SAS/STAT® 9.2 User's Guide
The PLS Procedure
(Book Excerpt)

Chapter 66

The PLS Procedure

Contents

Overview: PLS Procedure

The PLS procedure fits models by using any one of a number of linear predictive methods, including *partial least squares* (PLS). Ordinary least squares regression, as implemented in SAS/STAT procedures such as PROC GLM and PROC REG, has the single goal of minimizing sample response prediction error, seeking linear functions of the predictors that explain as much variation in each

response as possible. The techniques implemented in the PLS procedure have the additional goal of accounting for variation in the predictors, under the assumption that directions in the predictor space that are well sampled should provide better prediction for *new* observations when the predictors are highly correlated. All of the techniques implemented in the PLS procedure work by extracting successive linear combinations of the predictors, called *factors* (also called *components*, *latent vectors*, or *latent variables*), which optimally address one or both of these two goals—explaining response variation and explaining predictor variation. In particular, the method of partial least squares balances the two objectives, seeking factors that explain both response and predictor variation.

Note that the name "partial least squares" also applies to a more general statistical method that is *not* implemented in this procedure. The partial least squares method was originally developed in the 1960s by the econometrician Herman Wold (1966) for modeling "paths" of causal relation between any number of "blocks" of variables. However, the PLS procedure fits only *predictive* partial least squares models, with one "block" of predictors and one "block" of responses. If you are interested in fitting more general path models, you should consider using the CALIS procedure.

Basic Features

The techniques implemented by the PLS procedure are as follows:

- principal components regression, which extracts factors to explain as much predictor sample variation as possible

- reduced rank regression, which extracts factors to explain as much response variation as possible. This technique, also known as (maximum) redundancy analysis, differs from multivariate linear regression only when there are multiple responses.

- partial least squares regression, which balances the two objectives of explaining response variation and explaining predictor variation. Two different formulations for partial least squares are available: the original predictive method of Wold (1966) and the SIMPLS method of de Jong (1993).

The number of factors to extract depends on the data. Basing the model on more extracted factors improves the model fit to the observed data, but extracting too many factors can cause *overfitting*— that is, tailoring the model too much to the current data, to the detriment of future predictions. The PLS procedure enables you to choose the number of extracted factors by *cross validation*—that is, fitting the model to part of the data, minimizing the prediction error for the unfitted part, and iterating with different portions of the data in the roles of fitted and unfitted. Various methods of cross validation are available, including one-at-a-time validation and splitting the data into blocks. The PLS procedure also offers test set validation, where the model is fit to the entire primary input data set and the fit is evaluated over a distinct test data set.

You can use the general linear modeling approach of the GLM procedure to specify a model for your design, allowing for general polynomial effects as well as classification or ANOVA effects. You can save the model fit by the PLS procedure in a data set and apply it to new data by using the SCORE procedure.

The PLS procedure now uses ODS Graphics to create graphs as part of its output. For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." For specific information about the statistical graphics available with the PLS procedure, see the PLOTS options in the PROC PLS statements and the section "ODS Graphics" on page 4787.

Getting Started: PLS Procedure

Spectrometric Calibration

The example in this section illustrates basic features of the PLS procedure. The data are reported in Umetrics (1995); the original source is Lindberg, Persson, and Wold (1983). Suppose that you are researching pollution in the Baltic Sea, and you would like to use the spectra of samples of seawater to determine the amounts of three compounds present in samples from the Baltic Sea: lignin sulfonate (ls: pulp industry pollution), humic acids (ha: natural forest products), and optical whitener from detergent (dt). Spectrometric calibration is a type of problem in which partial least squares can be very effective. The predictors are the spectra emission intensities at different frequencies in sample spectrum, and the responses are the amounts of various chemicals in the sample.

For the purposes of calibrating the model, samples with known compositions are used. The calibration data consist of 16 samples of known concentrations of ls, ha, and dt, with spectra based on 27 frequencies (or, equivalently, wavelengths). The following statements create a SAS data set named Sample for these data.

```
data Sample;
   input obsnam $ v1-v27 ls ha dt @@;
   datalines;
EM1   2766 2610 3306 3630 3600 3438 3213 3051 2907 2844 2796
      2787 2760 2754 2670 2520 2310 2100 1917 1755 1602 1467
      1353 1260 1167 1101 1017         3.0110   0.0000    0.00
EM2   1492 1419 1369 1158  958  887  905  929  920  887  800
       710  617  535  451  368  296  241  190  157  128  106
        89   70   65   56   50         0.0000   0.4005    0.00
EM3   2450 2379 2400 2055 1689 1355 1109  908  750  673  644
       640  630  618  571  512  440  368  305  247  196  156
       120   98   80   61   50         0.0000   0.0000   90.63
EM4   2751 2883 3492 3570 3282 2937 2634 2370 2187 2070 2007
      1974 1950 1890 1824 1680 1527 1350 1206 1080  984  888
       810  732  669  630  582         1.4820   0.1580   40.00
EM5   2652 2691 3225 3285 3033 2784 2520 2340 2235 2148 2094
      2049 2007 1917 1800 1650 1464 1299 1140 1020  909  810
       726  657  594  549  507         1.1160   0.4104   30.45
EM6   3993 4722 6147 6720 6531 5970 5382 4842 4470 4200 4077
      4008 3948 3864 3663 3390 3090 2787 2481 2241 2028 1830
      1680 1533 1440 1314 1227         3.3970   0.3032   50.82
EM7   4032 4350 5430 5763 5490 4974 4452 3990 3690 3474 3357
```

```
        3300 3213 3147 3000 2772 2490 2220 1980 1779 1599 1440
        1320 1200 1119 1032  957          2.4280  0.2981  70.59
EM8     4530 5190 6910 7580 7510 6930 6150 5490 4990 4670 4490
        4370 4300 4210 4000 3770 3420 3060 2760 2490 2230 2060
        1860 1700 1590 1490 1380          4.0240  0.1153  89.39
EM9     4077 4410 5460 5857 5607 5097 4605 4170 3864 3708 3588
        3537 3480 3330 3192 2910 2610 2325 2064 1830 1638 1476
        1350 1236 1122 1044  963          2.2750  0.5040  81.75
EM10    3450 3432 3969 4020 3678 3237 2814 2487 2205 2061 2001
        1965 1947 1890 1776 1635 1452 1278 1128  981  867  753
         663  600  552  507  468          0.9588  0.1450 101.10
EM11    4989 5301 6807 7425 7155 6525 5784 5166 4695 4380 4197
        4131 4077 3972 3777 3531 3168 2835 2517 2244 2004 1809
        1620 1470 1359 1266 1167          3.1900  0.2530 120.00
EM12    5340 5790 7590 8390 8310 7670 6890 6190 5700 5380 5200
        5110 5040 4900 4700 4390 3970 3540 3170 2810 2490 2240
        2060 1870 1700 1590 1470          4.1320  0.5691 117.70
EM13    3162 3477 4365 4650 4470 4107 3717 3432 3228 3093 3009
        2964 2916 2838 2694 2490 2253 2013 1788 1599 1431 1305
        1194 1077  990  927  855          2.1600  0.4360  27.59
EM14    4380 4695 6018 6510 6342 5760 5151 4596 4200 3948 3807
        3720 3672 3567 3438 3171 2880 2571 2280 2046 1857 1680
        1548 1413 1314 1200 1119          3.0940  0.2471  61.71
EM15    4587 4200 5040 5289 4965 4449 3939 3507 3174 2970 2850
        2814 2748 2670 2529 2328 2088 1851 1641 1431 1284 1134
        1020  918  840  756  714          1.6040  0.2856 108.80
EM16    4017 4725 6090 6570 6354 5895 5346 4911 4611 4422 4314
        4287 4224 4110 3915 3600 3240 2913 2598 2325 2088 1917
        1734 1587 1452 1356 1257          3.1620  0.7012  60.00
;
```

Fitting a PLS Model

To isolate a few underlying spectral factors that provide a good predictive model, you can fit a PLS model to the 16 samples by using the following SAS statements:

```
proc pls data=sample;
   model ls ha dt = v1-v27;
run;
```

By default, the PLS procedure extracts at most 15 factors. The procedure lists the amount of variation accounted for by each of these factors, both individual and cumulative; this listing is shown in

Figure 66.1 PLS Variation Summary

```
                        The PLS Procedure

                   Percent Variation Accounted for
                   by Partial Least Squares Factors

        Number of
        Extracted          Model Effects        Dependent Variables
         Factors        Current      Total      Current       Total

            1           97.4607     97.4607     41.9155      41.9155
            2            2.1830     99.6436     24.2435      66.1590
            3            0.1781     99.8217     24.5339      90.6929
            4            0.1197     99.9414      3.7898      94.4827
            5            0.0415     99.9829      1.0045      95.4873
            6            0.0106     99.9935      2.2808      97.7681
            7            0.0017     99.9952      1.1693      98.9374
            8            0.0010     99.9961      0.5041      99.4415
            9            0.0014     99.9975      0.1229      99.5645
           10            0.0010     99.9985      0.1103      99.6747
           11            0.0003     99.9988      0.1523      99.8270
           12            0.0003     99.9991      0.1291      99.9561
           13            0.0002     99.9994      0.0312      99.9873
           14            0.0004     99.9998      0.0065      99.9938
           15            0.0002    100.0000      0.0062     100.0000
```

Note that all of the variation in both the predictors and the responses is accounted for by only 15 factors; this is because there are only 16 sample observations. More important, almost all of the variation is accounted for with even fewer factors—one or two for the predictors and three to eight for the responses.

Selecting the Number of Factors by Cross Validation

A PLS model is not complete until you choose the number of factors. You can choose the number of factors by using cross validation, in which the data set is divided into two or more groups. You fit the model to all groups except one, and then you check the capability of the model to predict responses for the group omitted. Repeating this for each group, you then can measure the overall capability of a given form of the model. The predicted residual sum of squares (PRESS) statistic is based on the residuals generated by this process.

To select the number of extracted factors by cross validation, you specify the CV= option with an argument that says which cross validation method to use. For example, a common method is split-sample validation, in which the different groups are composed of every nth observation beginning with the first, every nth observation beginning with the second, and so on. You can use the CV=SPLIT option to specify split-sample validation with $n = 7$ by default, as in the following SAS statements:

```
proc pls data=sample cv=split;
   model ls ha dt = v1-v27;
run;
```

The resulting output is shown in Figure 66.2 and Figure 66.3.

Figure 66.2 Split-Sample Validated PRESS Statistics for Number of Factors

```
                          The PLS Procedure

         Split-sample Validation for the Number of Extracted Factors

                         Number of          Root
                         Extracted          Mean
                          Factors           PRESS

                               0          1.107747
                               1          0.957983
                               2          0.931314
                               3          0.520222
                               4          0.530501
                               5          0.586786
                               6          0.475047
                               7          0.477595
                               8          0.483138
                               9          0.485739
                              10           0.48946
                              11          0.521445
                              12          0.525653
                              13          0.531049
                              14          0.531049
                              15          0.531049


                Minimum root mean PRESS              0.4750
                Minimizing number of factors              6
```

Figure 66.3 PLS Variation Summary for Split-Sample Validated Model

```
                          Percent Variation Accounted for
                          by Partial Least Squares Factors

         Number of
         Extracted          Model Effects          Dependent Variables
          Factors       Current       Total        Current       Total

               1         97.4607      97.4607       41.9155      41.9155
               2          2.1830      99.6436       24.2435      66.1590
               3          0.1781      99.8217       24.5339      90.6929
               4          0.1197      99.9414        3.7898      94.4827
               5          0.0415      99.9829        1.0045      95.4873
               6          0.0106      99.9935        2.2808      97.7681
```

The absolute minimum PRESS is achieved with six extracted factors. Notice, however, that this is not much smaller than the PRESS for three factors. By using the CVTEST option, you can perform a statistical model comparison suggested by van der Voet (1994) to test whether this difference is significant, as shown in the following SAS statements:

```
proc pls data=sample cv=split cvtest(seed=12345);
   model ls ha dt = v1-v27;
run;
```

The model comparison test is based on a rerandomization of the data. By default, the seed for this randomization is based on the system clock, but it is specified here. The resulting output is shown in Figure 66.4 and Figure 66.5.

Figure 66.4 Testing Split-Sample Validation for Number of Factors

```
                        The PLS Procedure

        Split-sample Validation for the Number of Extracted Factors

              Number of          Root
              Extracted          Mean                     Prob >
                Factors          PRESS          T**2        T**2

                     0        1.107747       9.272858     0.0010
                     1        0.957983      10.62305      <.0001
                     2        0.931314       8.950878     0.0010
                     3        0.520222       5.133259     0.1440
                     4        0.530501       5.168427     0.1340
                     5        0.586786       6.437266     0.0150
                     6        0.475047              0     1.0000
                     7        0.477595       2.809763     0.4750
                     8        0.483138       7.189526     0.0110
                     9        0.485739       7.931726     0.0070
                    10         0.48946       6.612597     0.0150
                    11        0.521445       6.666235     0.0130
                    12        0.525653       7.092861     0.0080
                    13        0.531049       7.538298     0.0030
                    14        0.531049       7.538298     0.0030
                    15        0.531049       7.538298     0.0030


        Minimum root mean PRESS                              0.4750
        Minimizing number of factors                              6
        Smallest number of factors with p > 0.1                   3
```

Figure 66.5 PLS Variation Summary for Tested Split-Sample Validated Model

```
                      Percent Variation Accounted for
                      by Partial Least Squares Factors

           Number of
           Extracted          Model Effects          Dependent Variables
             Factors       Current      Total        Current      Total

                   1        97.4607     97.4607       41.9155     41.9155
                   2         2.1830     99.6436       24.2435     66.1590
                   3         0.1781     99.8217       24.5339     90.6929
```

The *p*-value of 0.1430 in comparing the cross validated residuals from models with 6 and 3 factors indicates that the difference between the two models is insignificant; therefore, the model with fewer factors is preferred. The variation summary shows that over 99% of the predictor variation and over 90% of the response variation are accounted for by the three factors.

Predicting New Observations

Now that you have chosen a three-factor PLS model for predicting pollutant concentrations based on sample spectra, suppose that you have two new samples. The following SAS statements create a data set containing the spectra for the new samples:

```
data newobs;
   input obsnam $ v1-v27 @@;
   datalines;
EM17   3933 4518 5637 6006 5721 5187 4641 4149 3789
       3579 3447 3381 3327 3234 3078 2832 2571 2274
       2040 1818 1629 1470 1350 1245 1134 1050  987
EM25   2904 2997 3255 3150 2922 2778 2700 2646 2571
       2487 2370 2250 2127 2052 1713 1419 1200  984
        795  648  525  426  351  291  240  204  162
;
```

You can apply the PLS model to these samples to estimate pollutant concentration. To do so, append the new samples to the original 16, and specify that the predicted values for all 18 be output to a data set, as shown in the following statements:

```
data all; set sample newobs;
proc pls data=all nfac=3;
   model ls ha dt = v1-v27;
   output out=pred p=p_ls p_ha p_dt;
run;

proc print data=pred;
   where (obsnam in ('EM17','EM25'));
   var obsnam p_ls p_ha p_dt;
run;
```

The new observations are not used in calculating the PLS model, since they have no response values. Their predicted concentrations are shown in Figure 66.6.

Figure 66.6 Predicted Concentrations for New Observations

Obs	obsnam	p_ls	p_ha	p_dt
17	EM17	2.54261	0.31877	81.4174
18	EM25	-0.24716	1.37892	46.3212

Finally, if you enable ODS graphics, PLS also displays by default a plot of the amount of variation accounted for by each factor, as well as a correlations loading plot that summarizes the first two
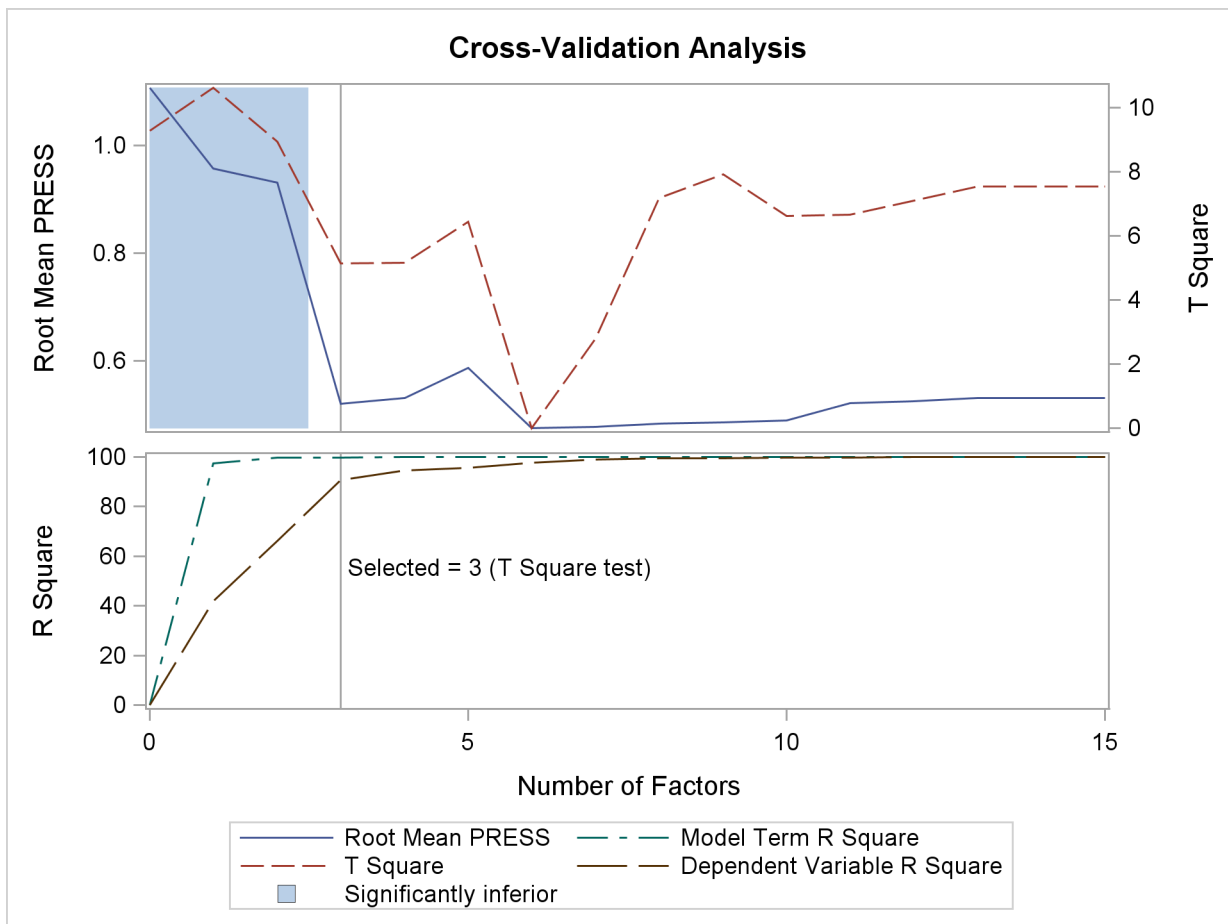
dimensions of the PLS model. The following statements, which are the same as the previous split-sample validation analysis but with ODS graphics enabled, additionally produce Figure 66.7 and Figure 66.8:
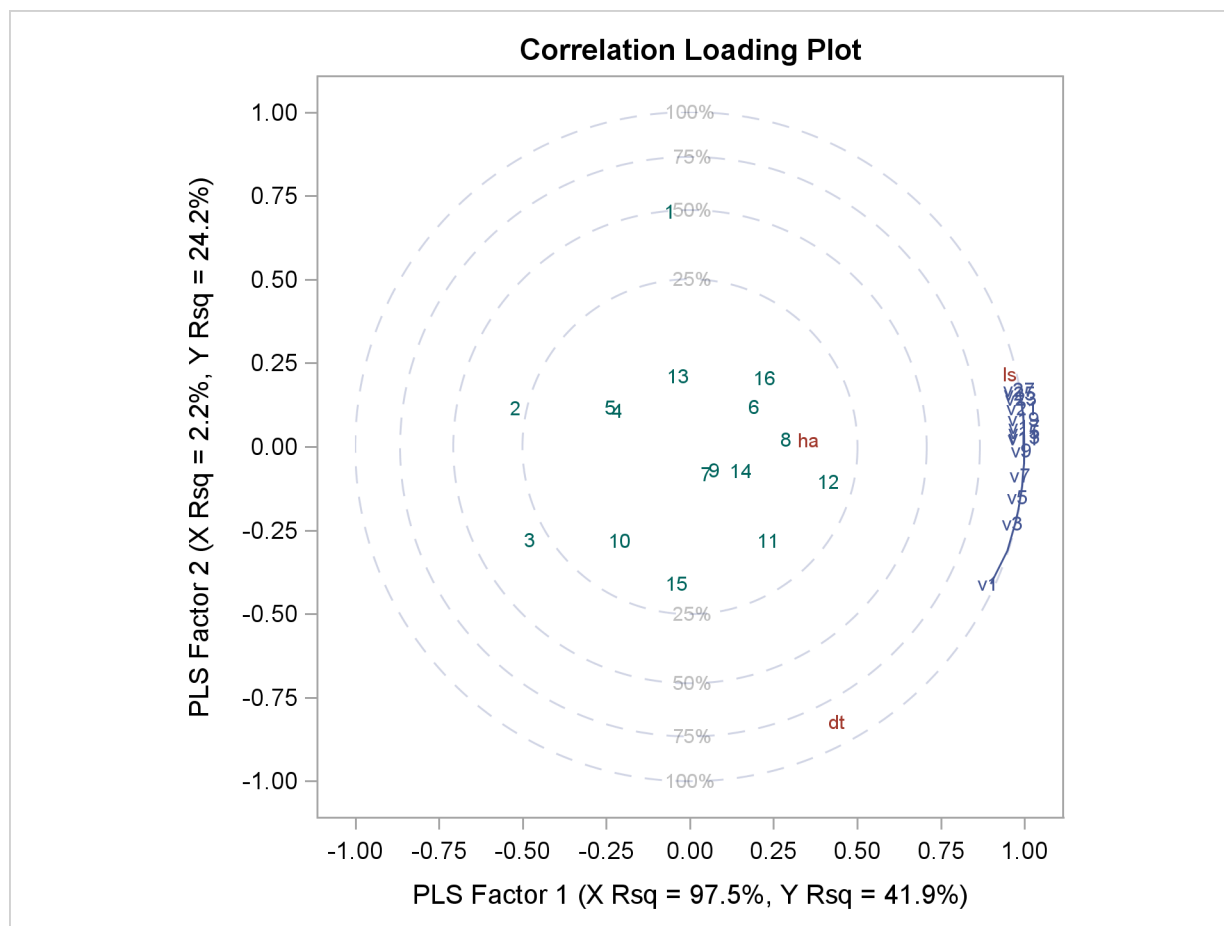
```
ods graphics on;

proc pls data=sample cv=split cvtest(seed=12345);
   model ls ha dt = v1-v27;
run;

ods graphics off;
```

Figure 66.7 Split-Sample Cross Validation Plot

Figure 66.8 Correlation Loadings Plot

The cross validation plot in Figure 66.7 gives a visual representation of the selection of the optimum number of factors discussed previously. The correlation loadings plot is a compact summary of many features of the PLS model. For example, it shows that the first factor is highly positively correlated with all spectral values, indicating that it is approximately an average of them all; the second factor is positively correlated with the lowest frequencies and negatively correlated with the highest, indicating that it is approximately a contrast between the two ends of the spectrum. The observations, represented by their number in the data set on this plot, are generally spaced well apart, indicating that the data give good information about these first two factors. For more details on the interpretation of the correlation loadings plot, see the section "ODS Graphics" on page 4787 and Example 66.1.

Syntax: PLS Procedure

The following statements are available in PROC PLS. Items within the angle brackets are optional.

> **PROC PLS** *< options >* **;**
>> **BY** *variables* **;**
>> **CLASS** *variables < / option >* **;**
>> **ID** *variables* **;**
>> **MODEL** *dependent-variables = effects < / options >* **;**
>> **OUTPUT OUT=***SAS-data-set < options >* **;**

To analyze a data set, you must use the PROC PLS and MODEL statements. You can use the other statements as needed.

PROC PLS Statement

> **PROC PLS** *< options >* **;**

You use the PROC PLS statement to invoke the PLS procedure and, optionally, to indicate the analysis data and method. The following options are available.

CENSCALE
> lists the centering and scaling information for each response and predictor.

CV=ONE
CV=SPLIT *< (n) >*
CV=BLOCK *< (n) >*
CV=RANDOM *< (cv-random-opts) >*
CV=TESTSET(*SAS-data-set***)**
> specifies the cross validation method to be used. By default, no cross validation is performed. The method CV=ONE requests one-at-a-time cross validation, CV=SPLIT requests that every nth observation be excluded, CV=BLOCK requests that n blocks of consecutive observations be excluded, CV=RANDOM requests that observations be excluded at random, and CV=TESTSET(*SAS-data-set*) specifies a test set of observations to be used for validation (formally, this is called "test set validation" rather than "cross validation"). You can, optionally, specify n for CV=SPLIT and CV=BLOCK; the default is $n = 7$. You can also specify the following optional *cv-random-options* in parentheses after the CV=RANDOM option:

> | | |
> |---|---|
> | **NITER=***n* | specifies the number of random subsets to exclude. The default value is 10. |
> | **NTEST=***n* | specifies the number of observations in each random subset chosen for exclusion. The default value is one-tenth of the total number of observations. |
> | **SEED=***n* | specifies an integer used to start the pseudo-random number generator for selecting the random test set. If you do not specify a seed, or specify |

a value less than or equal to zero, the seed is by default generated from reading the time of day from the computer's clock.

CVTEST < **(***cvtest-options***)** >

specifies that van der Voet's (1994) randomization-based model comparison test be performed to test models with different numbers of extracted factors against the model that minimizes the predicted residual sum of squares; see the section "Cross Validation" on page 4782 for more information. You can also specify the following *cv-test-options* in parentheses after the CVTEST option:

PVAL=*n*   specifies the cutoff probability for declaring an insignificant difference. The default value is 0.10.

STAT=*test-statistic*   specifies the test statistic for the model comparison. You can specify either T2, for Hotelling's T^2 statistic, or PRESS, for the predicted residual sum of squares. The default value is T2.

NSAMP=*n*   specifies the number of randomizations to perform. The default value is 1000.

SEED=*n*   specifies the seed value for randomization generation (the clock time is used by default).

DATA=*SAS-data-set*

names the SAS data set to be used by PROC PLS. The default is the most recently created data set.

DETAILS

lists the details of the fitted model for each successive factor. The details listed are different for different extraction methods; see the section "Displayed Output" on page 4785 for more information.

METHOD=PLS < **(** *PLS-options* **)** >
METHOD=SIMPLS
METHOD=PCR
METHOD=RRR

specifies the general factor extraction method to be used. The value PLS requests partial least squares, SIMPLS requests the SIMPLS method of de Jong (1993), PCR requests principal components regression, and RRR requests reduced rank regression. The default is METHOD=PLS. You can also specify the following optional *PLS-options* in parentheses after METHOD=PLS:

ALGORITHM=NIPALS | SVD | EIG | RLGW

names the specific algorithm used to compute extracted PLS factors. NIPALS requests the usual iterative NIPALS algorithm, SVD bases the extraction on the singular value decomposition of $X'Y$, EIG bases the extraction on the eigenvalue decomposition of $Y'XX'Y$, and RLGW is an iterative approach that is efficient when there are many predictors. ALGORITHM=SVD is the most accurate but least efficient approach; the default is ALGORITHM=NIPALS.

MAXITER=*n* specifies the maximum number of iterations for the NIPALS and RLGW algorithms. The default value is 200.

EPSILON=*n* specifies the convergence criterion for the NIPALS and RLGW algorithms. The default value is 10^{-12}.

MISSING=NONE

MISSING=AVG

MISSING=EM <(*EM-options*)>

 specifies how observations with missing values are to be handled in computing the fit. The default is MISSING=NONE, for which observations with any missing variables (dependent or independent) are excluded from the analysis. MISSING=AVG specifies that the fit be computed by filling in missing values with the average of the nonmissing values for the corresponding variable. If you specify MISSING=EM, then the procedure first computes the model with MISSING=AVG and then fills in missing values by their predicted values based on that model and computes the model again. For both methods of imputation, the imputed values contribute to the centering and scaling values, and the difference between the imputed values and their final predictions contributes to the percentage of variation explained. You can also specify the following optional *EM-options* in parentheses after MISSING=EM:

MAXITER=*n* specifies the maximum number of iterations for the imputation/fit loop. The default value is 1. If you specify a large value of MAXITER=, then the loop will iterate until it converges (as controlled by the EPSILON= option).

EPSILON=*n* specifies the convergence criterion for the imputation/fit loop. The default value for is 10^{-8}. This option is effective only if you specify a large value for the MAXITER= option.

NFAC=*n*

 specifies the number of factors to extract. The default is min$\{15, p, N\}$, where p is the number of predictors (the number of dependent variables for METHOD=RRR) and N is the number of runs (observations). This is probably more than you need for most applications. Extracting too many factors can lead to an overfit model, one that matches the training data too well, sacrificing predictive ability. Thus, if you use the default NFAC= specification, you should also either use the CV= option to select the appropriate number of factors for the final model or consider the analysis to be preliminary and examine the results to determine the appropriate number of factors for a subsequent analysis.

NOCENTER

 suppresses centering of the responses and predictors before fitting. This is useful if the analysis variables are already centered and scaled. See the section "Centering and Scaling" on page 4784 for more information.

NOCVSTDIZE

 suppresses re-centering and rescaling of the responses and predictors before each model is fit in the cross validation. See the section "Centering and Scaling" on page 4784 for more information.

NOPRINT

suppresses the normal display of results. This is useful when you want only the output statistics saved in a data set. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 20, "Using the Output Delivery System" for more information.

NOSCALE

suppresses scaling of the responses and predictors before fitting. This is useful if the analysis variables are already centered and scaled. See the section "Centering and Scaling" on page 4784 for more information.

PLOTS < *(global-plot-options)* > < = *plot-request* < *(options)* > >

PLOTS < *(global-plot-options)* > < = *(plot-request* < *(options)* > < ... *plot-request* < *(options)* > >*)* >

controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses from around the plot request. For example:

```
plots=none
plots=cvplot
plots=(diagnostics cvplot)
plots(unpack)=diagnostics
plots(unpack)=(diagnostics corrload(trace=off))
```

You must enable ODS Graphics before requesting plots—for example, like this:

```
ods graphics on;
proc pls data=pentaTrain;
   model log_RAI = S1-S5 L1-L5 P1-P5;
run;
ods graphics off;
```

For general information about ODS Graphics, see Chapter 21, "Statistical Graphics Using ODS." If you have enabled ODS Graphics but do not specify the PLOTS= option, then PROC PLS produces by default a plot of the R-square analysis and a correlation loading plot summarizing the first two factors. The global plot options include the following:

FLIP

interchanges the X-axis and Y-axis dimensions for the score, weight, and loading plots.

ONLY

suppresses the default plots. Only plots specifically requested are displayed.

UNPACKPANEL

UNPACK

suppresses paneling. By default, multiple plots can appear in some output panels. Specify UNPACKPANEL to get each plot in a separate panel. You can specify PLOTS(UNPACKPANEL) to unpack only the default plots. You can also specify UNPACKPANEL as a suboption for certain specific plots, as discussed in the following.

The plot requests include the following:

ALL

produces all appropriate plots. You can specify other options with ALL—for example, to request all plots and unpack only the residuals, specify PLOTS=(ALL RESIDUALS(UNPACK)).

CORRLOAD < (TRACE = ON | OFF) >

produces a correlation loading plot (default). The TRACE= option controls how points corresponding to the X-loadings in the correlation loadings plot are depicted. By default, these points are depicted by the name of the corresponding model effect if there are 20 or fewer of them; otherwise, they are depicted by a connected "trace" through the points. You can use this option to change this behavior.

CVPLOT

produces a cross validation and R-square analysis. This plot requires the CV= option to be specified, and is displayed by default in this case.

DIAGNOSTICS < (UNPACK) >

produces a summary panel of the fit for each dependent variable. The summary by default consists of a panel for each dependent variable, with plots depicting the distribution of residuals and predicted values. You can use the UNPACK suboption to specify that the subplots be produced separately.

DMOD

produces the DMODX, DMODY, and DMODXY plots.

DMODX

produces a plot of the distance of each observation to the X model.

DMODXY

produces plots of the distance of each observation to the X and Y models.

DMODY

produces a plot of the distance of each observation to the Y model.

FIT

produces both the fit diagnostics and the ParmProfiles plot.

NONE

suppresses the display of graphics.

PARMPROFILES

produces profiles of the regression coefficients.

SCORES < (UNPACK | FLIP) >

produces the XScores, YScores, XYScores, and DModXY plots. You can use the UNPACK suboption to specify that the subplots for scores be produced separately, and the FLIP option to interchange their default X-axis and Y-axis dimensions.

RESIDUALS < (UNPACK) >

plots the residuals for each dependent variable against each independent variable. Residual plots are by default composed of multiple plots combined into a single panel. You can use the UNPACK suboption to specify that the subplots be produced separately.

VIP

produces profiles of variable importance factors.

WEIGHTS < (UNPACK | FLIP) >

produces all X and Y loading and weight plots, as well as the VIP plot. You can use the UNPACK suboption to specify that the subplots for weights and loadings be produced separately, and the FLIP option to interchange their default X-axis and Y-axis dimensions.

XLOADINGPLOT < (UNPACK | FLIP) >

produces a scatter plot matrix of X-loadings against each other. Loading scatter plot matrices are by default composed of multiple plots combined into a single panel. You can use the UNPACK suboption to specify that the subplots be produced separately, and the FLIP option to interchange the default X-axis and Y-axis dimensions.

XLOADINGPROFILES

produces profiles of the X-loadings.

XSCORES < (UNPACK | FLIP) >

produces a scatter plot matrix of X-scores against each other. Score scatter plot matrices are by default composed of multiple plots combined into a single panel. You can use the UNPACK suboption to specify that the subplots be produced separately, and the FLIP option to interchange the default X-axis and Y-axis dimensions.

XWEIGHTPLOT < (UNPACK | FLIP) >

produces a scatter plot matrix of X-weights against each other. Weight scatter plot matrices are by default composed of multiple plots combined into a single panel. You can use the UNPACK suboption to specify that the subplots be produced separately, and the FLIP option to interchange the default X-axis and Y-axis dimensions.

XWEIGHTPROFILES

produces profiles of the X-weights.

XYSCORES < (UNPACK) >

produces a scatter plot matrix of X-scores against Y-scores. Score scatter plot matrices are by default composed of multiple plots combined into a single panel. You can use the UNPACK suboption to specify that the subplots be produced separately.

YSCORES < (UNPACK | FLIP) >

produces a scatter plot matrix of Y-scores against each other. Score scatter plot matrices are by default composed of multiple plots combined into a single panel. You can use the UNPACK suboption to specify that the subplots be produced separately, and the FLIP option to interchange the default X-axis and Y-axis dimensions.

YWEIGHTPLOT < (UNPACK | FLIP) >

> produces a scatter plot matrix of Y-weights against each other. Weight scatter plot matrices are by default composed of multiple plots combined into a single panel. You can use the UNPACK suboption to specify that the subplots be produced separately, and the FLIP option to interchange the default X-axis and Y-axis dimensions.

VARSCALE

> specifies that continuous model variables be centered and scaled prior to centering and scaling the model effects in which they are involved. The rescaling specified by the VARSCALE option is sometimes more appropriate if the model involves crossproducts between model variables; however, the VARSCALE option still might not produce the model you expect. See the section "Centering and Scaling" on page 4784 for more information.

VARSS

> lists, in addition to the average response and predictor sum of squares accounted for by each successive factor, the amount of variation accounted for in each response and predictor.

BY Statement

> **BY** *variables* ;

You can specify a BY statement with PROC PLS to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. The *variables* are one or more variables in the input data set.

If you specify more than one BY statement, the procedure uses only the latest BY statement and ignores any previous ones.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.

- Specify the option NOTSORTED or DESCENDING in the BY statement for the PLS procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.

- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about the BY statement, see *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the *Base SAS Procedures Guide*.

CLASS Statement

>**CLASS** *variables* < / *option* > ;

The CLASS statement names the classification variables to be used in the analysis. If the CLASS statement is used, it must appear before the MODEL statement.

Classification variables can be either character or numeric. By default, class levels are determined from the entire formatted values of the CLASS variables. Note that this represents a slight change from previous releases in the way in which class levels are determined. Prior to SAS 9, class levels were determined by using no more than the first 16 characters of the formatted values. If you want to revert to this previous behavior, you can use the TRUNCATE option in the CLASS statement. In any case, you can use formats to group values into levels. See the discussion of the FORMAT procedure in the *Base SAS Procedures Guide* and the discussions of the FORMAT statement and SAS formats in *SAS Language Reference: Dictionary*.

Any variable in the model that is not listed in the CLASS statement is assumed to be continuous. Continuous variables must be numeric.

You can specify the following option in the CLASS statement after a slash(/):

TRUNCATE
> specifies that class levels should be determined by using only up to the first 16 characters of the formatted values of CLASS variables. When formatted values are longer than 16 characters, you can use this option in order to revert to the levels as determined in releases prior to SAS 9.

ID Statement

>**ID** *variables* ;

The ID statement names variables whose values are used to label observations in plots. If you do not specify an ID statement, then each observations is labeled in plots by its corresponding observation number.

MODEL Statement

>**MODEL** *response-variables = predictor-effects* < / *options* > ;

The MODEL statement names the responses and the predictors, which determine the **Y** and **X** matrices of the model, respectively. Usually you simply list the names of the predictor variables as the model effects, but you can also use the effects notation of PROC GLM to specify polynomial effects and interactions; see the section "Specification of Effects" on page 2486 in Chapter 39, "The GLM Procedure" for further details. The MODEL statement is required. You can specify only one

MODEL statement (in contrast to the REG procedure, for example, which allows several MODEL statements in the same PROC REG run).

You can specify the following options in the MODEL statement after a slash (/).

INTERCEPT
> By default, the responses and predictors are centered; thus, no intercept is required in the model. You can specify the INTERCEPT option to override the default.

SOLUTION
> lists the coefficients of the final predictive model for the responses. The coefficients for predicting the centered and scaled responses based on the centered and scaled predictors are displayed, as well as the coefficients for predicting the raw responses based on the raw predictors.

OUTPUT Statement

> **OUTPUT** *OUT= SAS-data-set keyword=names <...keyword=names>* ;

You use the OUTPUT statement to specify a data set to receive quantities that can be computed for every input observation, such as extracted factors and predicted values. The following *keywords* are available:

PREDICTED	predicted values for responses
YRESIDUAL	residuals for responses
XRESIDUAL	residuals for predictors
XSCORE	extracted factors (X-scores, latent vectors, latent variables, T)
YSCORE	extracted responses (Y-scores, U)
STDY	standardized (centered and scaled) responses
STDX	standardized (centered and scaled) predictors
H	approximate leverage
PRESS	approximate predicted residuals
TSQUARE	scaled sum of squares of score values
STDXSSE	sum of squares of residuals for standardized predictors
STDYSSE	sum of squares of residuals for standardized responses

Suppose that there are N_x predictors and N_y responses and that the model has N_f selected factors.

- The keywords XRESIDUAL and STDX define an output variable for each predictor, so N_x names are required after each one.

- The keywords PREDICTED, YRESIDUAL, STDY, and PRESS define an output variable for each response, so N_y names are required after each of these keywords.

- The keywords XSCORE and YSCORE specify an output variable for each selected model factor. For these keywords, you provide only one base name, and the variables corresponding to each successive factor are named by appending the factor number to the base name. For example, if $N_f = 3$, then a specification of XSCORE=T would produce the variables T1, T2, and T3.

- Finally, the keywords H, TSQUARE, STDXSSE, and STDYSSE each specify a single output variable, so only one name is required after each of these keywords.

Details: PLS Procedure

Regression Methods

All of the predictive methods implemented in PROC PLS work essentially by finding linear combinations of the predictors (factors) to use to predict the responses linearly. The methods differ only in how the factors are derived, as explained in the following sections.

Partial Least Squares

Partial least squares (PLS) works by extracting one factor at a time. Let $\mathbf{X} = \mathbf{X}_0$ be the centered and scaled matrix of predictors and let $\mathbf{Y} = \mathbf{Y}_0$ be the centered and scaled matrix of response values. The PLS method starts with a linear combination $\mathbf{t} = \mathbf{X}_0\mathbf{w}$ of the predictors, where $\mathbf{t}$ is called a *score* vector and $\mathbf{w}$ is its associated *weight* vector. The PLS method predicts both $\mathbf{X}_0$ and $\mathbf{Y}_0$ by regression on $\mathbf{t}$:

$$
\begin{aligned}
\hat{\mathbf{X}}_0 &= \mathbf{t}\mathbf{p}', &\text{where} \quad \mathbf{p}' &= (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{X}_0 \\
\hat{\mathbf{Y}}_0 &= \mathbf{t}\mathbf{c}', &\text{where} \quad \mathbf{c}' &= (\mathbf{t}'\mathbf{t})^{-1}\mathbf{t}'\mathbf{Y}_0
\end{aligned}
$$

The vectors $\mathbf{p}$ and $\mathbf{c}$ are called the X- and Y-*loadings*, respectively.

The specific linear combination $\mathbf{t} = \mathbf{X}_0\mathbf{w}$ is the one that has maximum covariance $\mathbf{t}'\mathbf{u}$ with some response linear combination $\mathbf{u} = \mathbf{Y}_0\mathbf{q}$. Another characterization is that the X- and Y-weights $\mathbf{w}$ and $\mathbf{q}$ are proportional to the first left and right singular vectors of the covariance matrix $\mathbf{X}_0'\mathbf{Y}_0$ or, equivalently, the first eigenvectors of $\mathbf{X}_0'\mathbf{Y}_0\mathbf{Y}_0'\mathbf{X}_0$ and $\mathbf{Y}_0'\mathbf{X}_0\mathbf{X}_0'\mathbf{Y}_0$, respectively.

This accounts for how the first PLS factor is extracted. The second factor is extracted in the same way by replacing $\mathbf{X}_0$ and $\mathbf{Y}_0$ with the X- and Y-residuals from the first factor:

$$
\begin{aligned}
\mathbf{X}_1 &= \mathbf{X}_0 - \hat{\mathbf{X}}_0 \\
\mathbf{Y}_1 &= \mathbf{Y}_0 - \hat{\mathbf{Y}}_0
\end{aligned}
$$

These residuals are also called the *deflated* $\mathbf{X}$ and $\mathbf{Y}$ blocks. The process of extracting a score vector and deflating the data matrices is repeated for as many extracted factors as are wanted.

SIMPLS

Note that each extracted PLS factor is defined in terms of different X-variables $\mathbf{X}_i$. This leads to difficulties in comparing different scores, weights, and so forth. The SIMPLS method of de Jong (1993) overcomes these difficulties by computing each score $\mathbf{t}_i = \mathbf{X}\mathbf{r}_i$ in terms of the original (centered and scaled) predictors $\mathbf{X}$. The SIMPLS X-weight vectors r_i are similar to the eigenvectors of $\mathbf{SS}' = \mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$, but they satisfy a different orthogonality condition. The $\mathbf{r}_1$ vector is just the first eigenvector $\mathbf{e}_1$ (so that the first SIMPLS score is the same as the first PLS score), but whereas the second eigenvector maximizes

$$\mathbf{e}_1' S S' \mathbf{e}_2 \text{ subject to } \mathbf{e}_1' \mathbf{e}_2 = 0$$

the second SIMPLS weight $\mathbf{r}_2$ maximizes

$$\mathbf{r}_1' S S' \mathbf{r}_2 \text{ subject to } \mathbf{r}_1' X' X \mathbf{r}_2 = \mathbf{t}_1' \mathbf{t}_2 = 0$$

The SIMPLS scores are identical to the PLS scores for one response but slightly different for more than one response; see de Jong (1993) for details. The X- and Y-loadings are defined as in PLS, but since the scores are all defined in terms of $\mathbf{X}$, it is easy to compute the overall model coefficients $\mathbf{B}$:

$$
\begin{aligned}
\hat{\mathbf{Y}} &= \sum_i \mathbf{t}_i \mathbf{c}_i' \\
&= \sum_i \mathbf{X}\mathbf{r}_i \mathbf{c}_i' \\
&= \mathbf{XB}, \text{ where } \mathbf{B} = \mathbf{RC}'
\end{aligned}
$$

Principal Components Regression

Like the SIMPLS method, principal components regression (PCR) defines all the scores in terms of the original (centered and scaled) predictors $\mathbf{X}$. However, unlike both the PLS and SIMPLS methods, the PCR method chooses the X-weights/X-scores without regard to the response data. The X-scores are chosen to explain as much variation in $\mathbf{X}$ as possible; equivalently, the X-weights for the PCR method are the eigenvectors of the predictor covariance matrix $\mathbf{X}'\mathbf{X}$. Again, the X- and Y-loadings are defined as in PLS; but, as in SIMPLS, it is easy to compute overall model coefficients for the original (centered and scaled) responses $\mathbf{Y}$ in terms of the original predictors $\mathbf{X}$.

Reduced Rank Regression

As discussed in the preceding sections, partial least squares depends on selecting factors $\mathbf{t} = \mathbf{X}\mathbf{w}$ of the predictors and $\mathbf{u} = \mathbf{Y}\mathbf{q}$ of the responses that have maximum covariance, whereas principal components regression effectively ignores $\mathbf{u}$ and selects $\mathbf{t}$ to have maximum variance, subject to orthogonality constraints. In contrast, reduced rank regression selects $\mathbf{u}$ to account for as much variation in the *predicted* responses as possible, effectively ignoring the predictors for the purposes of factor extraction. In reduced rank regression, the Y-weights $\mathbf{q}_i$ are the eigenvectors of the covariance matrix $\hat{\mathbf{Y}}_{\text{LS}}' \hat{\mathbf{Y}}_{\text{LS}}$ of the responses predicted by ordinary least squares regression; the X-scores are the projections of the Y-scores $\mathbf{Y}\mathbf{q}_i$ onto the X space.

Relationships between Methods

When you develop a predictive model, it is important to consider not only the explanatory power of the model for current responses, but also how well sampled the predictive functions are, since this affects how well the model can extrapolate to future observations. All of the techniques implemented in the PLS procedure work by extracting successive factors, or linear combinations of the predictors, that optimally address one or both of these two goals—explaining response variation and explaining predictor variation. In particular, principal components regression selects factors that explain as much predictor variation as possible, reduced rank regression selects factors that explain as much response variation as possible, and partial least squares balances the two objectives, seeking for factors that explain both response and predictor variation.

To see the relationships between these methods, consider how each one extracts a single factor from the following artificial data set consisting of two predictors and one response:

```
data data;
   input x1 x2 y;
   datalines;
    3.37651   2.30716        0.75615
    0.74193  -0.88845        1.15285
    4.18747   2.17373        1.42392
    0.96097   0.57301        0.27433
   -1.11161  -0.75225       -0.25410
   -1.38029  -1.31343       -0.04728
    1.28153  -0.13751        1.00341
   -1.39242  -2.03615        0.45518
    0.63741   0.06183        0.40699
   -2.52533  -1.23726       -0.91080
    2.44277   3.61077       -0.82590
   ;

proc pls data=data nfac=1 method=rrr;
   model y = x1 x2;
run;

proc pls data=data nfac=1 method=pcr;
   model y = x1 x2;
run;

proc pls data=data nfac=1 method=pls;
   model y = x1 x2;
run;
```

The amount of model and response variation explained by the first factor for each method is shown in Figure 66.9 through Figure 66.11.

Figure 66.9 Variation Explained by First Reduced Rank Regression Factor

```
                         The PLS Procedure

                  Percent Variation Accounted for by
                   Reduced Rank Regression Factors

         Number of
         Extracted         Model Effects       Dependent Variables
          Factors       Current      Total     Current       Total

                 1       15.0661     15.0661    100.0000     100.0000
```

Figure 66.10 Variation Explained by First Principal Components Regression Factor

```
                         The PLS Procedure

          Percent Variation Accounted for by Principal Components

         Number of
         Extracted         Model Effects       Dependent Variables
          Factors       Current      Total     Current       Total

                 1       92.9996     92.9996     9.3787       9.3787
```

Figure 66.11 Variation Explained by First Partial Least Squares Regression Factor

```
                         The PLS Procedure

                  Percent Variation Accounted for
                  by Partial Least Squares Factors

         Number of
         Extracted         Model Effects       Dependent Variables
          Factors       Current      Total     Current       Total

                 1       88.5357     88.5357    26.5304      26.5304
```

Notice that, while the first reduced rank regression factor explains *all* of the response variation, it accounts for only about 15% of the predictor variation. In contrast, the first principal components regression factor accounts for most of the predictor variation (93%) but only 9% of the response variation. The first partial least squares factor accounts for only slightly less predictor variation than principal components but about three times as much response variation.

Figure 66.12 illustrates how partial least squares balances the goals of explaining response and predictor variation in this case.

Figure 66.12 Depiction of First Factors for Three Different Regression Methods
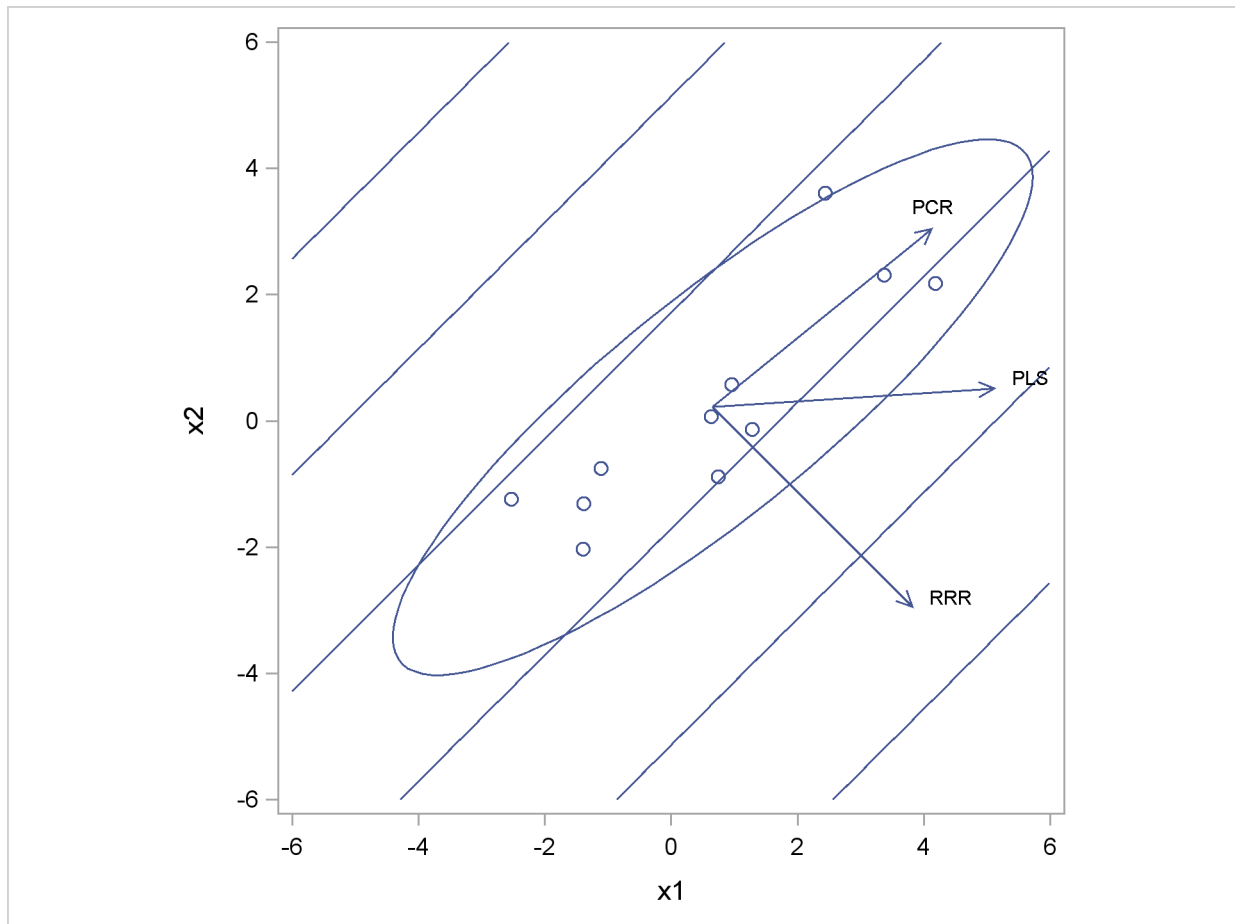
The ellipse shows the general shape of the 11 observations in the predictor space, with the contours of increasing y overlaid. Also shown are the directions of the first factor for each of the three methods. Notice that, while the predictors vary most in the x1 = x2 direction, the response changes most in the orthogonal x1 = -x2 direction. This explains why the first principal component accounts for little variation in the response and why the first reduced rank regression factor accounts for little variation in the predictors. The direction of the first partial least squares factor represents a compromise between the other two directions.

Cross Validation

None of the regression methods implemented in the PLS procedure fit the observed data any better than ordinary least squares (OLS) regression; in fact, all of the methods approach OLS as more factors are extracted. The crucial point is that, when there are many predictors, OLS can *overfit* the observed data; biased regression methods with fewer extracted factors can provide better predictability of *future* observations. However, as the preceding observations imply, the quality of the observed data fit cannot be used to choose the number of factors to extract; the number of extracted

factors must be chosen on the basis of how well the model fits observations not involved in the modeling procedure itself.

One method of choosing the number of extracted factors is to fit the model to only part of the available data (the *training set*) and to measure how well models with different numbers of extracted factors fit the other part of the data (the *test set*). This is called *test set validation*. However, it is rare that you have enough data to make both parts large enough for pure test set validation to be useful. Alternatively, you can make several different divisions of the observed data into training set and test set. This is called *cross validation*, and there are several different types. In *one-at-a-time* cross validation, the first observation is held out as a single-element test set, with all other observations as the training set; next, the second observation is held out, then the third, and so on. Another method is to hold out successive blocks of observations as test sets—for example, observations 1 through 7, then observations 8 through 14, and so on; this is known as *blocked* validation. A similar method is *split-sample* cross validation, in which successive groups of widely separated observations are held out as the test set—for example, observations {1, 11, 21, ... }, then observations {2, 12, 22, ... }, and so on. Finally, test sets can be selected from the observed data randomly; this is known as *random sample* cross validation.

Which validation you should use depends on your data. Test set validation is preferred when you have enough data to make a division into a sizable training set and test set that represent the predictive population well. You can specify that the number of extracted factors be selected by test set validation by using the CV=TESTSET(*data set*) option, where *data set* is the name of the data set containing the test set. If you do not have enough data for test set validation, you can use one of the cross validation techniques. The most common technique is one-at-a-time validation (which you can specify with the CV=ONE option or just the CV option), unless the observed data are serially correlated, in which case either blocked or split-sample validation might be more appropriate (CV=BLOCK or CV=SPLIT); you can specify the number of test sets in blocked or split-sample validation with a number in parentheses after the CV= option. Note that CV=ONE is the most computationally intensive of the cross validation methods, since it requires a recomputation of the PLS model for every input observation. Also, note that using random subset selection with CV=RANDOM might lead two different researchers to produce different PLS models on the same data (unless the same seed is used).

Whichever validation method you use, the number of factors chosen is usually the one that minimizes the predicted residual sum of squares (PRESS); this is the default choice if you specify any of the CV methods with PROC PLS. However, often models with fewer factors have PRESS statistics that are only marginally larger than the absolute minimum. To address this, van der Voet (1994) has proposed a statistical test for comparing the predicted residuals from different models; when you apply van der Voet's test, the number of factors chosen is the fewest with residuals that are insignificantly larger than the residuals of the model with minimum PRESS.

To see how van der Voet's test works, let $R_{i,jk}$ be the jth predicted residual for response k for the model with i extracted factors; the PRESS statistic is $\sum_{jk} R_{i,jk}^2$. Also, let $i_{\min}$ be the number of factors for which PRESS is minimized. The critical value for van der Voet's test is based on the differences between squared predicted residuals

$$D_{i,jk} = R_{i,jk}^2 - R_{i_{\min},jk}^2$$

One alternative for the critical value is $C_i = \sum_{jk} D_{i,jk}$, which is just the difference between the

PRESS statistics for i and $i_{\min}$ factors; alternatively, van der Voet suggests Hotelling's T^2 statistic $C_i = \mathbf{d}'_{i,.} S_i^{-1} \mathbf{d}_{i,.}$, where $\mathbf{d}_{i,.}$ is the sum of the vectors $\mathbf{d}_{i,j} = \{D_{i,j1}, \ldots, D_{i,jN_y}\}'$ and S_i is the sum of squares and crossproducts matrix

$$S_i = \sum_j \mathbf{d}_{i,j} \mathbf{d}'_{i,j}$$

Virtually, the significance level for van der Voet's test is obtained by comparing C_i with the distribution of values that result from randomly exchanging $R^2_{i,jk}$ and $R^2_{i_{\min},jk}$. In practice, a Monte Carlo sample of such values is simulated and the significance level is approximated as the proportion of simulated critical values that are greater than C_i. If you apply van der Voet's test by specifying the CVTEST option, then, by default, the number of extracted factors chosen is the least number with an approximate significance level that is greater than 0.10.

Centering and Scaling

By default, the predictors and the responses are centered and scaled to have mean 0 and standard deviation 1. Centering the predictors and the responses ensures that the criterion for choosing successive factors is based on how much *variation* they explain, in either the predictors or the responses or both. (See the section "Regression Methods" on page 4778 for more details on how different methods explain variation.) Without centering, both the mean variable value and the variation around that mean are involved in selecting factors. Scaling serves to place all predictors and responses on an equal footing relative to their variation in the data. For example, if Time and Temp are two of the predictors, then scaling says that a change of std(Time) in Time is roughly equivalent to a change of std(Temp) in Temp.

Usually, both the predictors and responses should be centered and scaled. However, if their values already represent variation around a nominal or target value, then you can use the NOCENTER option in the PROC PLS statement to suppress centering. Likewise, if the predictors or responses are already all on comparable scales, then you can use the NOSCALE option to suppress scaling.

Note that, if the predictors involve crossproduct terms, then, by default, the variables are *not* standardized before standardizing the crossproduct. That is, if the ith values of two predictors are denoted x_i^1 and x_i^2, then the default standardized ith value of the crossproduct is

$$\frac{x_i^1 x_i^2 - \text{mean}_j(x_j^1 x_j^2)}{\text{std}_j(x_j^1 x_j^2)}$$

If you want the crossproduct to be based instead on standardized variables

$$\frac{x_i^1 - m^1}{s^1} \times \frac{x_i^2 - m^2}{s^2}$$

where $m^k = \text{mean}_j(x_j^k)$ and $s^k = \text{std}_j(x_j^k)$ for $k = 1, 2$, then you should use the VARSCALE option in the PROC PLS statement. Standardizing the variables separately is usually a good idea, but unless the model also contains all crossproducts nested within each term, the resulting model

might not be equivalent to a simple linear model in the same terms. To see this, note that a model involving the crossproduct of two standardized variables

$$\frac{x_i^1 - m^1}{s^1} \times \frac{x_i^2 - m^2}{s^2} = x_i^1 x_i^2 \frac{1}{s^1 s^2} - x_i^1 \frac{m^2}{s^1 s^2} - x_i^2 \frac{m^1}{s^1 s^2} + \frac{m^1 m^2}{s^1 s^2}$$

involves both the crossproduct term and the linear terms for the unstandardized variables.

When cross validation is performed for the number of effects, there is some disagreement among practitioners as to whether each cross validation training set should be retransformed. By default, PROC PLS does so, but you can suppress this behavior by specifying the NOCVSTDIZE option in the PROC PLS statement.

Missing Values

By default, PROC PLS handles missing values very simply. Observations with any missing independent variables (including all classification variables) are excluded from the analysis, and no predictions are computed for such observations. Observations with no missing independent variables but any missing dependent variables are also excluded from the analysis, but predictions are computed.

However, the MISSING= option in the PROC PLS statement provides more sophisticated ways of modeling in the presence of missing values. If you specify MISSING=AVG or MISSING=EM, then all observations in the input data set contribute to both the analysis and the OUTPUT OUT= data set. With MISSING=AVG, the fit is computed by filling in missing values with the average of the nonmissing values for the corresponding variable. With MISSING=EM, the procedure first computes the model with MISSING=AVG, then fills in missing values with their predicted values based on that model and computes the model again. Alternatively, you can specify MISSING=EM(MAXITER=*n*) with a large value of *n* in order to perform this imputation/fit loop until convergence.

Displayed Output

By default, PROC PLS displays just the amount of predictor and response variation accounted for by each factor.

If you perform a cross validation for the number of factors by specifying the CV option in the PROC PLS statement, then the procedure displays a summary of the cross validation for each number of factors, along with information about the optimal number of factors.

If you specify the DETAILS option in the PROC PLS statement, then details of the fitted model are displayed for each successive factor. These details for each number of factors include the following:

- the predictor loadings
- the predictor weights

- the response weights

- the coded regression coefficients (for METHOD=SIMPLS, PCR, or RRR)

If you specify the CENSCALE option in the PROC PLS statement, then centering and scaling information for each response and predictor is displayed.

If you specify the VARSS option in the PROC PLS statement, the procedure displays, in addition to the average response and predictor sum of squares accounted for by each successive factor, the amount of variation accounted for in each response and predictor.

If you specify the SOLUTION option in the MODEL statement, then PROC PLS displays the coefficients of the final predictive model for the responses. The coefficients for predicting the centered and scaled responses based on the centered and scaled predictors are displayed, as well as the coefficients for predicting the raw responses based on the raw predictors.

ODS Table Names

PROC PLS assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in Table 66.1. For more information about ODS, see Chapter 20, "Using the Output Delivery System."

Table 66.1   ODS Tables Produced by PROC PLS

ODS Table Name	Description	Statement	Option
CVResults	Results of cross validation	PROC	CV
CenScaleParms	Parameter estimates for centered and scaled data	MODEL	SOLUTION
CodedCoef	Coded coefficients	PROC	DETAILS
MissingIterations	Iterations for missing value imputation	PROC	MISSING=EM
ModelInfo	Model information	PROC	default
NObs	Number of observations	PROC	default
ParameterEstimates	Parameter estimates for raw data	MODEL	SOLUTION
PercentVariation	Variation accounted for by each factor	PROC	default
ResidualSummary	Residual summary from cross validation	PROC	CV
XEffectCenScale	Centering and scaling information for predictor effects	PROC	CENSCALE
XLoadings	Loadings for independents	PROC	DETAILS
XVariableCenScale	Centering and scaling information for predictor variables	PROC	CENSCALE and VARSCALE
XWeights	Weights for independents	PROC	DETAILS
YVariableCenScale	Centering and scaling information for responses	PROC	CENSCALE
YWeights	Weights for dependents	PROC	DETAILS

ODS Graphics

This section describes the use of ODS for creating statistical graphs with the PLS procedure. To request these graphs you must specify the ODS GRAPHICS statement. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

When the ODS GRAPHICS are in effect, by default the PLS procedure produces a plot of the variation accounted for by each extracted factor, as well as a *correlation loading plot* for the first two extracted factors (if the final model has at least two factors). The plot of the variation accounted for can take several forms:

- If the PLS analysis does not include cross validation, then the plot shows the total R square for both model effects and the dependent variables against the number of factors.

- If you specify the CV= option to select the number of factors in the final model by cross validation, then the plot shows the R-square analysis discussed previously as well as the root mean PRESS from the cross validation analysis, with the selected number of factors identified by a vertical line.

The correlation loading plot for the first two factors summarizes many aspects of the two most significant dimensions of the model. It consists of overlaid scatter plots of the scores of the first two factors, the loadings of the model effects, and the loadings of the dependent variables. The loadings are scaled so that the amount of variation in the variables that is explained by the model is proportional to the distance from the origin; circles indicating various levels of explained variation are also overlaid on the correlation loading plot. Also, the correlation between the model approximations for any two variables is proportional to the length of the projection of the point corresponding to one variable on a line through the origin passing through the point corresponding to the other variable; the sign of the correlation corresponds to which side of the origin the projected point falls on.

The R square and the first two correlation loadings are plotted by default when the ODS GRAPHICS are in effect, but you can produce many other plots for the PROC PLS analysis.

ODS Graph Names

PROC PLS assigns a name to each graph it creates using ODS. You can use these names to reference the graphs when using ODS. The names are listed in Table 66.2.

To request these graphs you must specify the ODS GRAPHICS statement. For more information about the ODS GRAPHICS statement, see Chapter 21, "Statistical Graphics Using ODS."

Table 66.2 ODS Graphics Produced by PROC GLM

ODS Graph Name	Plot Description	Option
CorrLoadPlot	Correlation loading plot (default)	PLOT=CORRLOAD(*option*)
CVPlot	Cross validation and R-square analysis (default, as appropriate)	CV=

Table 66.2 *continued*

ODS Graph Name	Plot Description	Option
DModXPlot	Distance of each observation to the X model	PLOT=DMODX
DModXYPlot	Distance of each observation to the X and Y models	PLOT=DMODXY
DModYPlot	Distance of each observation to the Y model	PLOT=DMODY
DiagnosticsPanel	Panel of summary diagnostics for the fit	PLOT=DIAGNOSTICS
AbsResidualByPredicted	Absolute residual by predicted values	PLOT=DIAGNOSTICS(UNPACK)
ObservedByPredicted	Observed by predicted	PLOT=DIAGNOSTICS(UNPACK)
QQPlot	Residual Q-Q plot	PLOT=DIAGNOSTICS(UNPACK)
ResidualByPredicted	Residual by predicted values	PLOT=DIAGNOSTICS(UNPACK)
ResidualHistogram	Residual histogram	PLOT=DIAGNOSTICS(UNPACK)
RFPlot	RF plot	PLOT=DIAGNOSTICS(UNPACK)
ParmProfiles	Profiles of regression coefficients	PLOT=PARMPROFILES
R2Plot	R-square analysis (default, as appropriate)	
ResidualPlots	Residuals for each dependent variable	PLOT=RESIDUALS
VariableImportancePlot	Profile of variable importance factors	PLOT=VIP
XLoadingPlot	Scatter plot matrix of X-loadings against each other	PLOT=XLOADINGPLOT
XLoadingProfiles	Profiles of the X-loadings	PLOT=XLOADINGPROFILES
XScorePlot	Scatter plot matrix of X-scores against each other	PLOT=XSCORES
XWeightPlot	Scatter plot matrix of X-weights against each other	PLOT=XWEIGHTPLOT
XWeightProfiles	Profiles of the X-weights	PLOT=XWEIGHTPROFILES
XYScorePlot	Scatter plot matrix of X-scores against Y-scores	PLOT=XYSCORES
YScorePlot	Scatter plot matrix of Y-scores against each other	PLOT=YSCORES
YWeightPlot	Scatter plot matrix of Y-weights against each other	PLOT=YWEIGHTPLOT

Examples: PLS Procedure

Example 66.1: Examining Model Details

This example, from Umetrics (1995), demonstrates different ways to examine a PLS model. The data come from the field of drug discovery. New drugs are developed from chemicals that are biologically active. Testing a compound for biological activity is an expensive procedure, so it is useful to be able to predict biological activity from cheaper chemical measurements. In fact, computational chemistry makes it possible to calculate certain chemical measurements without even making the compound. These measurements include size, lipophilicity, and polarity at various sites on the molecule. The following statements create a data set named pentaTrain, which contains these data.

```
data pentaTrain;
   input obsnam $ S1 L1 P1 S2 L2 P2
                 S3 L3 P3 S4 L4 P4
                 S5 L5 P5  log_RAI @@;
   n = _n_;
   datalines;
VESSK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
          1.9607 -1.6324  0.5746   1.9607 -1.6324   0.5746
          2.8369  1.4092 -3.1398                     0.00
VESAK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
          1.9607 -1.6324  0.5746   0.0744 -1.7333   0.0902
          2.8369  1.4092 -3.1398                     0.28
VEASK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
          0.0744 -1.7333  0.0902   1.9607 -1.6324   0.5746
          2.8369  1.4092 -3.1398                     0.20
VEAAK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
          0.0744 -1.7333  0.0902   0.0744 -1.7333   0.0902
          2.8369  1.4092 -3.1398                     0.51
VKAAK    -2.6931 -2.5271 -1.2871   2.8369   1.4092 -3.1398
          0.0744 -1.7333  0.0902   0.0744 -1.7333   0.0902
          2.8369  1.4092 -3.1398                     0.11
VEWAK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
         -4.7548  3.6521  0.8524   0.0744 -1.7333   0.0902
          2.8369  1.4092 -3.1398                     2.73
VEAAP    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
          0.0744 -1.7333  0.0902   0.0744 -1.7333   0.0902
         -1.2201  0.8829  2.2253                     0.18
VEHAK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
          2.4064  1.7438  1.1057   0.0744 -1.7333   0.0902
          2.8369  1.4092 -3.1398                     1.53
VAAAK    -2.6931 -2.5271 -1.2871   0.0744 -1.7333   0.0902
          0.0744 -1.7333  0.0902   0.0744 -1.7333   0.0902
          2.8369  1.4092 -3.1398                    -0.10
GEAAK     2.2261 -5.3648  0.3049   3.0777   0.3891 -0.0701
          0.0744 -1.7333  0.0902   0.0744 -1.7333   0.0902
          2.8369  1.4092 -3.1398                    -0.52
```

```
LEAAK    -4.1921 -1.0285 -0.9801   3.0777   0.3891 -0.0701
          0.0744 -1.7333  0.0902   0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                     0.40
FEAAK    -4.9217  1.2977  0.4473   3.0777   0.3891 -0.0701
          0.0744 -1.7333  0.0902   0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                     0.30
VEGGK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
          2.2261 -5.3648  0.3049   2.2261 -5.3648  0.3049
          2.8369  1.4092 -3.1398                    -1.00
VEFAK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
         -4.9217  1.2977  0.4473   0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                     1.57
VELAK    -2.6931 -2.5271 -1.2871   3.0777   0.3891 -0.0701
         -4.1921 -1.0285 -0.9801   0.0744 -1.7333  0.0902
          2.8369  1.4092 -3.1398                     0.59
;
```

You would like to study the relationship between these measurements and the activity of the compound, represented by the logarithm of the relative Bradykinin activating activity (log_RAI). Notice that these data consist of many predictors relative to the number of observations. Partial least squares is especially appropriate in this situation as a useful tool for finding a few underlying predictive factors that account for most of the variation in the response. Typically, the model is fit for part of the data (the "training" or "work" set), and the quality of the fit is judged by how well it predicts the other part of the data (the "test" or "prediction" set). For this example, the first 15 observations serve as the training set and the rest constitute the test set (refer to Ufkes et al. 1978, 1982).

When you fit a PLS model, you hope to find a few PLS factors that explain most of the variation in both predictors and responses. Factors that explain response variation provide good predictive models for new responses, and factors that explain predictor variation are well represented by the observed values of the predictors. The following statements fit a PLS model with two factors and save predicted values, residuals, and other information for each data point in a data set named outpls.

```
proc pls data=pentaTrain;
   model log_RAI = S1-S5 L1-L5 P1-P5;
run;
```

The PLS procedure displays a table, shown in Output 66.1.1, showing how much predictor and response variation is explained by each PLS factor.

Output 66.1.1 Amount of Training Set Variation Explained

```
                        The PLS Procedure

                  Percent Variation Accounted for
                  by Partial Least Squares Factors

        Number of
        Extracted        Model Effects      Dependent Variables
         Factors      Current      Total     Current      Total

             1        16.9014     16.9014    89.6399      89.6399
             2        12.7721     29.6735     7.8368      97.4767
             3        14.6554     44.3289     0.4636      97.9403
             4        11.8421     56.1710     0.2485      98.1889
             5        10.5894     66.7605     0.1494      98.3383
             6         5.1876     71.9481     0.2617      98.6001
             7         6.1873     78.1354     0.2428      98.8428
             8         7.2252     85.3606     0.1926      99.0354
             9         6.7285     92.0891     0.0725      99.1080
            10         7.9076     99.9967     0.0000      99.1080
            11         0.0033    100.0000     0.0099      99.1179
            12         0.0000    100.0000     0.0000      99.1179
            13         0.0000    100.0000     0.0000      99.1179
            14         0.0000    100.0000     0.0000      99.1179
            15         0.0000    100.0000     0.0000      99.1179
```

From Output 66.1.1, note that 97% of the response variation is already explained by just two factors, but only 29% of the predictor variation is explained.

The graphics in PROC PLS, available when ODS Graphics is in effect, make it easier to see features of the PLS model.

If you enable ODS Graphics, then in addition to the tables discussed previously, PROC PLS displays a graphical depiction of the R-square analysis as well as a correlation loadings plot summarizing the model based on the first two PLS factors. The following statements perform the previous analysis with ODS Graphics enabled, producing Output 66.1.2 and Output 66.1.3.

```
ods graphics on;

proc pls data=pentaTrain;
   model log_RAI = S1-S5 L1-L5 P1-P5;
run;

ods graphics off;
```

Output 66.1.2 Plot of Proportion of Variation Accounted For
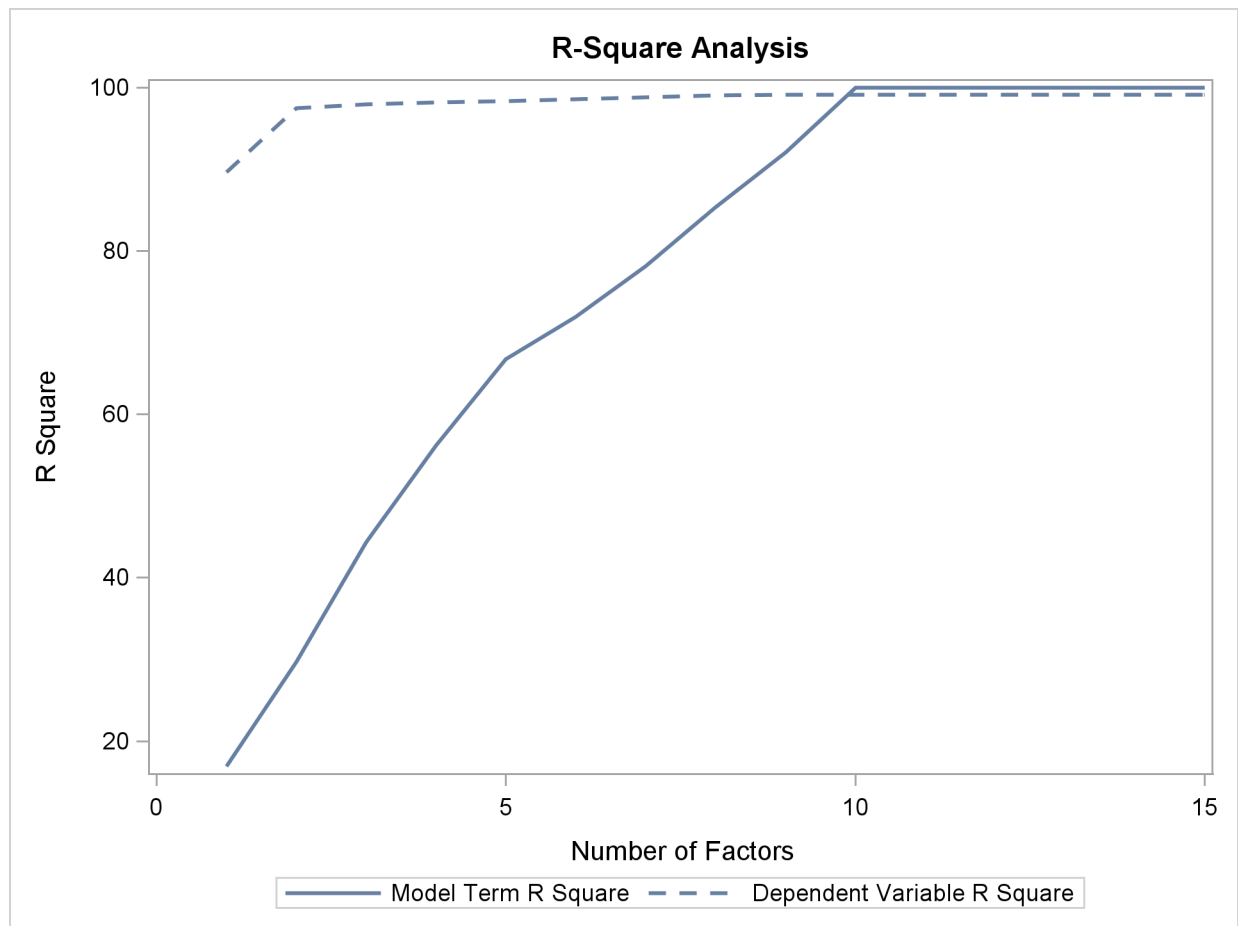
Output 66.1.3 Correlation Loadings Plot
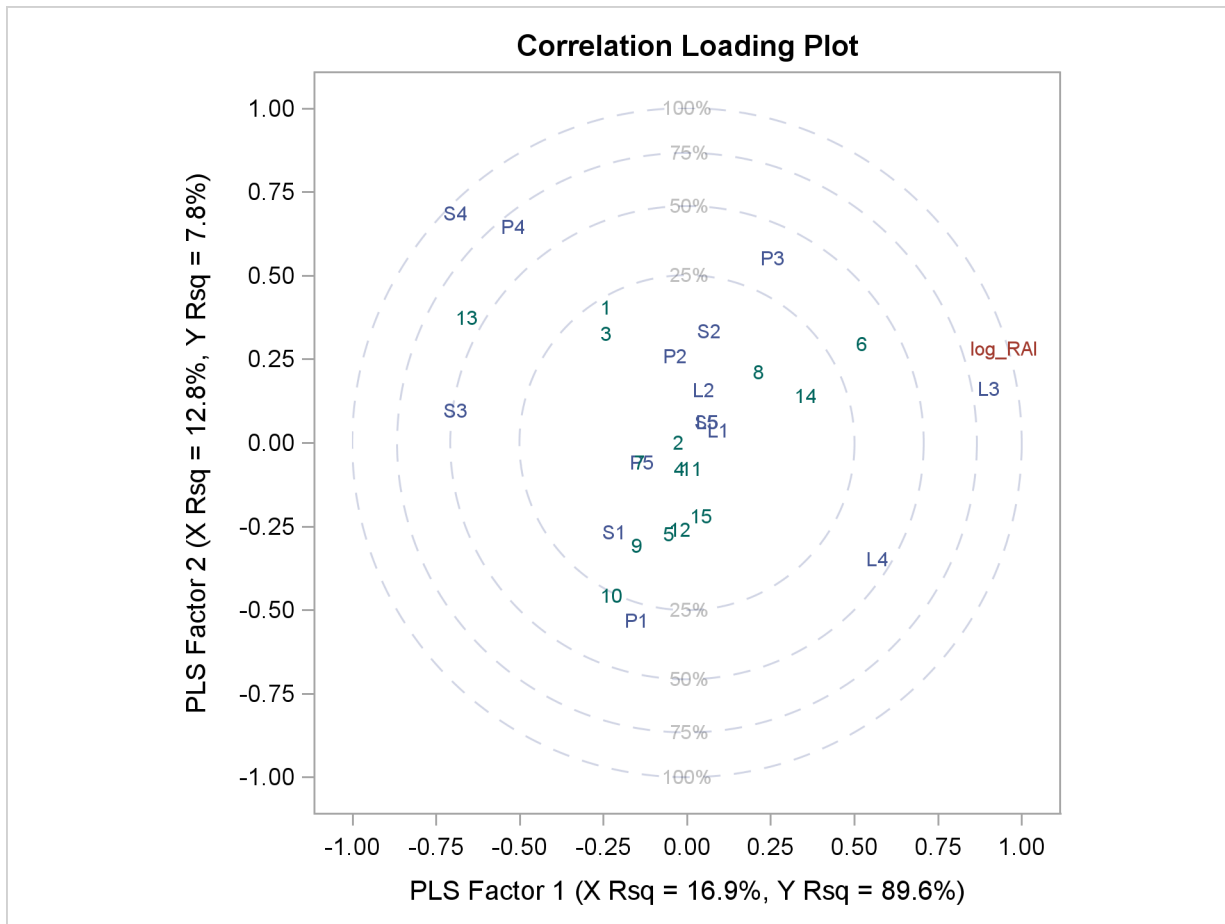
The plot in Output 66.1.2 of the proportion of variation explained (or R square) makes it clear that there is a plateau in the response variation after two factors are included in the model. The correlation loading plot in Output 66.1.3 summarizes many features of this two-factor model, including the following:

- The X-scores are plotted as numbers for each observation. You should look for patterns or clearly grouped observations. If you see a curved pattern, for example, you might want to add a quadratic term. Two or more groupings of observations indicate that it might be better to analyze the groups separately, perhaps by including classification effects in the model. This plot appears to show most of the observations close together, with a few being more spread out with larger positive X-scores for factor 2. There are no clear grouping patterns, but observation 13 stands out.

- The loadings show how much variation in each variable is accounted for by the first two factors, jointly by the distance of the corresponding point from the origin and individually by the distance for the projections of this point onto the horizontal and vertical axes. That the dependent variable is well explained by the model is reflected in the fact that the point for log_RAI is near the 100% circle.

- You can also use the projection interpretation to relate variables to each other. For example, projecting other variables' points onto the line that runs through the log_RAI point and the origin, you can see that the PLS approximation for the predictor L3 is highly positively correlated with log_RAI, S3 is somewhat less correlated but in the negative direction, and several predictors including L1, L5, and S5 have very little correlation with log_RAI.

Other graphics enable you to explore more of the features of the PLS model. For example, you can examine the X-scores versus the Y-scores to explore how partial least squares chooses successive factors. For a good PLS model, the first few factors show a high correlation between the X- and Y-scores. The correlation usually decreases from one factor to the next. When ODS Graphics is in effect, you can plot the X-scores versus the Y-scores by using the PLOT=XYSCORES option, as shown in the following statements.

```
ods graphics on;

proc pls data=pentaTrain nfac=4 plot=XYScores;
   model log_RAI = S1-S5 L1-L5 P1-P5;
run;

ods graphics off;
```

The plot of the X-scores versus the Y-scores for the first four factors is shown in Output 66.1.4.

Output 66.1.4 X-Scores versus Y-Scores
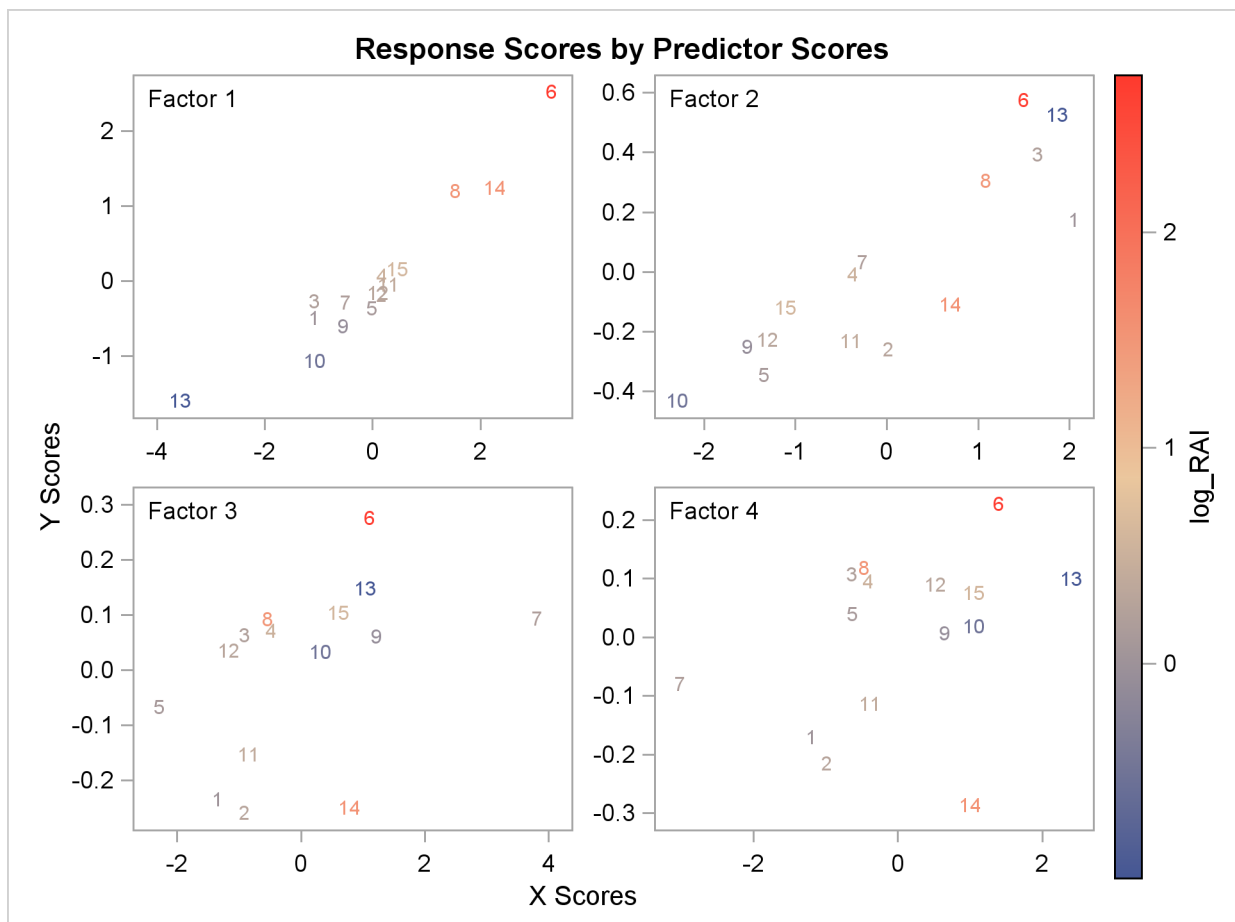
For this example, Output 66.1.4 shows high correlation between X- and Y-scores for the first factor but somewhat lower correlation for the second factor and sharply diminishing correlation after that. This adds strength to the judgment that NFAC=2 is the right number of factors for these data and this model. Note that observation 13 is again extreme in the first two plots. This run might be overly influential for the PLS analysis; thus, you should check to make sure it is reliable.

As explained earlier, you can draw some inferences about the relationship between individual predictors and the dependent variable from the correlation loading plot. However, the regression coefficient profile and the variable importance plot give a more direct indication of which predictors are most useful for predicting the dependent variable. The regression coefficients represent the importance each predictor has in the prediction of just the response. The variable importance plot, on the other hand, represents the contribution of each predictor in fitting the PLS model for both predictors and response. It is based on the *Variable Importance for Projection* (VIP) statistic of Wold (1994), which summarizes the contribution a variable makes to the model. If a predictor has a relatively small coefficient (in absolute value) *and* a small value of VIP, then it is a prime candidate for deletion. Wold in Umetrics (1995) considers a value less than 0.8 to be "small" for the VIP. The following statements fit a two-factor PLS model and display these two additional plots.

```
ods graphics on;

proc pls data=pentaTrain nfac=2 plot=(ParmProfiles VIP);
   model log_RAI = S1-S5 L1-L5 P1-P5;
run;

ods graphics off;
```

The additional graphics are shown in Output 66.1.5 and Output 66.1.6.

Output 66.1.5 Variable Importance Plots
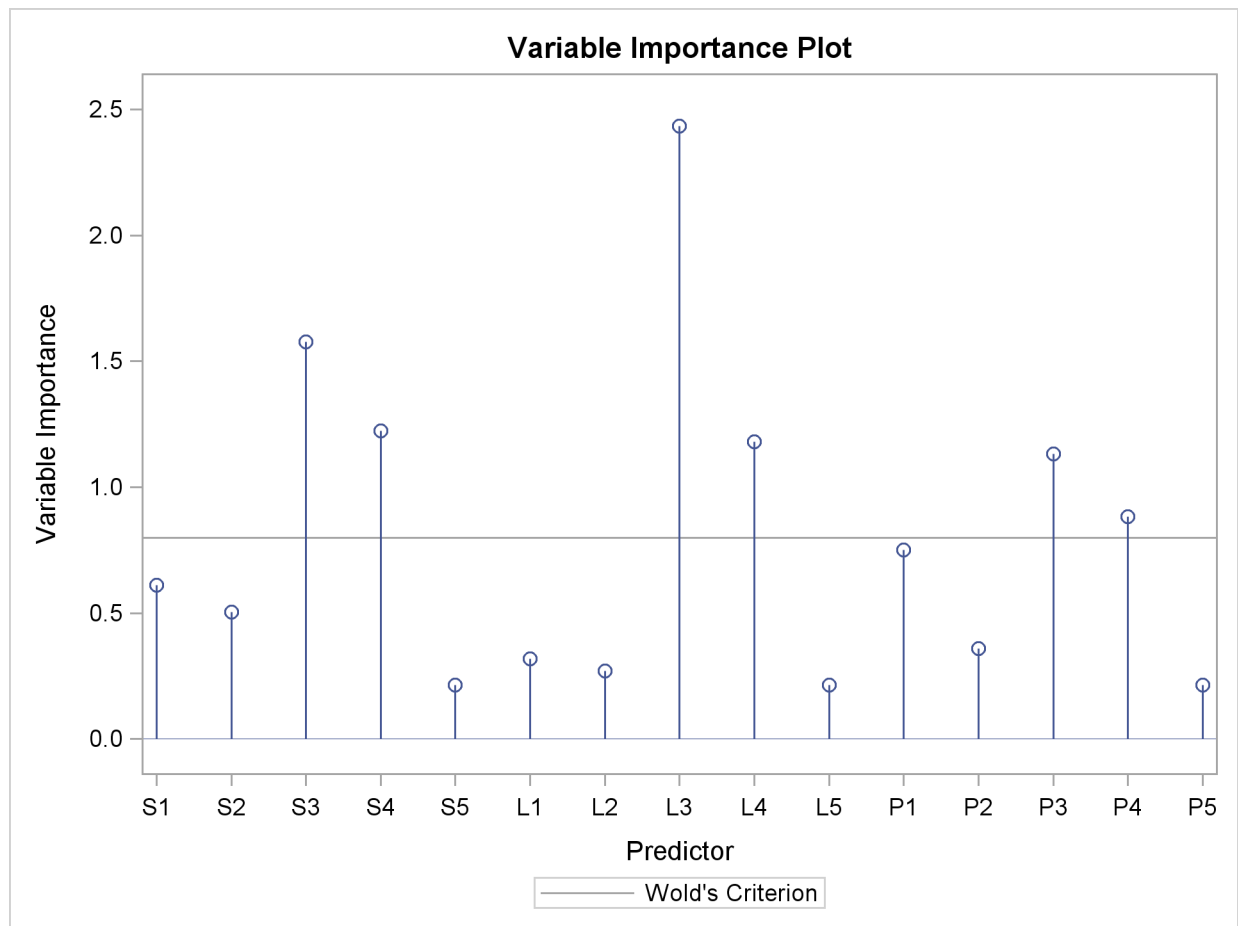
Output 66.1.6 Regression Parameter Profile
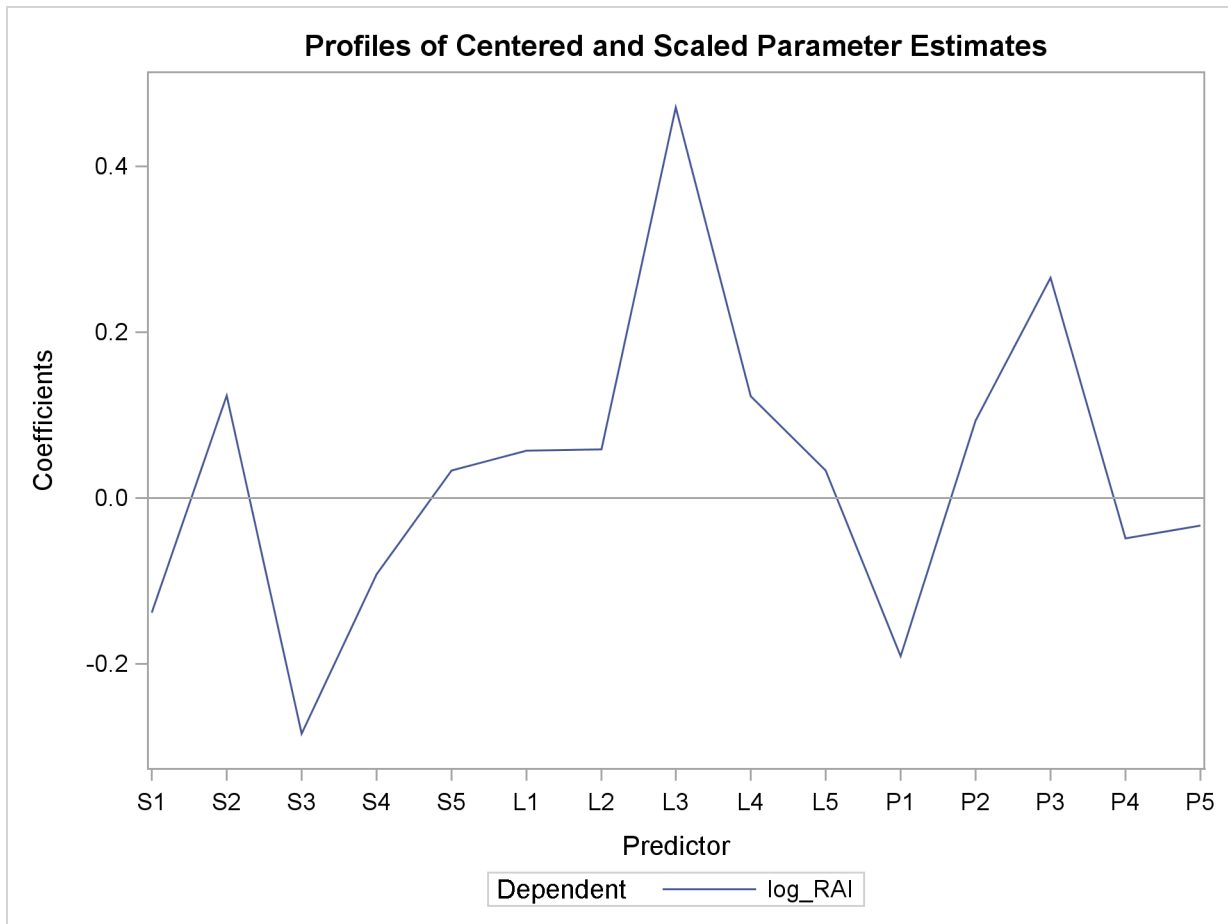
In these two plots, the variables L1, L2, P2, S5, L5, and P5 have small absolute coefficients and small VIP. Looking back at the correlation loadings plot in Output 66.1.2, you can see that these variables tend to be the ones near zero for both PLS factors. You should consider dropping these variables from the model.

Example 66.2: Examining Outliers

This example is a continuation of Example 66.1.

Standard diagnostics for statistical models focus on the response, allowing you to look for patterns that indicate the model is inadequate or for outliers that do not seem to follow the trend of the rest of the data. However, partial least squares effectively models the predictors as well as the responses, so you should consider the pattern of the fit for both. The DModX and DModY statistics give the distance from each point to the PLS model with respect to the predictors and the responses, respectively, and ODS Graphics enables you to plot these values. No point should be dramatically farther from the model than the rest. If there is a group of points that are all farther from the

model than the rest, they might have something in common, in which case they should be analyzed separately.

The following statements fit a reduced model to the data discussed in Example 66.1 and plot a panel of standard diagnostics as well as the distances of the observations to the model.

```
ods graphics on;

proc pls data=pentaTrain nfac=2 plot=(diagnostics dmod);
   model log_RAI = S1    P1
                   S2
                   S3 L3 P3
                   S4 L4   ;
run;

ods graphics off;
```

The plots are shown in Output 66.2.1 and Output 66.2.2.

Output 66.2.1 Model Fit Diagnostics
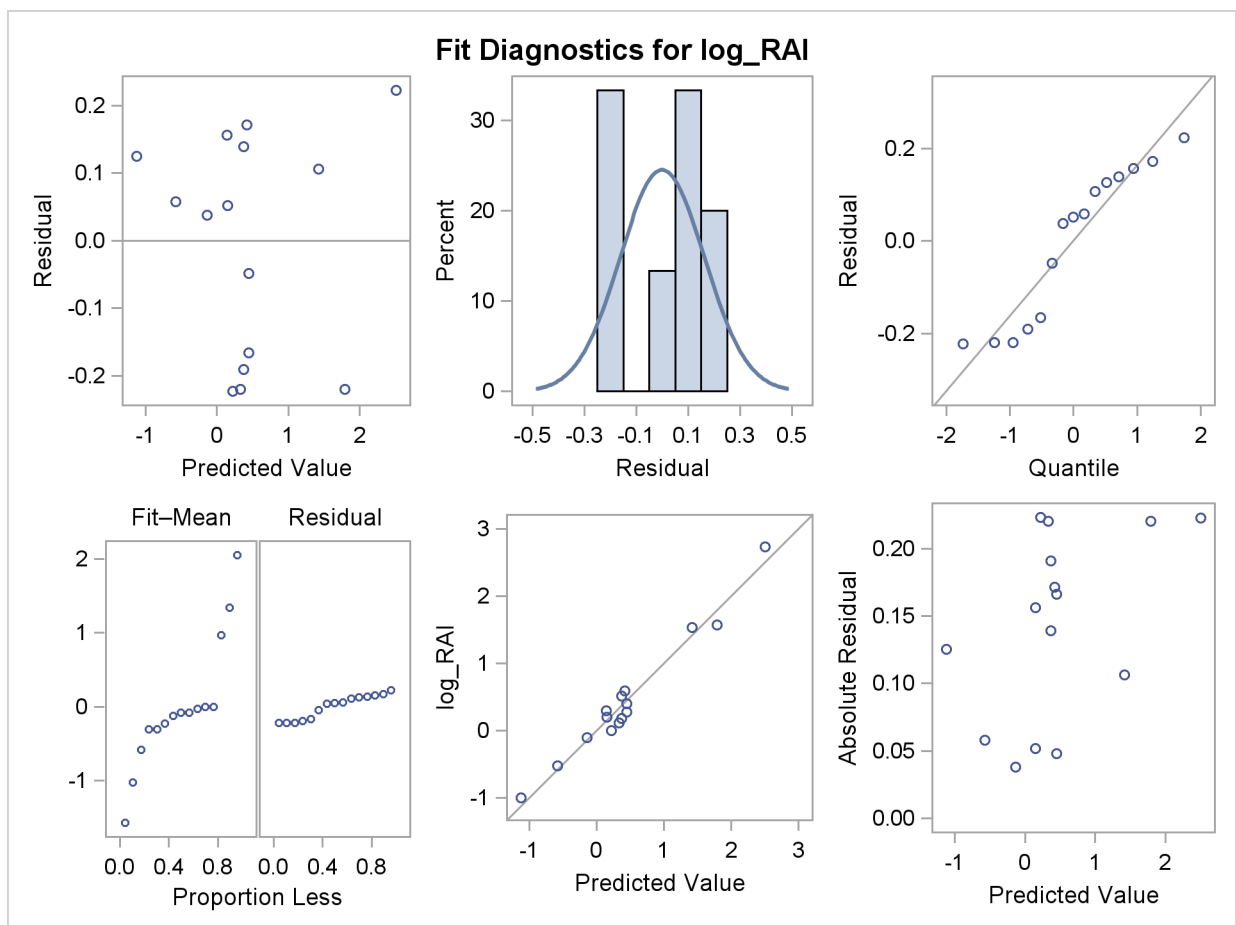
Output 66.2.2 Predictor versus Response Distances to the Model
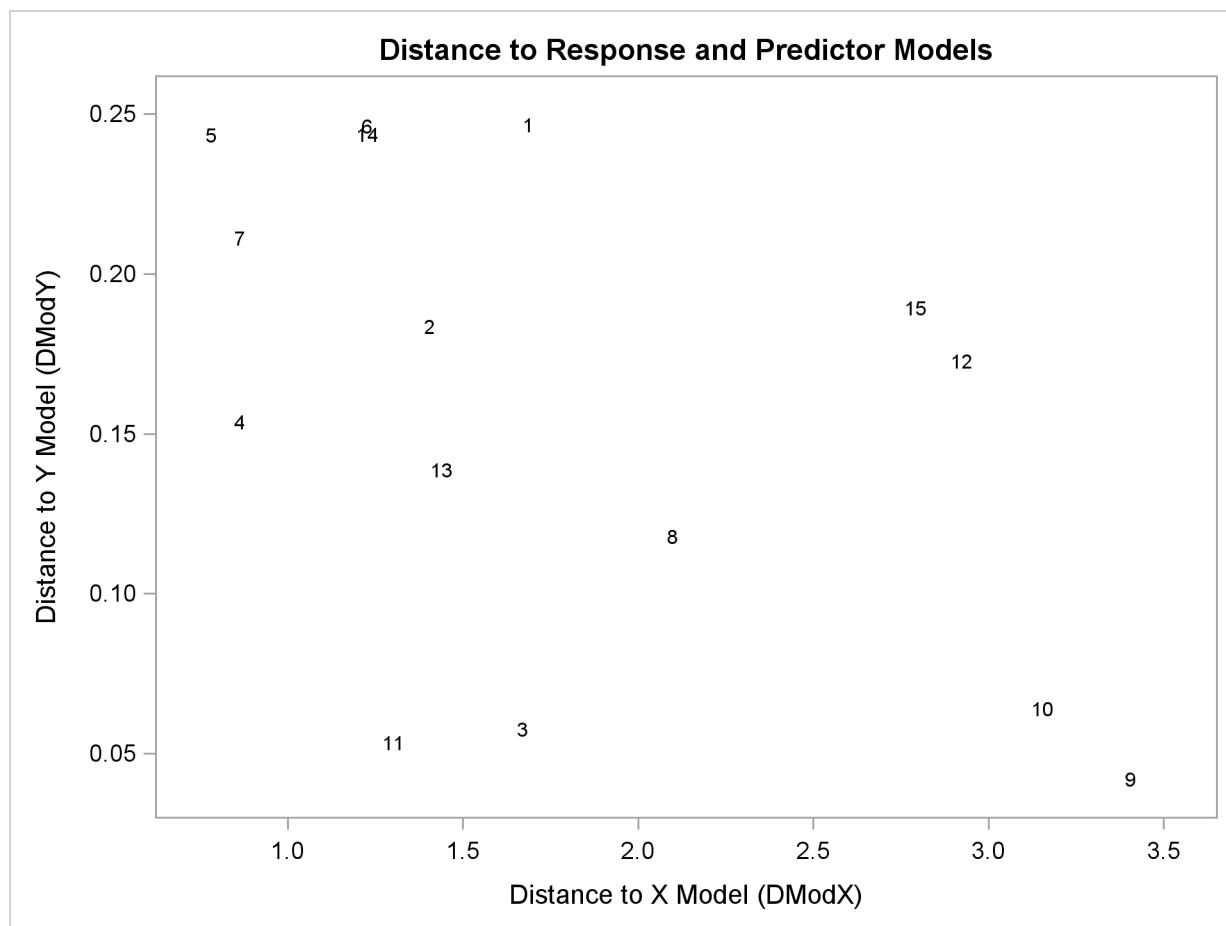
There appear to be no profound outliers in either the predictor space or the response space.

Example 66.3: Choosing a PLS Model by Test Set Validation

This example demonstrates issues in spectrometric calibration. The data (Umetrics 1995) consist of spectrographic readings on 33 samples containing known concentrations of two amino acids, tyrosine and tryptophan. The spectra are measured at 30 frequencies across the overall range of frequencies. For example, Figure 66.3.1 shows the observed spectra for three samples, one with only tryptophan, one with only tyrosine, and one with a mixture of the two, all at a total concentration of 10^{-6}.

Output 66.3.1 Spectra for Three Samples of Tyrosine and Tryptophan
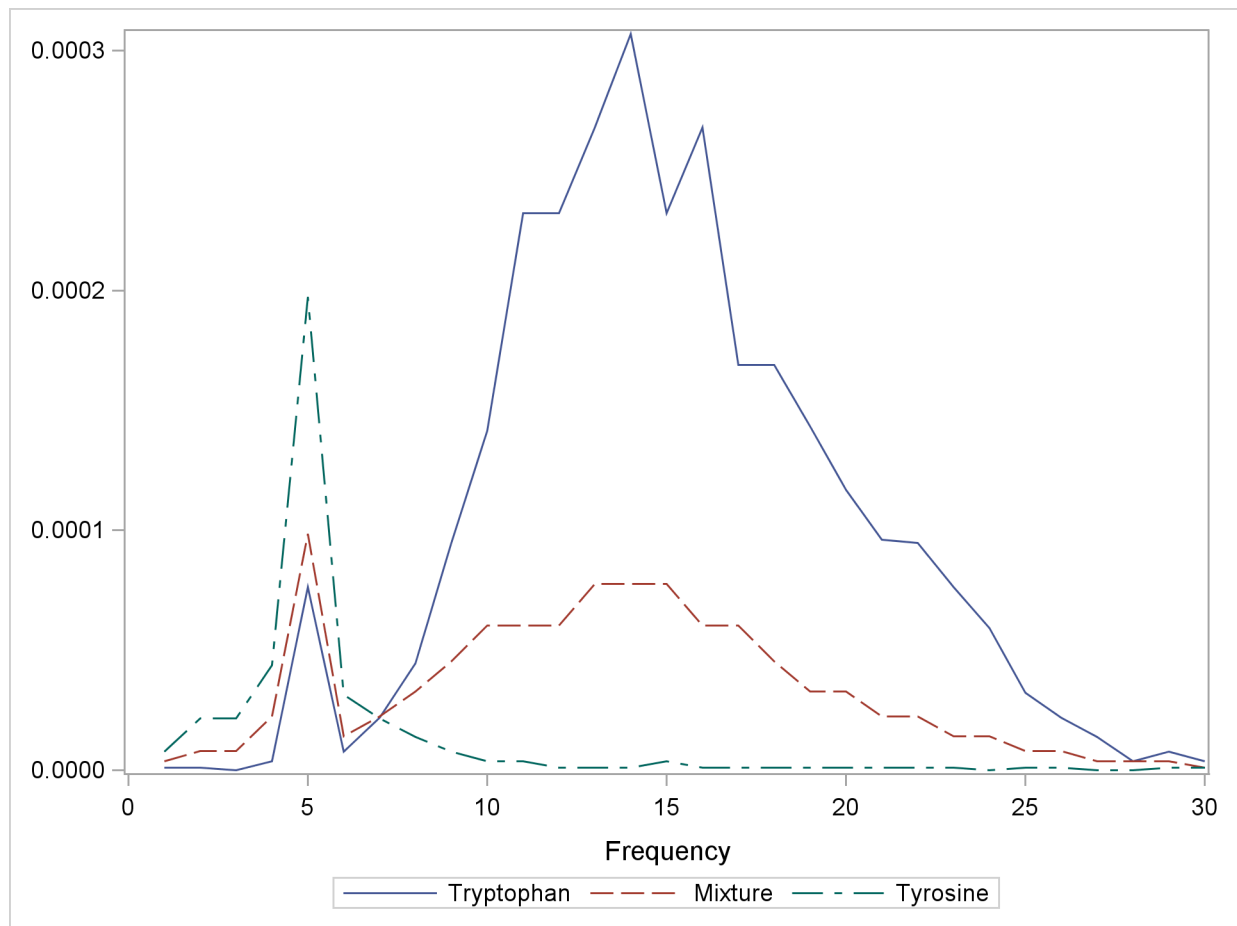
Of the 33 samples, 18 are used as a training set and 15 as a test set. The data originally appear in McAvoy et al. (1989).

These data were created in a lab, with the concentrations fixed in order to provide a wide range of applicability for the model. You want to use a linear function of the logarithms of the spectra to predict the logarithms of tyrosine and tryptophan concentration, as well as the logarithm of the total concentration. Actually, because of the possibility of zeros in both the responses and the predictors, slightly different transformations are used. The following statements create SAS data sets containing the training and test data, named ftrain and ftest, respectively.

```
data ftrain;
   input obsnam $ tot tyr f1-f30 @@;
   try = tot - tyr;
   if (tyr) then tyr_log = log10(tyr); else tyr_log = -8;
   if (try) then try_log = log10(try); else try_log = -8;
   tot_log = log10(tot);
   datalines;
17mix35 0.00003 0
 -6.215 -5.809 -5.114 -3.963 -2.897 -2.269 -1.675 -1.235
 -0.900 -0.659 -0.497 -0.395 -0.335 -0.315 -0.333 -0.377
 -0.453 -0.549 -0.658 -0.797 -0.878 -0.954 -1.060 -1.266
 -1.520 -1.804 -2.044 -2.269 -2.496 -2.714
```

```
19mix35 0.00003 3E-7
 -5.516 -5.294 -4.823 -3.858 -2.827 -2.249 -1.683 -1.218
 -0.907 -0.658 -0.501 -0.400 -0.345 -0.323 -0.342 -0.387
 -0.461 -0.554 -0.665 -0.803 -0.887 -0.960 -1.072 -1.272
 -1.541 -1.814 -2.058 -2.289 -2.496 -2.712
21mix35 0.00003 7.5E-7
 -5.519 -5.294 -4.501 -3.863 -2.827 -2.280 -1.716 -1.262
 -0.939 -0.694 -0.536 -0.444 -0.384 -0.369 -0.377 -0.421
 -0.495 -0.596 -0.706 -0.824 -0.917 -0.988 -1.103 -1.294
 -1.565 -1.841 -2.084 -2.320 -2.521 -2.729

... more lines ...

mix6    0.0001 0.00009
 -1.140 -0.757 -0.497 -0.362 -0.329 -0.412 -0.513 -0.647
 -0.772 -0.877 -0.958 -1.040 -1.104 -1.162 -1.233 -1.317
 -1.425 -1.543 -1.661 -1.804 -1.877 -1.959 -2.034 -2.249
 -2.502 -2.732 -2.964 -3.142 -3.313 -3.576
;
data ftest;
   input obsnam $ tot tyr f1-f30 @@;
   try = tot - tyr;
   if (tyr) then tyr_log = log10(tyr); else tyr_log = -8;
   if (try) then try_log = log10(try); else try_log = -8;
   tot_log = log10(tot);
   datalines;
43trp6  1E-6 0
 -5.915 -5.918 -6.908 -5.428 -4.117 -5.103 -4.660 -4.351
 -4.023 -3.849 -3.634 -3.634 -3.572 -3.513 -3.634 -3.572
 -3.772 -3.772 -3.844 -3.932 -4.017 -4.023 -4.117 -4.227
 -4.492 -4.660 -4.855 -5.428 -5.103 -5.428
59mix6  1E-6 1E-7
 -5.903 -5.903 -5.903 -5.082 -4.213 -5.083 -4.838 -4.639
 -4.474 -4.213 -4.001 -4.098 -4.001 -4.001 -3.907 -4.001
 -4.098 -4.098 -4.206 -4.098 -4.213 -4.213 -4.335 -4.474
 -4.639 -4.838 -4.837 -5.085 -5.410 -5.410
51mix6  1E-6 2.5E-7
 -5.907 -5.907 -5.415 -4.843 -4.213 -4.843 -4.843 -4.483
 -4.343 -4.006 -4.006 -3.912 -3.830 -3.830 -3.755 -3.912
 -4.006 -4.001 -4.213 -4.213 -4.335 -4.483 -4.483 -4.642
 -4.841 -5.088 -5.088 -5.415 -5.415 -5.415

... more lines ...

tyro2   0.0001 0.0001
 -1.081 -0.710 -0.470 -0.337 -0.327 -0.433 -0.602 -0.841
 -1.119 -1.423 -1.750 -2.121 -2.449 -2.818 -3.110 -3.467
 -3.781 -4.029 -4.241 -4.366 -4.501 -4.366 -4.501 -4.501
 -4.668 -4.668 -4.865 -4.865 -5.109 -5.111
;
```

The following statements fit a PLS model with 10 factors.

```
proc pls data=ftrain nfac=10;
   model tot_log tyr_log try_log = f1-f30;
run;
```

The table shown in Output 66.3.2 indicates that only three or four factors are required to explain almost all of the variation in both the predictors and the responses.

Output 66.3.2 Amount of Training Set Variation Explained

```
                          The PLS Procedure

                     Percent Variation Accounted for
                     by Partial Least Squares Factors

         Number of
         Extracted        Model Effects        Dependent Variables
          Factors      Current      Total      Current      Total

             1         81.1654     81.1654     48.3385     48.3385
             2         16.8113     97.9768     32.5465     80.8851
             3          1.7639     99.7407     11.4438     92.3289
             4          0.1951     99.9357      3.8363     96.1652
             5          0.0276     99.9634      1.6880     97.8532
             6          0.0132     99.9765      0.7247     98.5779
             7          0.0052     99.9817      0.2926     98.8705
             8          0.0053     99.9870      0.1252     98.9956
             9          0.0049     99.9918      0.1067     99.1023
            10          0.0034     99.9952      0.1684     99.2707
```

In order to choose the optimal number of PLS factors, you can explore how well models based on the training data with different numbers of factors fit the test data. To do so, use the CV=TESTSET option, with an argument pointing to the test data set ftest. The following statements also employ the ODS Graphics features in PROC PLS to display the cross validation results in a plot.

```
ods graphics on;

proc pls data=ftrain nfac=10 cv=testset(ftest)
                        cvtest(stat=press seed=12345);
   model tot_log tyr_log try_log = f1-f30;
run;

ods graphics off;
```

The tabular results of the test set validation are shown in Output 66.3.3, and the graphical results are shown in Output 66.3.4. They indicate that, although five PLS factors give the minimum predicted residual sum of squares, the residuals for four factors are insignificantly different from those for five. Thus, the smaller model is preferred.

Output 66.3.3 Test Set Validation for the Number of PLS Factors

```
                         The PLS Procedure

          Test Set Validation for the Number of Extracted Factors

              Number of         Root
              Extracted         Mean       Prob >
               Factors         PRESS        PRESS

                     0        3.056797      <.0001
                     1        2.630561      <.0001
                     2         1.00706      0.0070
                     3        0.664603      0.0020
                     4        0.521578      0.3800
                     5        0.500034      1.0000
                     6        0.513561      0.5100
                     7        0.501431      0.6870
                     8        1.055791      0.1530
                     9        1.435085      0.1010
                    10        1.720389      0.0320


            Minimum root mean PRESS                        0.5000
            Minimizing number of factors                        5
            Smallest number of factors with p > 0.1            4


                    Percent Variation Accounted for
                    by Partial Least Squares Factors

        Number of
        Extracted         Model Effects         Dependent Variables
         Factors      Current      Total        Current       Total

               1       81.1654     81.1654       48.3385      48.3385
               2       16.8113     97.9768       32.5465      80.8851
               3        1.7639     99.7407       11.4438      92.3289
               4        0.1951     99.9357        3.8363      96.1652
```

Output 66.3.4 Test Set Validation Plot
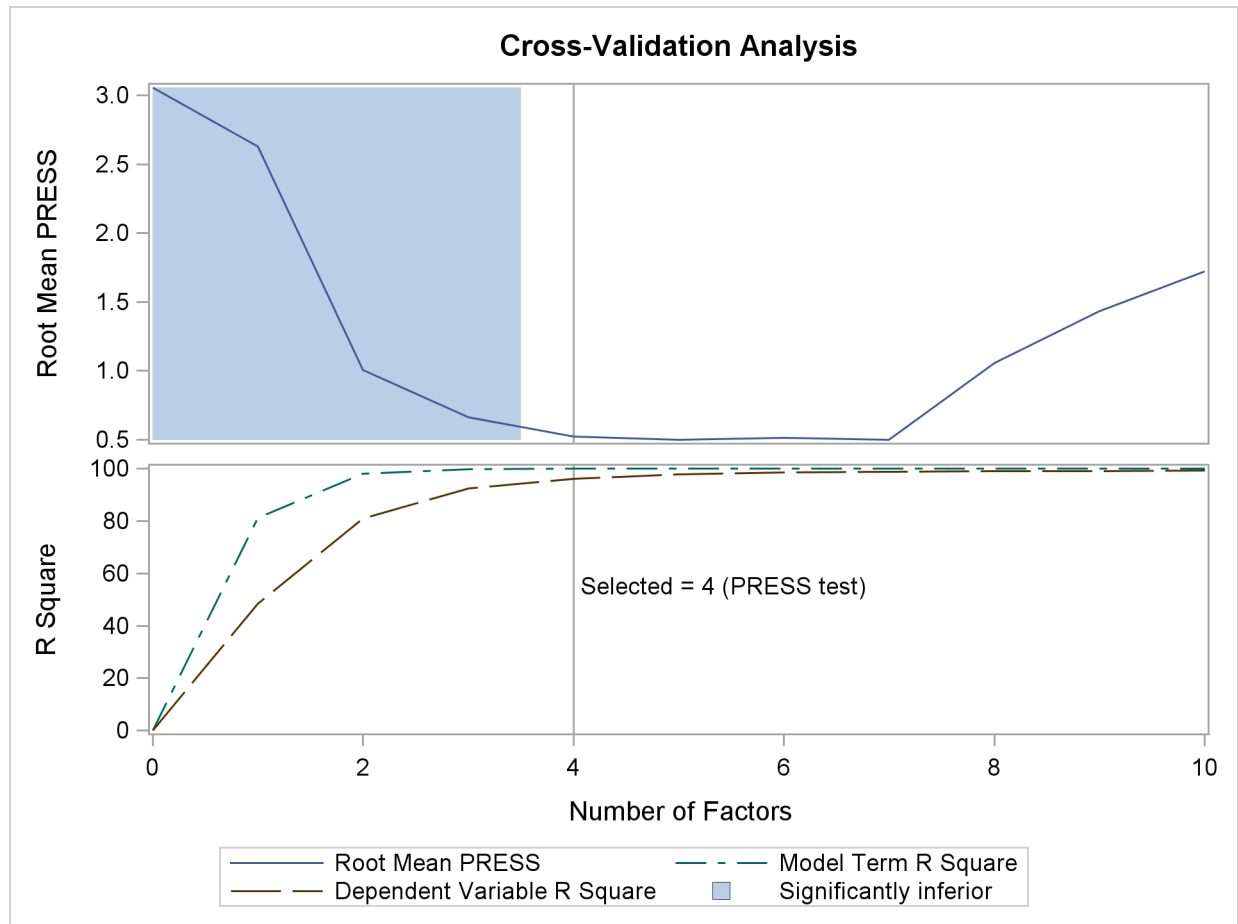
The factor loadings show how the PLS factors are constructed from the centered and scaled predictors. For spectral calibration, it is useful to plot the loadings against the frequency. In many cases, the physical meanings that can be attached to factor loadings help to validate the scientific interpretation of the PLS model. You can use ODS Graphics with PROC PLS to plot the loadings for the four PLS factors against frequency, as shown in the following statements.

```
ods graphics on;

proc pls data=ftrain nfac=4 plot=XLoadingProfiles;
   model tot_log tyr_log try_log = f1-f30;
run;

ods graphics off;
```

The resulting plot is shown in Output 66.3.5.

Output 66.3.5 Predictor Loadings across Frequencies
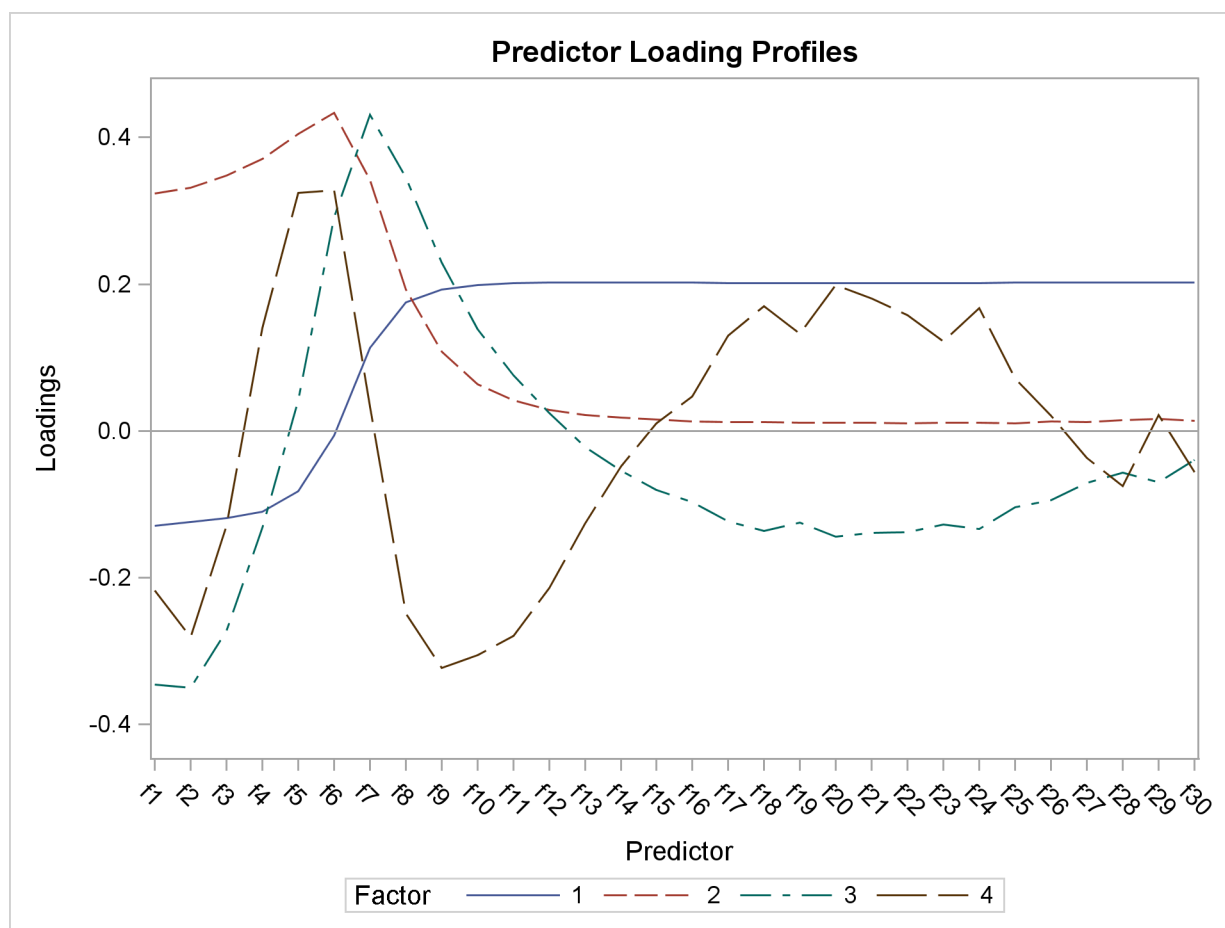
Notice that all four factors handle frequencies below and above about 7 or 8 differently. For example, the first factor is very nearly a simple contrast between the averages of the two sets of frequencies, and the second factor appears to be approximately a weighted sum of only the frequencies in the first set.

References

de Jong, S. (1993), "SIMPLS: An Alternative Approach to Partial Least Squares Regression," *Chemometrics and Intelligent Laboratory Systems*, 18, 251–263.

de Jong, S. and Kiers, H. (1992), "Principal Covariates Regression," *Chemometrics and Intelligent Laboratory Systems*, 14, 155–164.

Dijkstra, T. (1983), "Some Comments on Maximum Likelihood and Partial Least Squares Methods," *Journal of Econometrics*, 22, 67–90.

Dijkstra, T. (1985), *Latent Variables in Linear Stochastic Models: Reflections on Maximum Likelihood and Partial Least Squares Methods.*, Second Edition, Amsterdam, The Netherlands: Sociometric Research Foundation.

Frank, I. and Friedman, J. (1993), "A Statistical View of Some Chemometrics Regression Tools," *Technometrics*, 35, 109–135.

Geladi, P. and Kowalski, B. (1986), "Partial Least-Squares Regression: A Tutorial," *Analytica Chimica Acta*, 185, 1–17.

Haykin, S. (1994), *Neural Networks: A Comprehensive Foundation*, New York: Macmillan.

Helland, I. (1988), "On the Structure of Partial Least Squares Regression," *Communications in Statistics, Simulation and Computation*, 17, 581–607.

Hoerl, A. and Kennard, R. (1970), "Ridge Regression: Biased Estimation for Non-orthogonal Problems," *Technometrics*, 12, 55–67.

Lindberg, W., Persson, J.-A., and Wold, S. (1983), "Partial Least-Squares Method for Spectrofluorimetric Analysis of Mixtures of Humic Acid and Ligninsulfonate," *Analytical Chemistry*, 55, 643–648.

McAvoy, T. J., Wang, N. S., Naidu, S., Bhat, N., Gunter, J., and Simmons, M. (1989), "Interpreting Biosensor Data via Backpropagation," *International Joint Conference on Neural Networks*, 1, 227–233.

Naes, T. and Martens, H. (1985), "Comparison of Prediction Methods for Multicollinear Data," *Communications in Statistics, Simulation and Computation*, 14, 545–576.

Ränner, S., Lindgren, F., Geladi, P., and Wold, S. (1994), "A PLS Kernel Algorithm for Data Sets with Many Variables and Fewer Objects," *Journal of Chemometrics*, 8, 111–125.

Sarle, W. S. (1994), "Neural Networks and Statistical Models," in *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Shao, J. (1993), "Linear Model Selection by Cross-Validation," *Journal of the American Statistical Association*.

Tobias, R. (1995), "An Introduction to Partial Least Squares Regression," in *Proceedings of the Twentieth Annual SAS Users Group International Conference*, Cary, NC: SAS Institute Inc.

Ufkes, J. G. R., Visser, B. J., Heuver, G., and Van Der Meer, C. (1978), "Structure-Activity Relationships of Bradykinin-Potentiating Peptides," *European Journal of Pharmacology*, 50, 119.

Ufkes, J. G. R., Visser, B. J., Heuver, G., Wynne, H. J., and Van Der Meer, C. (1982), "Further Studies on the Structure-Activity Relationships of Bradykinin-Potentiating Peptides," *European Journal of Pharmacology*, 79, 155.

Umetrics (1995), *Multivariate Analysis (3-day course)*, Winchester, MA.

van den Wollenberg, A. L. (1977), "Redundancy Analysis—An Alternative to Canonical Correlation Analysis," *Psychometrika*, 42, 207–219.

van der Voet, H. (1994), "Comparing the Predictive Accuracy of Models Using a Simple Random-
ization Test," *Chemometrics and Intelligent Laboratory Systems*, 25, 313–323.

Wold, H. (1966), "Estimation of Principal Components and Related Models by Iterative Least
Squares," in P. R. Krishnaiah, ed., *Multivariate Analysis*, New York: Academic Press.

Wold, S. (1994), "PLS for Multivariate Linear Modeling," *QSAR: Chemometric Methods in Molec-
ular Design. Methods and Principles in Medicinal Chemistry*.

Subject Index

Syntax Index

Your Turn

We welcome your feedback.

- If you have comments about this book, please send them to **yourturn@sas.com**. Include the full title and page numbers (if applicable).

- If you have comments about the software, please send them to **suggest@sas.com**.

SAS® Publishing Delivers!

Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

SAS® Press

Need to learn the basics? Struggling with a programming problem? You'll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

support.sas.com/saspress

SAS® Documentation

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – **free** on the Web.
- Hard-copy books.

support.sas.com/publishing

SAS® Publishing News

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

support.sas.com/spn

§sas | THE POWER TO KNOW®