derived components.

Pane Options


2D Component Plot Options

Plot
○ Factor Scores
● Component Weights

X-Axis Component: 1

Y-Axis Component: 2

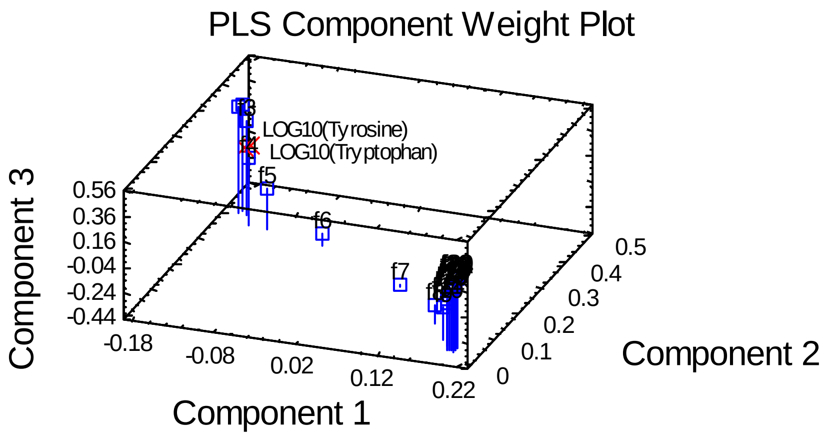
- **Plot** – Select columns of either the factor scores matrix *T* or the component weights matrix *W*.

- **X-Axis Component**: Select one of the *c* components to plot on the horizontal axis.

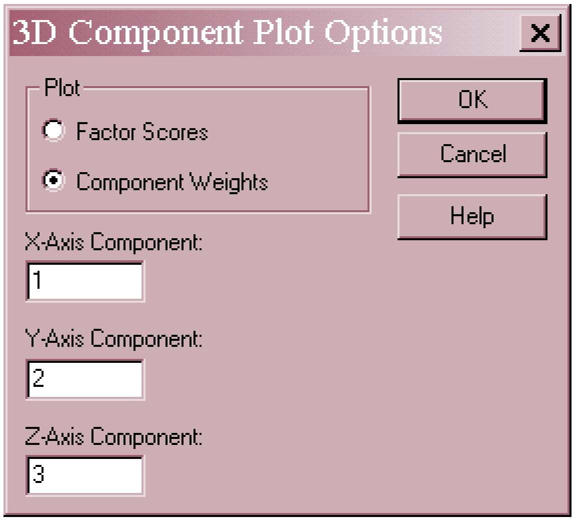
- **Y-Axis Component**: Select one of the *c* components to plot on the vertical axis.

3D Component Plots

The *3D Component Plots* option parallels the 2D plot except that three components are selected.



Pane Options



3D Component Plot Options

Plot

○ Factor Scores

● Component Weights

X-Axis Component: 1

Y-Axis Component: 2

Z-Axis Component: 3

- **Plot –** Select columns of either the factor scores matrix *T* or the component weights matrix *W*.

- **X-Axis Component**: Select one of the *c* components to plot on the horizontal axis.

- **Y-Axis Component**: Select one of the *c* components to plot on the axis extending back into the screen.

- **Z-Axis Component**: Select one of the *c* components to plot on the vertical axis.

Predictions and Residuals

The *Predictions and Residuals* pane will display information for observations in the training set, observations in the test set, and/or any new rows that have been added to the datasheet containing values for the independent variables but missing values for *Y*. The last option allows you to exercise the model and make predictions for observations not included in either the training or test set.

The table below shows part of the output for the example data:

Predictions and Residuals

				Standardized
Row	*LOG10(Tryptophan)*	*Predicted*	*Residual*	*Residual*
1	-4.52288	-4.49803	-0.024852	-0.768533
2	-4.52724	-4.5206	-0.0066395	-0.234679
3	-4.53387	-4.57756	0.04369	1.73365
4	-4.54516	-4.52187	-0.0232803	-0.622566
…	…	…	…	…

A separate table is included for each of the dependent variables. Included in the table are:

- *Row* – the row number in the datasheet.

- *Y* – the observed value of the dependent variable, if any.

- *Predicted* – the predicted value $\hat{Y}$ from the fitted model.

- *Residual* – the residual value for the i-th observation of the j-th dependent variable is calculated from

$$e_{ij} = Y_{ij} - \hat{Y}_{ij} \qquad (10)$$

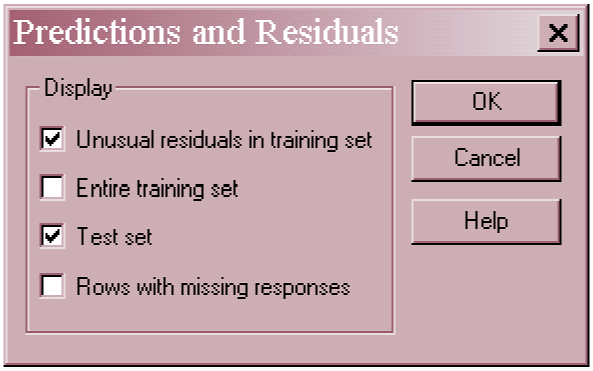
- *Standardized Residual* – for cases in the training set, an internally Studentized residual calculated by dividing each residual by an estimate of its standard error, given by

$$r_{ij} = \frac{e_{ij}}{\sqrt{MSE_j(1-h_i)}} \qquad (11)$$

  where h_i is the leverage of the i-th case.

© 2009 by StatPoint Technologies, Inc.

Pane Options



Predictions and Residuals

Display

☑ Unusual residuals in training set

□ Entire training set

☑ Test set

□ Rows with missing responses

The rows displayed may include:

1. *Unusual residuals in the training set*: any rows in the training set with standardized residuals exceeding 2 in absolute value.

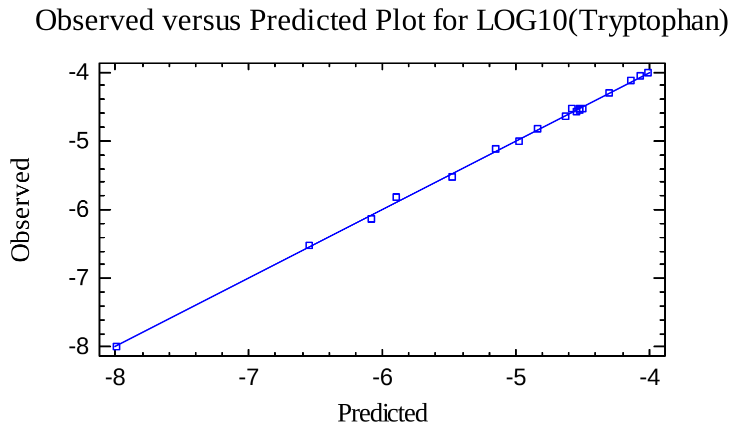
2. *Entire training set*: all rows in the training set.

3. *Test set*: all rows in the test set.

4. *Rows with missing responses*: rows with missing values for one or more of the dependent variables.

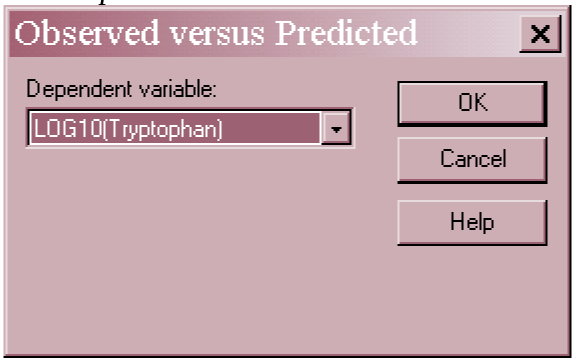
Observed versus Predicted

This plot shows the observed values of a selected dependent variable versus the values predicted by the fitted model:



If the model fits well, the points should line along the diagonal line.

Pane Options



Dependent variable: LOG10(Tryptophan)

Select the desired dependent variable to plot.

Leverages

In fitting a PLS model, all observations do not have an equal influence on the coefficient estimates in the fitted model. Those with unusual values of the independent variables tend to have more influence than the others. The *Leverages* pane displays any observations that have unusually high influence on the fitted model:

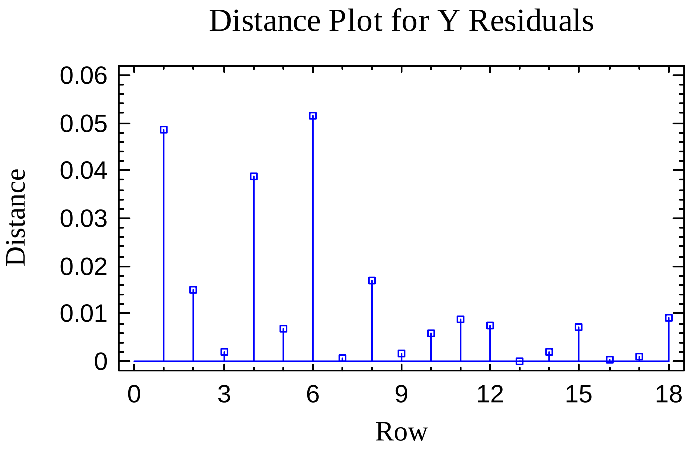
```
Leverages
Row   Leverage
Average leverage of single data point = 0.388889
```

Leverage is a statistic that measures the influence of each observation on the final model. Observations are placed on the list if they have more than 3 times the leverage of an average point. Observations with high leverage should be examined closely to be sure that they are valid, since a high leverage point that is also an outlier can badly distort the estimated model.

In the sample data, there are no high leverage points.

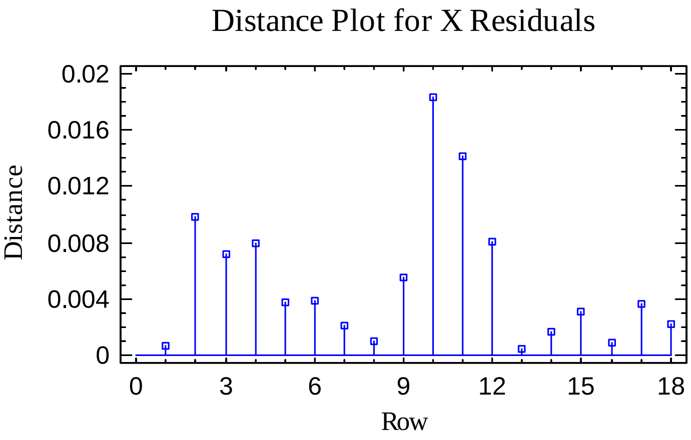
Residual Distance Graphs

The *Residual Distance Graphs* plot the distance from the origin to the X or Y residuals corresponding to each case in the training set. The plots may be used to determine which cases deviate most from the predicted values.



Distances are expressed as the sum of squares of the difference between the observed and predicted values of the standardized variables. For the *Y* variables, the residuals are elements of the *n* by *m* matrix *E* in the equation

$$Y = X\beta + E \tag{12}$$



For the *X* variables, the residuals are elements of the *n* by *p* matrix *F* in the equation

$$X = TP + F \tag{13}$$

Save Results

The following results may be saved to the data sheet:

1. *Predicted values* – the predicted values of the dependent variable(s).
2. *Y Residuals* – the residuals for each dependent variable.
3. *Standardized Y Residuals* – the standardized residuals for each dependent variable.
4. *PRESS residuals* – the PRESS residuals for each dependent variable.
5. *X Residuals* – the residuals for each independent variable.
6. *Leverages* – the leverages for each of the *n* cases.
7. *Y Distance* – the residual *Y* distance for each of the *n* cases.
8. *X Distance* – the residual *X* distance for each of the *n* cases.
9. *Component Weights* – the weight matrix *W*.
10. *Y Factor Loadings* – the factor loading matrix *Q*.
11. *X Factor Loadings* – the factor loading matrix *P*.
12. *Scores* – the score matrix *T*.

Calculations

The program uses the NIPALS (Nonlinear Iterative Partial Least Squares) algorithm to extract the components, after first transforming each variable so that it has a mean equal to 0 and a standard deviation equal to 1.