Discovering Partial Least Squares with JMP®

Ian Cox and Marie Gaudard

Contents

5 Predicting Biological Activity

Background

The example in this chapter comes from the field of drug discovery. New drugs are developed from chemicals that are biologically active. Because testing a compound for biological activity is expensive, chemists attempt to predict biological activity from other cheaper chemical measurements. In fact, computational chemistry makes it possible to calculate likely values for certain chemical properties without even making the compound.

In this example, you study the relationship between the size, hydrophobicity, and polarity of key chemical groups at various sites on the molecule, and the activity of the compound. The latter is represented by the logarithm of the relative Bradykinin potentiating activity. We develop a model based on a set of data from one study and then we apply the model to a separate data set from another study. For the first study, you learn that PLS is a useful tool for finding a few underlying factors that account for most of the variation in the response. However, you will also see that the model developed based on the first study's data set does not extend well to the data set from the second study.

The Data

Data Table Description

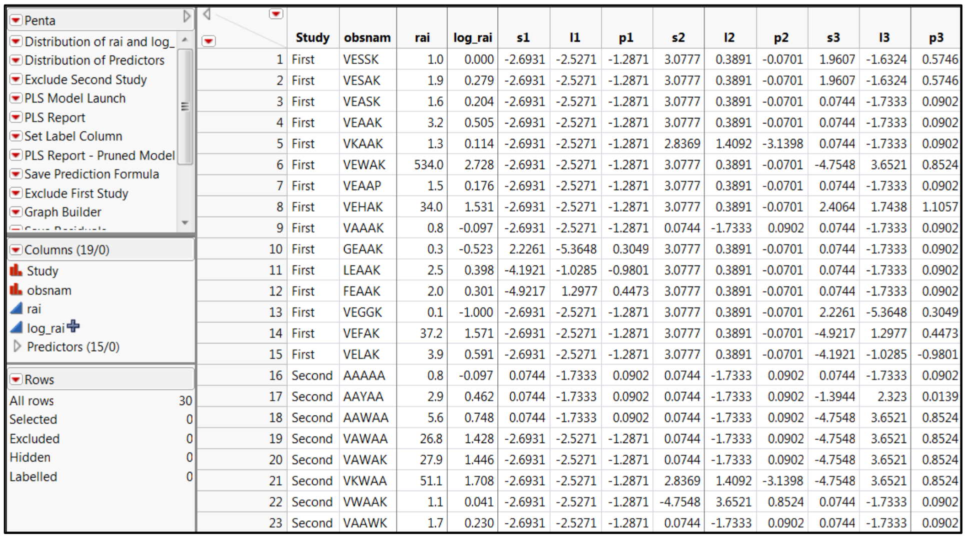
Open the data table Penta.jmp, partially shown in Figure 5.1, by clicking on the link in the master journal. This table contains 30 rows of observations.

The column obsnam contains an identification code. Each record in Penta.jmp represents a peptide chain of five amino acids. Each amino acid name is coded using a single letter and each chain is represented by five letters, as shown in the column obsnam. The amino acid coding is described in Table 1 of Hellberg et al. (1986).

The response of interest is rai, a relative measure of Bradykinin potentiating activity. (See Table 1 in both Ufkes et al. 1978 and Ufkes et al. 1982). However, rai is highly skewed, and so log_rai, the base 10 logarithm of rai, is used as the response of interest in the analysis. Note that log_rai is given by a formula; click the + sign next to log_rai in the **Columns** panel to view the formula.

The first column in the data table, Study, indicates the study of origin for the given row. The first 15 observations in the table were studied in Ufkes et al. (1978) and the last 15 in Ufkes et al. (1982).

Figure 5.1: Partial View of Penta.jmp



All rows	30
Selected	0
Excluded	0
Hidden	0
Labelled	0

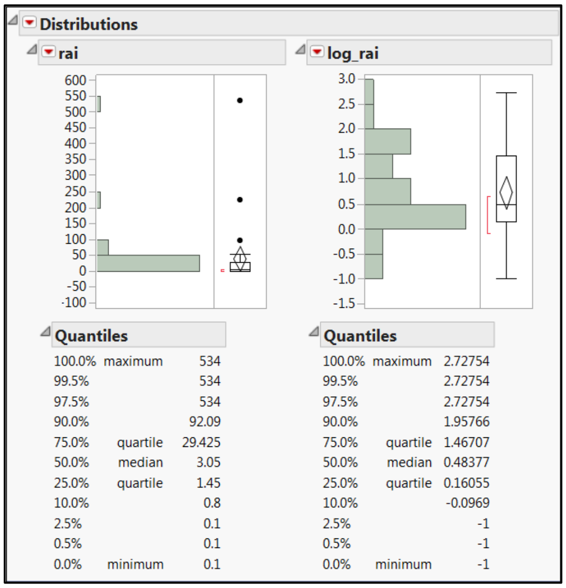
	Study	obsnam	rai	log_rai	s1	l1	p1	s2	l2	p2	s3	l3	p3
1	First	VESSK	1.0	0.000	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	1.9607	-1.6324	0.5746
2	First	VESAK	1.9	0.279	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	1.9607	-1.6324	0.5746
3	First	VEASK	1.6	0.204	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902
4	First	VEAAK	3.2	0.505	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902
5	First	VKAAK	1.3	0.114	-2.6931	-2.5271	-1.2871	2.8369	1.4092	-3.1398	0.0744	-1.7333	0.0902
6	First	VEWAK	534.0	2.728	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.7548	3.6521	0.8524
7	First	VEAAP	1.5	0.176	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902
8	First	VEHAK	34.0	1.531	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	2.4064	1.7438	1.1057
9	First	VAAAK	0.8	-0.097	-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902
10	First	GEAAK	0.3	-0.523	2.2261	-5.3648	0.3049	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902
11	First	LEAAK	2.5	0.398	-4.1921	-1.0285	-0.9801	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902
12	First	FEAAK	2.0	0.301	-4.9217	1.2977	0.4473	3.0777	0.3891	-0.0701	0.0744	-1.7333	0.0902
13	First	VEGGK	0.1	-1.000	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	2.2261	-5.3648	0.3049
14	First	VEFAK	37.2	1.571	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.9217	1.2977	0.4473
15	First	VELAK	3.9	0.591	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701	-4.1921	-1.0285	-0.9801
16	Second	AAAAA	0.8	-0.097	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902
17	Second	AAYAA	2.9	0.462	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	-1.3944	2.323	0.0139
18	Second	AAWAA	5.6	0.748	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902	-4.7548	3.6521	0.8524
19	Second	VAWAA	26.8	1.428	-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902	-4.7548	3.6521	0.8524
20	Second	VAWAK	27.9	1.446	-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902	-4.7548	3.6521	0.8524
21	Second	VKWAA	51.1	1.708	-2.6931	-2.5271	-1.2871	2.8369	1.4092	-3.1398	-4.7548	3.6521	0.8524
22	Second	VWAAK	1.1	0.041	-2.6931	-2.5271	-1.2871	-4.7548	3.6521	0.8524	0.0744	-1.7333	0.0902
23	Second	VAAWK	1.7	0.230	-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902

The data used in this example and a discussion can be found in SAS documentation (*SAS/STAT 9.3 User's Guide*, "The PLS Procedure"). To facilitate comparisons with SAS output, our analysis broadly follows the steps used in the PROC PLS example (Example 69.1). Further background on the data can be found in Ufkes et al. (1978) and Ufkes et al. (1982), Sjostrom and Wold (1985), and Hellberg et al. (1986).

Initial Data Visualization

Let's start by visualizing the data. Run the first saved script, Distribution of rai and log_rai. The plot for rai in Figure 5.2 shows that rai is highly skewed, with some large outlying values.

Figure 5.2: Distribution Reports for rai and log_rai



Quantiles

100.0%	maximum	534
99.5%		534
97.5%		534
90.0%		92.09
75.0%	quartile	29.425
50.0%	median	3.05
25.0%	quartile	1.45
10.0%		0.8
2.5%		0.1
0.5%		0.1
0.0%	minimum	0.1

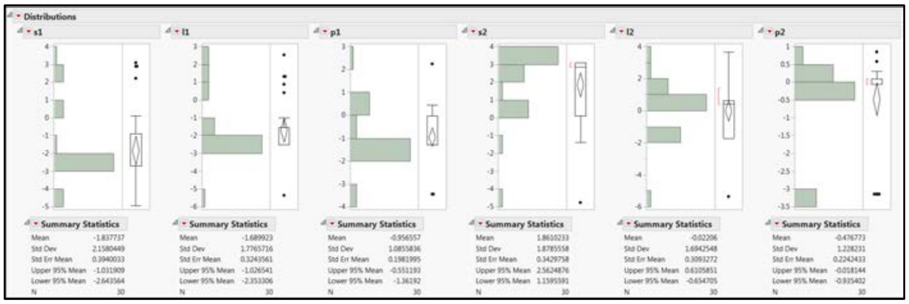
Quantiles

100.0%	maximum	2.72754
99.5%		2.72754
97.5%		2.72754
90.0%		1.95766
75.0%	quartile	1.46707
50.0%	median	0.48377
25.0%	quartile	0.16055
10.0%		-0.0969
2.5%		-1
0.5%		-1
0.0%	minimum	-1

Although PLS does not rely on distributional assumptions of normality, it is still good practice to look at the univariate distributions of the variables and assess whether one of the familiar transformations can make the distribution of a variable assume more of a symmetric, humped shape. In some cases, points that appear to be outliers appear much less extreme on the transformed scale. We see that this is the case here. When rai, the actual response of interest is transformed using a logarithmic function, the distribution of the transformed variable, log_rai, is much more symmetric and well-behaved.

Next, let's visualize the predictors one at a time. You can select **Analyze > Distribution**, or you can run the script called Distribution of Predictors that has been saved to the data table. The plots enable you to get a feel for the data. (See Figure 5.3.) Although some of the distributions appear unruly, given that these are measurements on predictors from a structured study, there is nothing to cause concern.

Figure 5.3: Partial View of Distribution Reports for Penta.jmp Predictors



s1

Summary Statistics	
Mean	-1.837737
Std Dev	2.1580449
Std Err Mean	0.3940033
Upper 95% Mean	-1.031909
Lower 95% Mean	-2.643564
N	30

l1

Summary Statistics	
Mean	-1.689923
Std Dev	1.7765716
Std Err Mean	0.3243561
Upper 95% Mean	-1.026541
Lower 95% Mean	-2.353306
N	30

p1

Summary Statistics	
Mean	-0.956557
Std Dev	1.0855836
Std Err Mean	0.1981995
Upper 95% Mean	-0.551193
Lower 95% Mean	-1.36192
N	30

s2

Summary Statistics	
Mean	1.8610233
Std Dev	1.8785558
Std Err Mean	0.3429758
Upper 95% Mean	2.5624876
Lower 95% Mean	1.1595591
N	30

l2

Summary Statistics	
Mean	-0.02206
Std Dev	1.6942548
Std Err Mean	0.3093272
Upper 95% Mean	0.6105851
Lower 95% Mean	-0.654705
N	30

p2

Summary Statistics	
Mean	-0.476773
Std Dev	1.228231
Std Err Mean	0.2242433
Upper 95% Mean	-0.018144
Lower 95% Mean	-0.935402
N	30

A First PLS Model

Our Plan

We begin by developing a PLS model for the data from the first study, namely the first 15 rows. Recall that these are the amino acid chains that were studied in the Ufkes et al. (1978) paper. We then apply the model we develop to the data from the second study.

To restrict our analysis to only the first 15 rows, we exclude and hide rows 16 to 30. To do this, you can select rows 16 to 30, and then right-click in the highlighted area next to the row numbers and select **Hide and Exclude**. Or, you can run the script Exclude Second Study.

Performing the Analysis

In JMP, PLS is accessed by selecting **Analyze > Multivariate Methods > Partial Least Squares**. It can also be accessed this way in JMP Pro.

1. Enter log_rai as **Y, Response**.
2. Enter all Predictors as **X, Factor**.
3. Click **OK**.

In JMP Pro, PLS can also be accessed through **Fit Model**. Select **Analyze > Fit Model**.

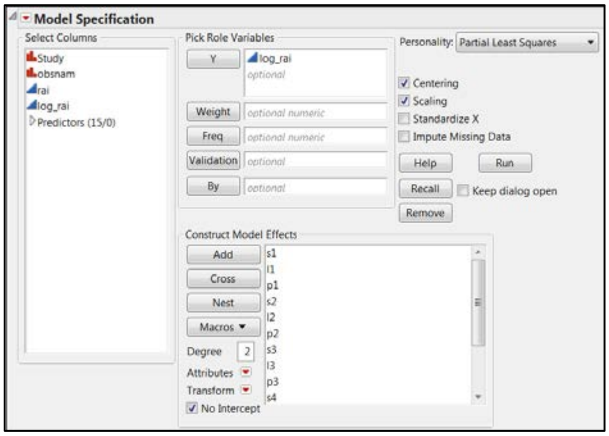
1. Enter log_rai as **Y**.
2. Enter all Predictors as **Model Effects**.
3. Select **Partial Least Squares** as the **Personality**.
4. Deselect the **Standardize X** option.

The **Standardize X** option centers and scales columns that are involved in higher-order terms. Leaving it checked in this case, where we have no higher-order terms, only affects reporting of the model coefficients for the original data. If you access PLS using **Analyze > Multivariate Methods > Partial Least Squares**, you are not able to add higher-order terms, and so the **Standardize X** option is not available.

Your Fit Model window should appear as shown in Figure 5.4.

5. Click **Run**.

Figure 5.4: Fit Model Window



Model Specification

Select Columns: Study, obsnam, rai, log_rai, Predictors (15/0)

Pick Role Variables

Y: log_rai

Personality: Partial Least Squares

☑ Centering
☑ Scaling
☐ Standardize X
☐ Impute Missing Data

☐ Keep dialog open

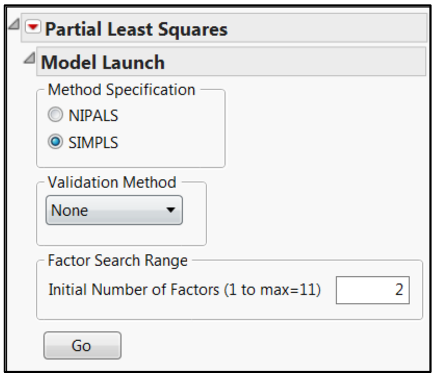
Construct Model Effects: s1, l1, p1, s2, l2, p2, s3, l3, p3, s4

Degree: 2

☑ No Intercept

Either of the menu approaches opens the Partial Least Squares Model Launch control panel. In the PLS Model Launch control panel, select **SIMPLS** as the **Method Specification**, choose **None** as the **Validation Method**, and request an **Initial Number of Factors** equal to **2**, as shown in Figure 5.5. You can also run the script PLS Model Launch to obtain the PLS Model Launch control panel.

Figure 5.5: Specification for PLS Model Launch Control Panel



Partial Least Squares

Model Launch

Method Specification

○ NIPALS

● SIMPLS

Validation Method: None

Factor Search Range

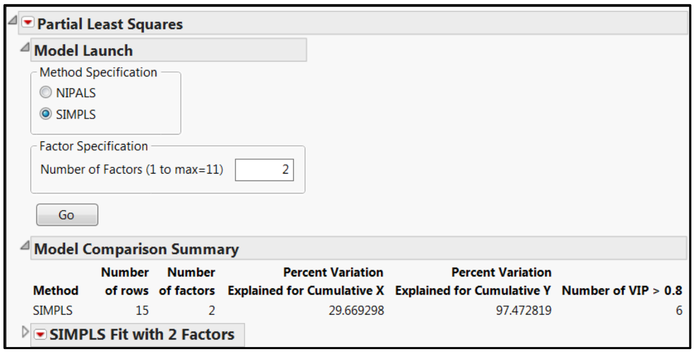
Initial Number of Factors (1 to max=11): 2

These choices imply that: you are using the SIMPLS algorithm to estimate the model parameters; you are not using model validation within the platform; and you are specifically fitting only two factors to encompass the variation between the model effects (**X**s) and **Y**. Note that we would obtain an identical model had we used NIPALS, because there is only one response. The VIP values, though, would differ slightly.

The Partial Least Squares Report

Clicking **Go** adds additional content to the report. (The script is PLS Report.) Two report sections are appended: A **Model Comparison Summary** report, which is updated when new fits are performed, and a **SIMPLS Fit with 2 Factors** report, showing extensive details about the fit just performed (shown closed in Figure 5.6).

Figure 5.6: PLS Report for Two-Factor Fit



Partial Least Squares

Model Launch

Method Specification

○ NIPALS

● SIMPLS

Factor Specification

Number of Factors (1 to max=11): 2

Model Comparison Summary

Method	Number of rows	Number of factors	Percent Variation Explained for Cumulative X	Percent Variation Explained for Cumulative Y	Number of VIP > 0.8
SIMPLS	15	2	29.669298	97.472819	6

SIMPLS Fit with 2 Factors

The **Model Comparison Summary** confirms that the model was fit as requested and that the two factors account for about 30% of the variation in the **X**s and 97% of the variation in **Y**. The last column, **Number of VIP > 0.8**, indicates that 6 of the 15 predictors are influential in determining the two factors. This suggests that there might be an opportunity for refining the model by dropping some of the predictors.

Just in passing, we note that the content of JMP reports can be customized using **Preferences**. To customize your PLS options, select **File > Preferences > Platforms > Partial Least Squares Fit** for report options, and **File > Preferences > Platforms > Partial Least Squares** for model launch options. Within the PLS report itself, options can be found in menus obtained by clicking red triangles.

The SIMPLS Fit Report

Now let's look at the report for the fit, **SIMPLS Fit with 2 Factors** (Figure 5.7). The **X-Y Scores Plots** show what we would expect, namely, a correlation between the X and Y scores. In a good PLS model, the first few factors should show a high correlation between the X and Y scores. The correlation usually (but not always) decreases from one factor to the next. The **X-Y Scores Plots** exhibit this behavior nicely.

It would be useful if you could see each observation's amino acid coding as a tooltip when you mouse over a point in the **X-Y Scores Plots**, as illustrated in the plot for the first scores in Figure 5.7. To accomplish this, assign the column obsnam the **Label** attribute. To do this, right-click obsnam in the **Columns** panel and select **Label/Unlabel**. (Alternatively, run the script Set Label Column.)

Incidentally, you can have labels displayed on plots for all or only certain rows by applying the **Label** attribute to the rows of interest. To do this, select the rows in the data table, right-click in the highlighted area, and then select **Label/Unlabel**.

The **Percent Variation Explained** report displays the variation in the **X**s and in **Y** that is explained by each factor. The **Cumulative X** and **Cumulative Y** values must agree with the corresponding figures in the **Model Comparison Summary**.

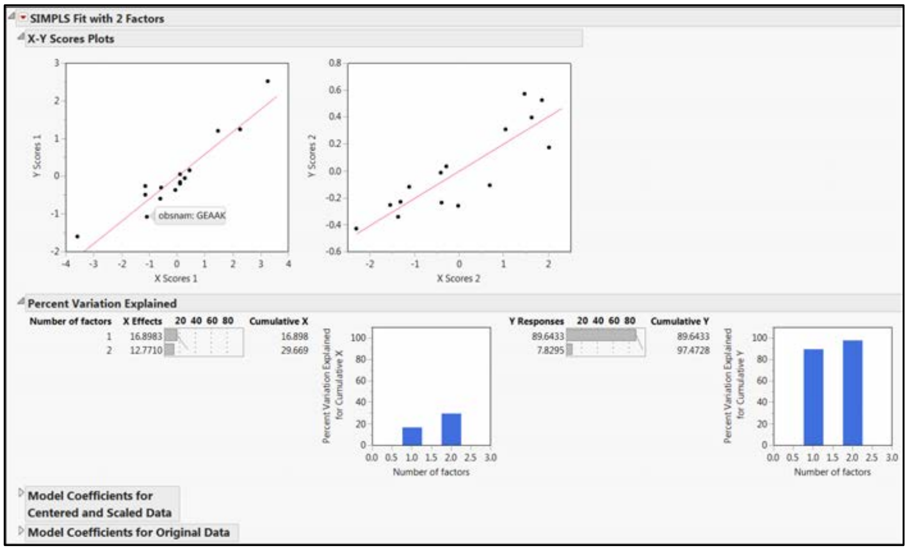
The two **Model Coefficients** reports give the estimated model coefficients for predicting log_rai.

- The report **Model Coefficients for Centered and Scaled Data** gives the coefficients that apply when the **X**s and **Y** have been centered to have mean zero and standard deviation one.

- The report **Model Coefficients for Original Data** gives coefficients for the model expressed in terms of the raw data values. (Had you checked the **Standardize X** option on the Fit Model launch window, the coefficients in this report would be for a model given in terms of the raw **Y** values but standardized **X** values.)

This second set of model coefficient estimates is often of secondary interest in terms of the analysis.

Figure 5.7: The SIMPLS Fit with 2 Factors Report



SIMPLS Fit with 2 Factors

X-Y Scores Plots

Percent Variation Explained

Number of factors	X Effects	Cumulative X
1	16.8983	16.898
2	12.7710	29.669

Y Responses	Cumulative Y
89.6433	89.6433
7.8295	97.4728

Model Coefficients for Centered and Scaled Data

Model Coefficients for Original Data

Y Scores 1
X Scores 1
obsnam: GEAAK
Y Scores 2
X Scores 2
20 40 60 80
Percent Variation Explained for Cumulative X
Percent Variation Explained for Cumulative Y
Number of factors

Other Options

A number of useful diagnostic and other tools are available as options from the **SIMPLS Fit with 2 Factors** red triangle menu. We explore a few of these in this section.

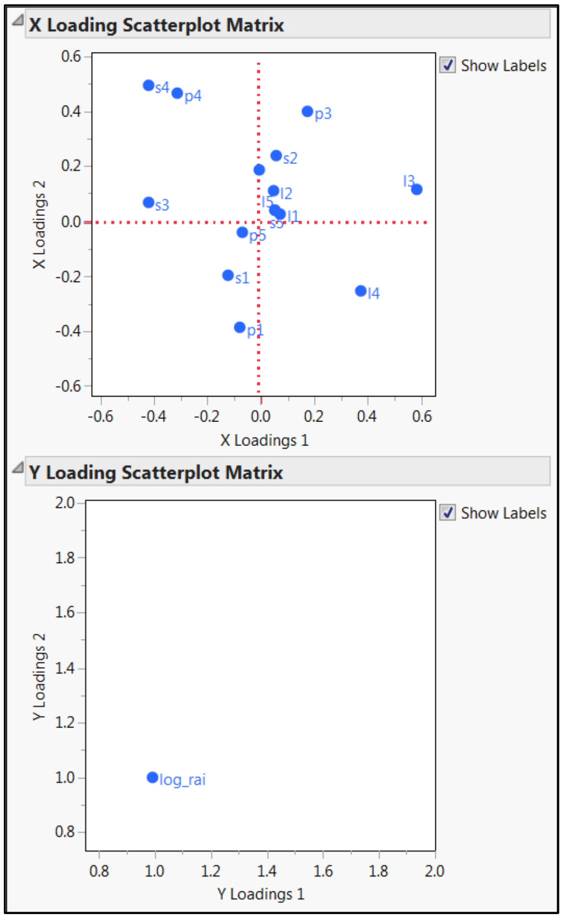
Loading Scatterplot Matrices

You can plot the loadings against each other. Recall that, for a given factor, the X loadings reflect the strength of the correlation relationship between the **X** variables and that factor. Similarly, the Y loadings reflect the correlation relationship between the **Y** variables and that factor. Because we have only one **Y** variable, there is a single **Y** loading for each factor. But because we have 15 **X**s, we have 15 loadings for each of the two factors.

Also, keep in mind that loadings are scaled so that, for a given factor, the vector of loadings has length one. This normalization enables us to compare loadings for **X**s and **Y**s across factors.

Select **Loading Scatterplot Matrices** from the red triangle menu for the fit. Figure 5.8 shows the resulting plots.

Figure 5.8: X and Y Loading Scatterplot Matrices



Recall that each column in the projection matrix **P** describes the strength of the correlation between a factor and the predictors. (See Equation (4.1).) In a general situation, the **X Loading Scatterplot Matrix** gives views of these loadings for two factors at a time.

In the **X Loading Scatterplot Matrix** in Figure 5.8, we see that l3 has a high positive loading on the first factor, but a small loading on the second factor. So it is highly correlated with the first factor, but not the second. On the other hand, p1 has a relatively large negative loading on the second factor, but is hardly correlated with the first factor. In fact, the first factor seems to be characterized primarily by the amino acids in the 3rd and 4th positions, while the second factor seems to be characterized by those in the 1st, 3rd, and 4th positions. However, some of these predictors, such as those for the 4th position, have positive correlations with one factor and negative correlations with the other.

This plot also shows a cluster of **X** variables with loadings on both factors that are close to zero. These variables, representing primarily the 2nd and 5th positions, are not explaining much of the variation in the **X** variables.

The **Y Loading Scatterplot Matrix** plots values that represent the scaled correlation between the **Y** variables and each of the two factors. In this case the plot is not particularly informative because there is only one **Y** variable. Because both correlations are scaled to have length one, its loading on both factors is 1.

Loading Plots

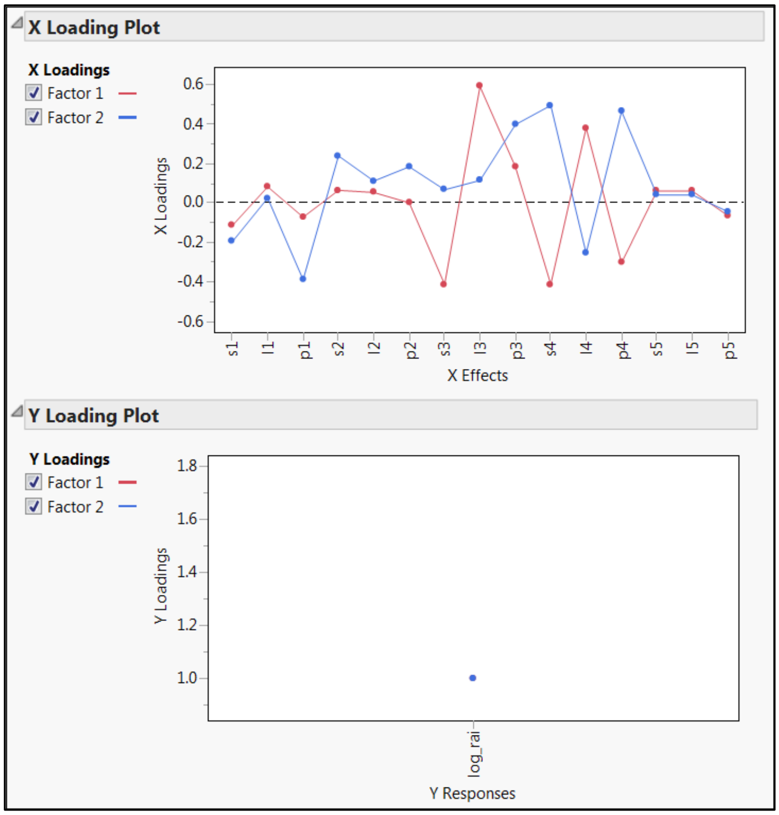
Loading plots give another way to view the relationships between the **X**s and **Y**s and the PLS factors. Loading plots are overlay plots that enable you to choose the factors whose loadings you want to display.

Select **Loading Plots** from the red triangle menu for the **SIMPLS Fit with 2 Factors** report. Figure 5.9 shows the two resulting plots. We have added a reference line at 0 to the **X Loading Plot** to help with interpretation. (To do this, double-click on the vertical axis, click **Add** in the **Reference Lines** panel, and then click **OK**.)

The **X Loading Plot** shows that not all predictors impact the factors highly, echoing our conclusions based on the **X Loading Scatterplot Matrix**. In fact, the loadings on both factors for l1, s5, l5, and p5 are close to zero. The predictors s3, l3, s4 and l4, for example, have high loadings, in absolute value, on **Factor 1**, whereas s4, l4, and p4 have high loadings, in absolute value, on both **Factor 1** and **Factor 2**, but in different directions. We think of the two factors as capturing this distinction among these predictors.

Again, because there is only one response variable, the **Y Loading Plot** is uninteresting.

Figure 5.9: Loading Plots



Score Scatterplot Matrices

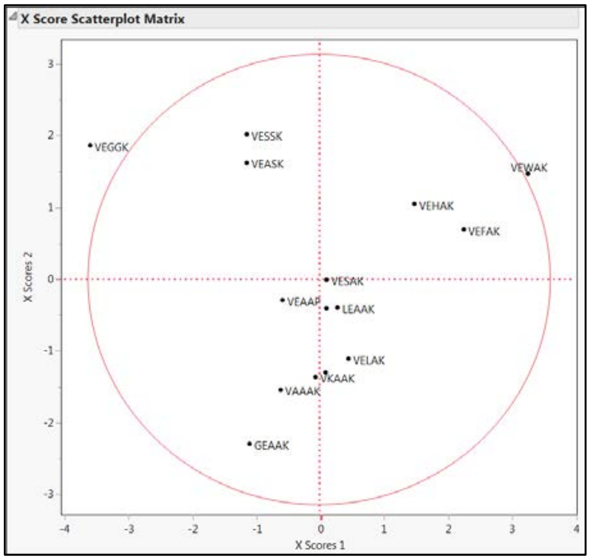
The X and Y scores express the data in terms of the factors. In this model, there are two factors, so there are two sets of X scores and two sets of Y scores. To look for irregularities relative to the projections of the data to the X and Y spaces, it is useful to plot the X scores and the Y scores against each other. 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 to the model. Two or more groupings of observations indicate that it might be better to analyze the groups separately.

To plot scores in this fashion, select **Score Scatterplot Matrices** from the red triangle menu for the **SIMPLS Fit with 2 Factors.** The scatterplot for the X scores is shown in Figure 5.10. The scatterplot matrix has a single cell, because only two factors were extracted. A 95% confidence ellipsoid, calculated using the orthogonality of the X scores, is shown on the plot.

To identify the observations in the scatterplot matrix, select rows 1–15 in the data table, right-click in the highlighted area, and select **Label/Unlabel**. This option tells JMP to label all the selected rows in appropriate plots (Figure 5.10).

Two peptide chains, "VEGGK" and "VEWAK" fall slightly outside the confidence ellipse, although "VEWAK" is essentially on the ellipse boundary. The chain "VEGGK" lies marginally beyond the plotted ellipse, indicating that it might be influential in the PLS analysis. You should check this observation to make sure that it is reliable. Note also that the plot shows some clustering of peptide chains with similar amino acid positional compositions.

Figure 5.10: X Score Scatterplot Matrix

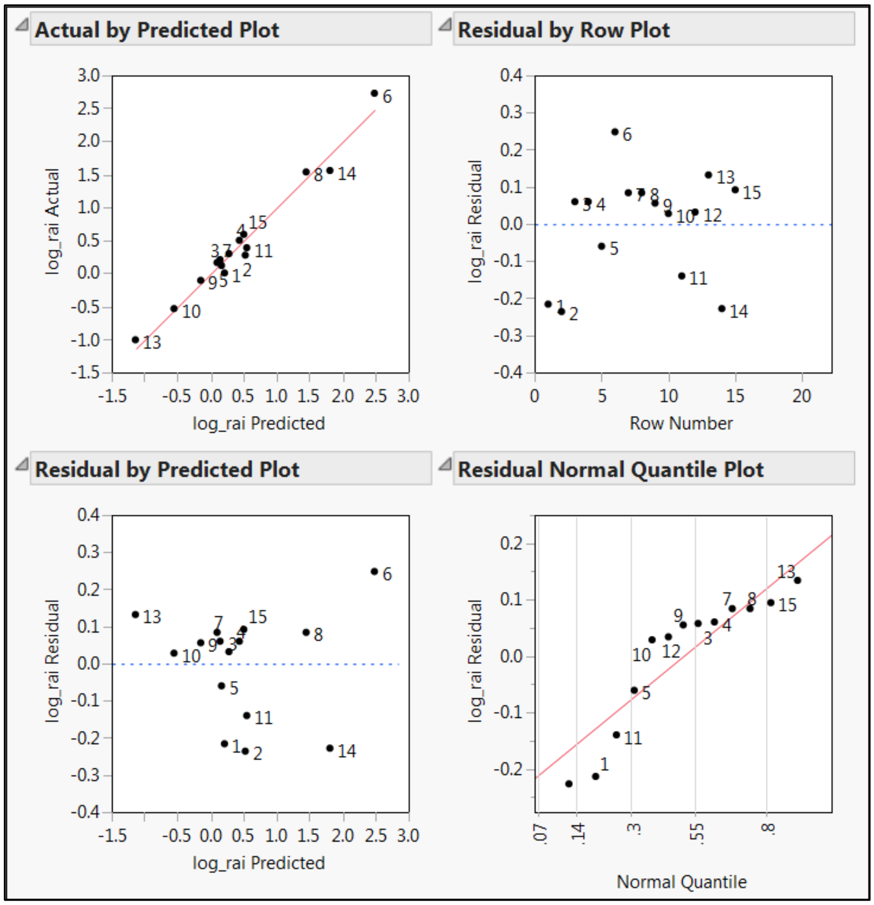


Diagnostics Plots

Selecting **Diagnostics Plots** from the fit's red triangle menu reveals four plots that help in detecting outliers or patterns that might be affecting the fit (Figure 5.11). These plots help you to detect non-normality, autocorrelation, and non-constant variance, all of which can signal problems for the fit.

At this point, we switch to using the row number as a label column. In the **Columns** panel, right-click on obsnam and select **Label/Unlabel** to remove the label attribute from this column. Do not make any changes to the **Label** attribute that is currently applied to rows 1–15. When rows are given the **Label** row state and no column is selected as a **Label** column, the default is to label by row number.

Figure 5.11: Diagnostics Plots



The **Actual by Predicted Plot** shows good agreement between the actual values of log_rai and the values predicted by the two-factor PLS model.

An ideal residual plot looks like a rectangular point cloud with most of the points falling in the vertical middle third of the plot. For these data, neither the **Residual by Predicted Plot** nor the **Residual by Row Plot** shows anything unusual.

In an ideal normal quantile plot, the points fall on a straight line. Here, the **Residual Normal Quantile Plot** shows that several observations are more extreme at the lower end than would be expected. However, this deviation from normality is not serious enough to cause concern.

Variable Reduction in PLS

Recall that PLS models **X** and **Y** by using extracted factors, and then relates **Y** to **X** by fitting a regression model **Y** = **XB** that is derived using those factors. The estimate of the matrix **B** involves the Y loadings as well as the extracted factors. (See Equation (4.2) and Equation (4.3).) It follows that a predictor can be important in the model in at least one of two ways: It can be important in connection with characterizing the factors used to model **X**; or it can be important in terms of the regression model that relates **Y** to **X**. A predictor could be useful in explaining variation in the **X** variables as well as their correlation to **Y**, thus helping to characterize the factors, and yet not be directly useful in predicting **Y**.

The *Variable Importance for the Projection* (VIP) statistic, discussed in Wold (1995, p. 213) and in Wold et al. (2001, p. 123), is defined as a weighted sum of squares of the weights, $\mathbf{W}$. (See Appendix 1 and Pérez-Enciso and Tenenhaus 2003.) Being based on the weights, it measures a predictor's contribution to characterizing the factors used in the PLS model, or, equivalently, to defining the projection.

Wold (1995) indicates that predictors that have both small VIP values and regression coefficients near zero can be considered for deletion from the model. Cut-off values for the VIP vary throughout the literature, but there is some agreement that values greater than 1.0 indicate predictors that are important, whereas values below 0.8 indicate predictors that can be deleted, assuming that their regression coefficients are small. In fact, Wold (see Eriksson et al. 2006) suggests a VIP cut-off of 0.8.

Let's summarize our discussion:

- A predictor's VIP represents its importance in determining the PLS projection model for both predictors and responses.
- A predictor's regression coefficient represents that variable's importance in predicting the response.
- If a predictor has a small VIP value and a relatively small regression coefficient (in absolute value), then it is a prime candidate for removal.

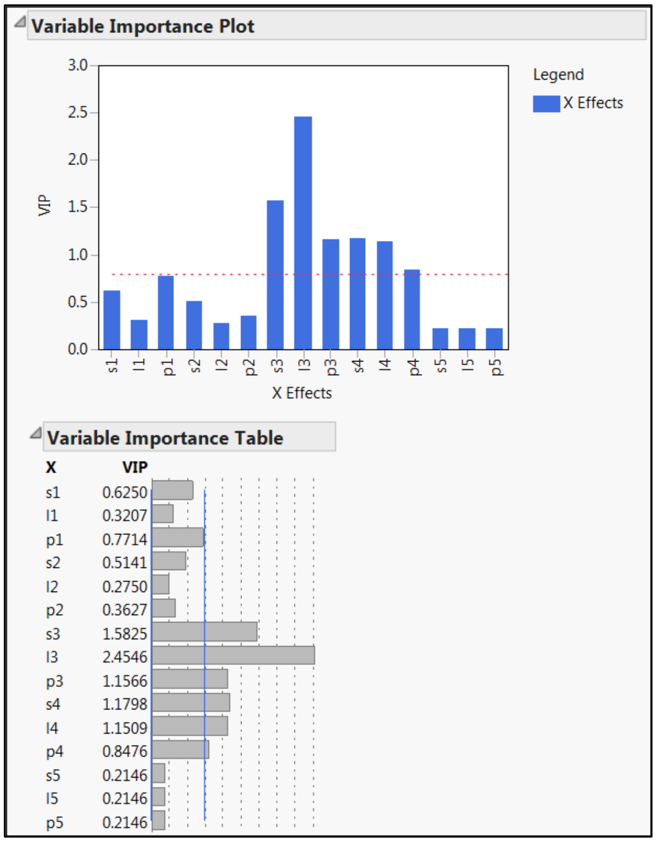
There has been considerable study in recent years of variable reduction procedures as they relate to PLS. One should engage in variable deletion cautiously. (See Appendix 2.) Use of cross validation to validate pruned models is prudent.

Variable Importance Plots

We illustrate how variable reduction can be accomplished in our example. We want to look at the VIP for each predictor and at the regression coefficients that make up the **B** matrix (which, in this case, is a column vector because there is only one **Y** and no constant term).

From the red triangle menu for the **SIMPLS Fit with 2 Factors** report, select **Variable Importance Plot**. This option provides both a **Variable Importance Plot** and a **Variable Importance Table** (Figure 5.12). The plot shows the VIP values across the predictors, with a dashed horizontal reference line at 0.8. The **Variable Importance Table** gives a similar plot, but also provides the actual VIP values. You can place these values into a data table by right-clicking in that report, and selecting **Make into Data Table**.

Figure 5.12: Variable Importance Plot and Table



Variable Importance Table

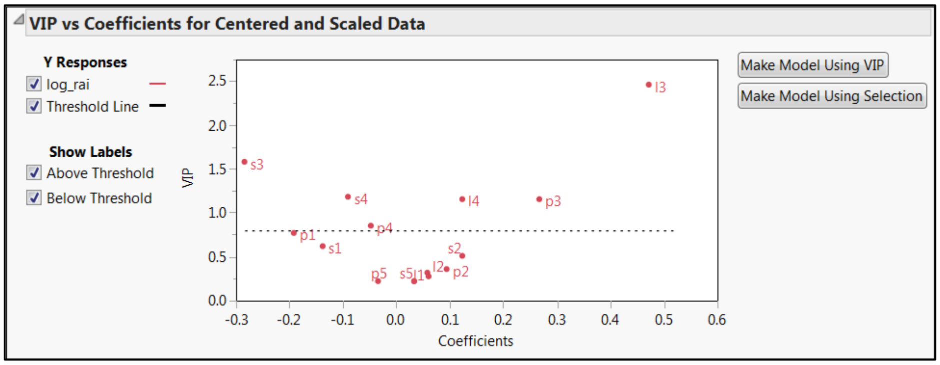
X	VIP
s1	0.6250
l1	0.3207
p1	0.7714
s2	0.5141
l2	0.2750
p2	0.3627
s3	1.5825
l3	2.4546
p3	1.1566
s4	1.1798
l4	1.1509
p4	0.8476
s5	0.2146
l5	0.2146
p5	0.2146

The **Variable Importance Plot** shows six predictors with VIP values exceeding 0.8. One can conclude that these are important for the modeling of both **X** and **Y**. Note that these are measures corresponding to the amino acids in positions 3 and 4 in the peptide chain, which suggests that the amino acids in these positions are important in modeling Bradykinin potentiating energy. The impact of these positions is detailed in Ufkes et al. (1978) and the significance of position 3 is acknowledged in Ufkes et al. (1982).

Now select the option **VIP vs Coefficients Plots** from the red triangle menu for the **SIMPLS Fit with 2 Factors** report. Two plots are shown, one for centered and scaled data and one for the original data. Figure 5.13 shows the plot for centered and scaled data.

The **VIP vs Coefficients** plots show the estimates of the regression coefficients for model terms on the horizontal axis and their VIP values on the vertical axis. Thus, these plots simultaneously give information about how each model term contributes to the regression and to the latent structure. Keep in mind that the values of the regression coefficients are affected by the centering and scaling of the measurements. So, unless there is a compelling reason to do otherwise, we recommend that the user focus on the plot in the **VIP vs Coefficients for Centered and Scaled Data** report, rather than the plot in the **VIP vs Coefficients for Original Data** report, in making decisions about variable reduction.

Figure 5.13: VIP versus Coefficients Plot for Centered and Scaled Data



Y Responses
☑ log_rai
☑ Threshold Line

Show Labels
☑ Above Threshold
☑ Below Threshold

To the right of the plot entitled **VIP vs Coefficients for Centered and Scaled Data**, you see two selection buttons: **Make Model Using VIP** and **Make Model Using Selection**. These buttons offer convenient ways to specify reduced models. If you have fit your model using the Partial Least Squares platform under the Multivariate Methods menu, clicking **Make Model Using VIP** opens a Partial Least Squares launch window where: the **X, Factor** list is populated with the effects whose VIPs exceed 0.8; the **Y, Response** list includes the previously selected responses. If you have fit your model using the Partial Least Squares personality under Fit Model, clicking **Make Model Using VIP** opens an appropriately populated Fit Model launch window.

Alternatively, you might prefer to select predictors directly in the plot by dragging a rectangle or by using the *Lasso* tool. Then, clicking on **Make Model Using Selection** opens a Partial Least Squares launch window where the **X, Factor** list is populated with these selected effects, or a Fit Model launch window where the **Construct Model Effects** list is populated with the selected effects.

We see that the predictors l1, s2, l2, p2, s5, l5, and p5 have small absolute coefficients and small VIPs. Looking back at the **Loading Scatterplot Matrices**, you see that these variables have loadings near zero for both PLS components, indicating that they don't have much influence on the factors that were used to construct the model.

A Pruned PLS Model

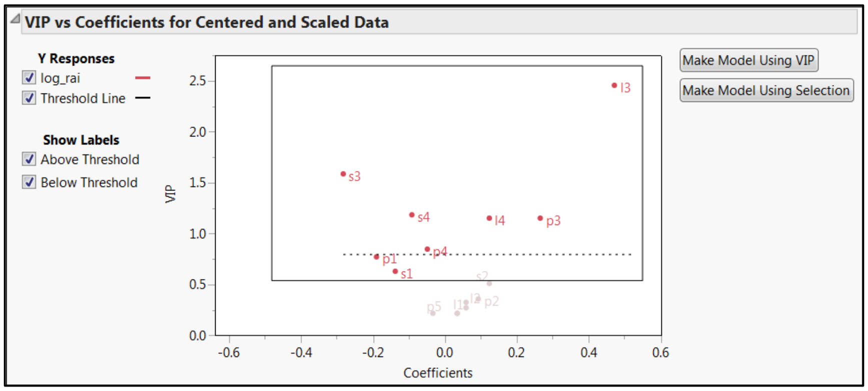
Model Fit

Based on our study of VIPs and regression coefficients, we remove the variables l1, s2, l2, p2, s5, l5, and p5 from the PLS model. Note that we do not remove s1 and p1, although their VIPs are below 0.8. These two variables have regression coefficients that are perhaps not negligible, and we prefer to err in the direction of not removing potentially active predictors. So we refit the model with the remaining eight predictors: s1, p1, s3, p3, l3, s4, l4, and p4.

1. In the plot entitled **VIP vs Coefficients for Centered and Scaled Data**, drag a rectangle starting at the upper right of the plot, above and to the right of l3, to include all eight of these variables (s1, p1, s3, p3, l3, s4, l4, and p4), as shown in Figure 5.14. Note that we have resized the plot to make it easier to select the desired variables by dragging a rectangle.
2. Click **Make Model Using Selection**. In the launch window that appears, make sure that you selected the correct variables. If you accidentally missed a variable or added an undesired variable, you can make an adjustment in this window.
3. Click **Run** or **OK**.

Alternatively, you can add the eight variables directly by selecting **Analyze > Multivariate Methods > Partial Least Squares** in both JMP and JMP Pro, and **Analyze > Fit Model** in JMP Pro.

Figure 5.14: Selection of Eight Variables for Pruned Model



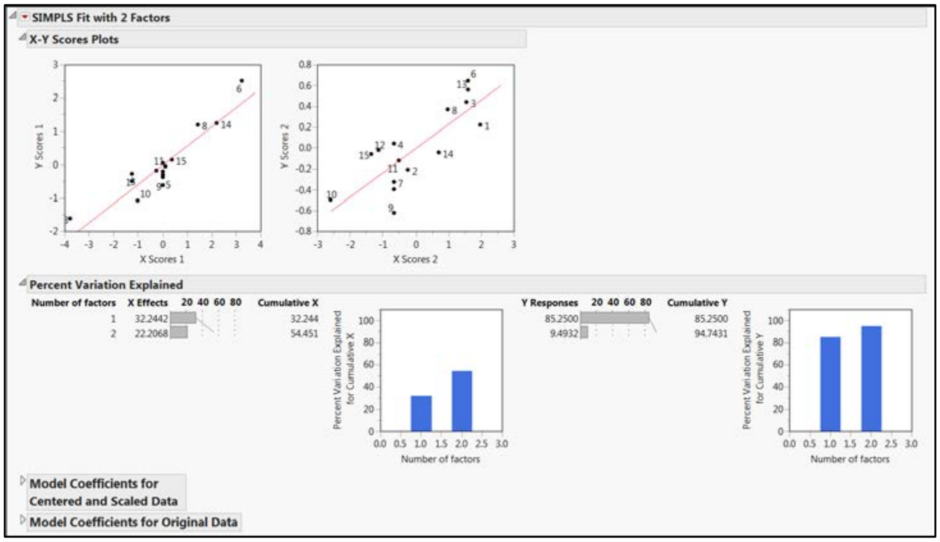
VIP vs Coefficients for Centered and Scaled Data

Y Responses
☑ log_rai
☑ Threshold Line

Show Labels
☑ Above Threshold
☑ Below Threshold

In the PLS Model Launch window, as before, select **SIMPLS** as the **Method Specification**, choose **None** as the **Validation Method**, and request an **Initial Number of Factors** equal to **2**. These choices produce the report shown in Figure 5.15. You can also simply run the script PLS Report – Pruned Model to obtain the report shown in Figure 5.15.

Figure 5.15: SIMPLS Fit with 2 Factors Report for Reduced Model



SIMPLS Fit with 2 Factors

X-Y Scores Plots

Y Scores 1
X Scores 1
Y Scores 2
X Scores 2
Percent Variation Explained for Cumulative X
Percent Variation Explained for Cumulative Y
Number of factors

Percent Variation Explained

Number of factors	X Effects	Cumulative X	Y Responses	Cumulative Y
1	32.2442	32.244	85.2500	85.2500
2	22.2068	54.451	9.4932	94.7431

Model Coefficients for Centered and Scaled Data

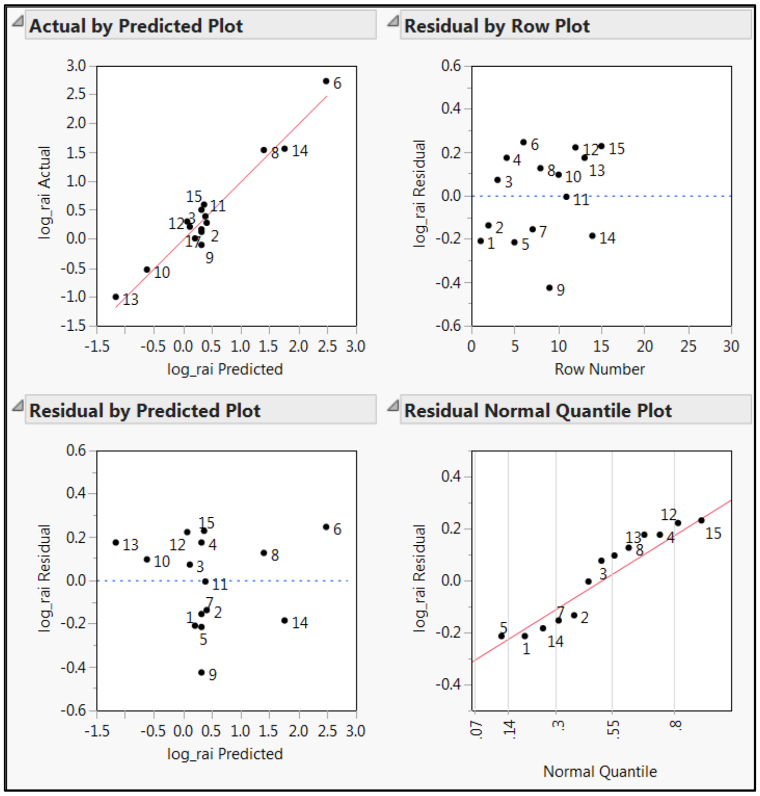
Model Coefficients for Original Data

The variation explained in **Y** by the pruned model is about 95%, a slight drop from 97% in the model with all predictors. However, the variation explained in **X** by the pruned model is about 54%, compared with about 30% in the full model. It appears that, by dropping predictors that are not highly related to **Y**, the new PLS factors provide a better representation of the variability in the reduced **X** space.

Diagnostics

From the red triangle menu next to the model fit, select **Diagnostics Plots**. The **Actual by Predicted Plot** shows that the model is predicting well. The **Residual by Predicted Plot** shows no anomalies or patterns. We note that the **Residual Normal Quantile Plot** is closer to being linear than it was for the full model (Figure 5.16).

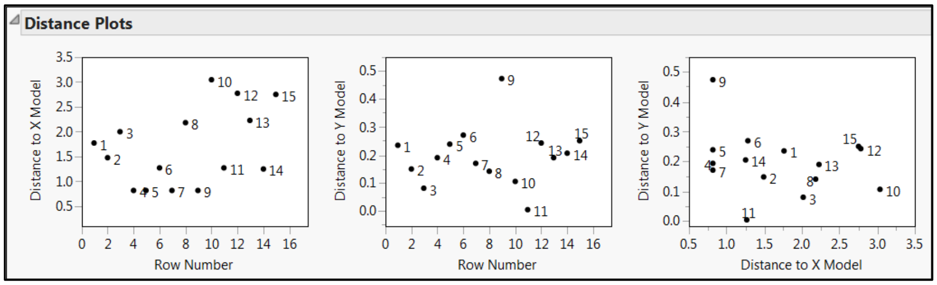
Figure 5.16: Diagnostics Plots for Pruned Model



Another way to check for outliers in the model is to look at the Euclidean distance from each observation to the PLS model in both the **X** and **Y** spaces. No observation should be dramatically farther from the model than the rest. Such behavior might indicate that the point is unduly influencing the fit of the model. If there is a group of points that are all farther from the model than the rest, it might be that they have something in common and should be analyzed separately.

Select **Distance Plots** from the red triangle menu to obtain the plots shown in Figure 5.17. With the possible exception of row 9 (which could be further investigated), there appear to be no outliers. Note that these distances to the model are called *DModX* and *DModY* by Umetrics and others (Eriksson et al. 2006).

Figure 5.17: Distance Plots for Pruned Model



Performance on Data from Second Study

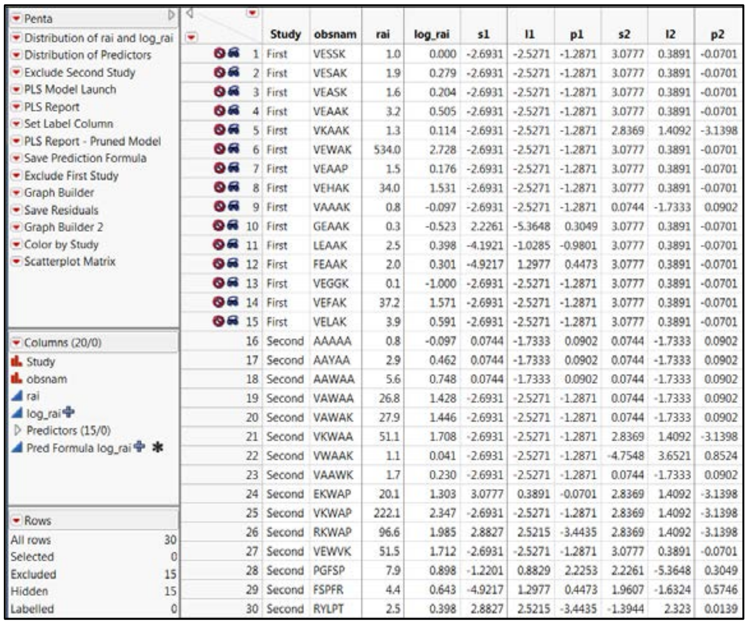
Comparing Predicted Values for the Second Study to Actual Values

The reduced model appears to be more satisfactory than the model including all predictors. So let's see how it performs on the observations that are currently hidden and excluded.

Select **Save Columns > Save Prediction Formula** from the red triangle menu for the **SIMPLS Fit with 2 Factors** report. (The script is Save Prediction Formula.) This option adds a new formula column called Pred Formula log_rai to the data table. Note that predicted values for all 30 observations appear in the data table, because a formula is being saved to the column.

We want to evaluate performance on rows 16–30, which contain the data from the later 1982 study. So, select rows 16–30 in the data table. Right-click on one of the highlighted rows and select **Clear Row States** from the menu that appears. Next, select rows 1–15 in the data table. Right-click in the highlighted area and select **Hide and Exclude**. (Alternatively, run the script Exclude First Study.) Your data table should appear as shown in Figure 5.18.

Figure 5.18: Row States That Define the Second Study



Penta
Distribution of rai and log_rai
Distribution of Predictors
Exclude Second Study
PLS Model Launch
PLS Report
Set Label Column
PLS Report - Pruned Model
Save Prediction Formula
Exclude First Study
Graph Builder
Save Residuals
Graph Builder 2
Color by Study
Scatterplot Matrix

Columns (20/0)
Study
obsnam
rai
log_rai
Predictors (15/0)
Pred Formula log_rai

Rows	
All rows	30
Selected	0
Excluded	15
Hidden	15
Labelled	0

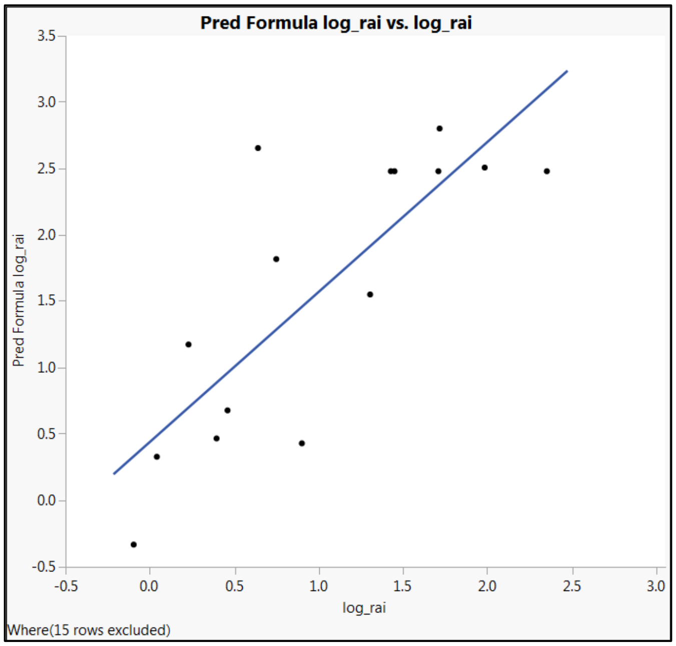
	Study	obsnam	rai	log_rai	s1	l1	p1	s2	l2	p2
1	First	VESSK	1.0	0.000	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
2	First	VESAK	1.9	0.279	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
3	First	VEASK	1.6	0.204	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
4	First	VEAAK	3.2	0.505	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
5	First	VKAAK	1.3	0.114	-2.6931	-2.5271	-1.2871	2.8369	1.4092	-3.1398
6	First	VEWAK	534.0	2.728	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
7	First	VEAAP	1.5	0.176	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
8	First	VEHAK	34.0	1.531	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
9	First	VAAAK	0.8	-0.097	-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902
10	First	GEAAK	0.3	-0.523	2.2261	-5.3648	0.3049	3.0777	0.3891	-0.0701
11	First	LEAAK	2.5	0.398	-4.1921	-1.0285	-0.9801	3.0777	0.3891	-0.0701
12	First	FEAAK	2.0	0.301	-4.9217	1.2977	0.4473	3.0777	0.3891	-0.0701
13	First	VEGGK	0.1	-1.000	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
14	First	VEFAK	37.2	1.571	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
15	First	VELAK	3.9	0.591	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
16	Second	AAAAA	0.8	-0.097	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902
17	Second	AAYAA	2.9	0.462	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902
18	Second	AAWAA	5.6	0.748	0.0744	-1.7333	0.0902	0.0744	-1.7333	0.0902
19	Second	VAWAA	26.8	1.428	-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902
20	Second	VAWAK	27.9	1.446	-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902
21	Second	VKWAA	51.1	1.708	-2.6931	-2.5271	-1.2871	2.8369	1.4092	-3.1398
22	Second	VWAAK	1.1	0.041	-2.6931	-2.5271	-1.2871	-4.7548	3.6521	0.8524
23	Second	VAAWK	1.7	0.230	-2.6931	-2.5271	-1.2871	0.0744	-1.7333	0.0902
24	Second	EKWAP	20.1	1.303	3.0777	0.3891	-0.0701	2.8369	1.4092	-3.1398
25	Second	VKWAP	222.1	2.347	-2.6931	-2.5271	-1.2871	2.8369	1.4092	-3.1398
26	Second	RKWAP	96.6	1.985	2.8827	2.5215	-3.4435	2.8369	1.4092	-3.1398
27	Second	VEWVK	51.5	1.712	-2.6931	-2.5271	-1.2871	3.0777	0.3891	-0.0701
28	Second	PGFSP	7.9	0.898	-1.2201	0.8829	2.2253	2.2261	-5.3648	0.3049
29	Second	FSPFR	4.4	0.643	-4.9217	1.2977	0.4473	1.9607	-1.6324	0.5746
30	Second	RYLPT	2.5	0.398	2.8827	2.5215	-3.4435	-1.3944	2.323	0.0139

To visually compare log_rai and Pred Formula log_rai:

1. Select **Graph > Graph Builder**.
2. Drag log_rai from the **Variables** list to the area under the graph template indicated by **X**.
3. Drag Pred Formula log_rai to the **Y** position next to the vertical axis at the left of the graph template.
4. Click the **Line of Fit** icon above the graph template (third icon from the left).
5. From the **Line of Fit** panel at the lower left, deselect **Confidence of Fit**, as it is not relevant in this case.

This produces the graph shown in Figure 5.19.

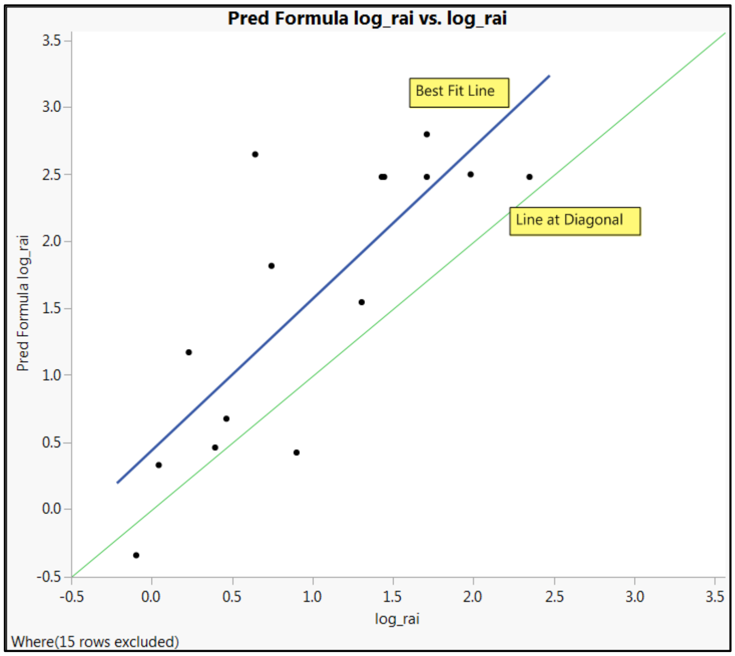
Figure 5.19: Predicted versus Actual Values of log_rai for Pruned Model



If Pred Formula log_rai were predicting log_rai exactly, the actual and predicted values would fall on a diagonal line. But, in fact, the line along which the points fall has slope greater than one.

For a better view, run the script Graph Builder. The resulting plot (Figure 5.20) shows a green line plotted at the diagonal. If the model were to fit the data well, the points should fall along this line. On average, the predicted values are higher than the actual values, some by as much as two log_rai units. Although the ultimate value of the model depends on whether differences of this magnitude are important, it appears that the model that was developed using data from the first study shows bias relative to predicting the data from the second study.

Figure 5.20: Predicted versus Actual Values with Line at Diagonal



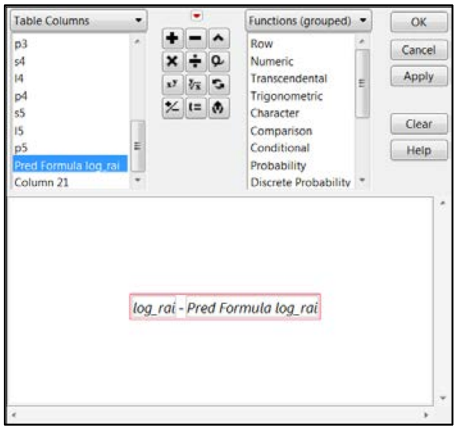
Comparing Residuals for Both Studies

We can gain additional insight by comparing the predicted values to the actual values for both sets of data. There are many ways to do this. We construct a residual plot showing both sets of observations.

To make a new formula column to calculate the residuals for all the data, complete the following steps. (The script is Save Residuals.)

1. To the right of the existing columns in the data table, double-click in the header area to add a new column.
2. Right-click in the header area and select **Formula**.
3. Enter the formula shown in Figure 5.21. Select log_rai in the **Table Columns** list, click the **–** sign on the operator pad to the right of the column list, and then select Pred Formula log_rai from the **Table Columns** list.
4. Click **OK**.
5. Click on the header for the new column and name it Residuals.

Figure 5.21: Formula for Residuals Column



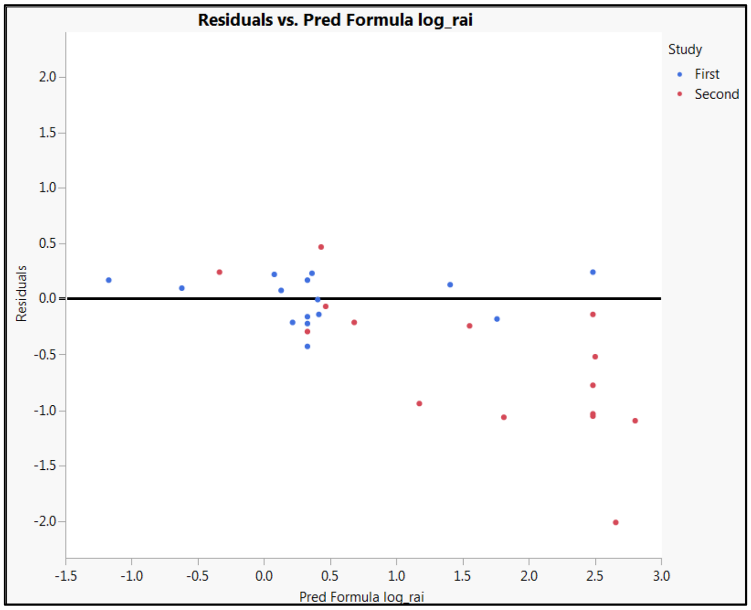
log_rai - Pred Formula log_rai

Now let’s use **Graph Builder** to construct the plot shown in Figure 5.22. (The script is Graph Builder 2.)

1. Select **Rows > Clear Row States** to remove the excluded and hidden states.
2. Select **Graph > Graph Builder**.
3. Drag Pred Formula log_rai from the **Variables** list to the area under the graph template indicated by **X**.
4. Drag Residuals to the **Y** position next to the vertical axis at the left of the graph template.
5. Drag Study to the **Color** box to the right of the graph template.
6. Deselect the **Smoother** icon above the graph template (second icon from the left).
7. Double-click on the vertical axis to open the **Y Axis Specification** menu. In the **Reference Lines** panel at the bottom, click **Add** to add a reference line at 0. Click **OK**.
8. If you want, drag the vertical axis settings to center the line at 0.
9. If you want to make the markers larger, right-click in the graph, select **Graph > Marker Size**, and then select the desired size. (We have selected **3, Large**.)
10. From the red triangle menu, deselect **Show Control Panel** to remove the control panel.

This produces the graph shown in Figure 5.22. It is clear that the model fits the data from the first study better than the data from the second study. Again we see that, for the second-study data, the predicted values tend to be too high, especially for larger predicted values.

Figure 5.22: Residual Plot for First and Second Study Data



Obtaining Additional Insight

This observation leads us to suspect that the two groups of observations are systematically different in some sense. To gain some insight, we construct a scatterplot matrix. But first, let's color the rows so that we can differentiate between data from the two studies in our scatterplot matrix. (The script is Color by Study, but note that the script does *not* create a legend window.)

1. Select **Rows > Color or Mark by Column**.
2. In the window that opens, select the column Study.
3. In that same window, check **Make Window with Legend**.
4. Click **OK**.

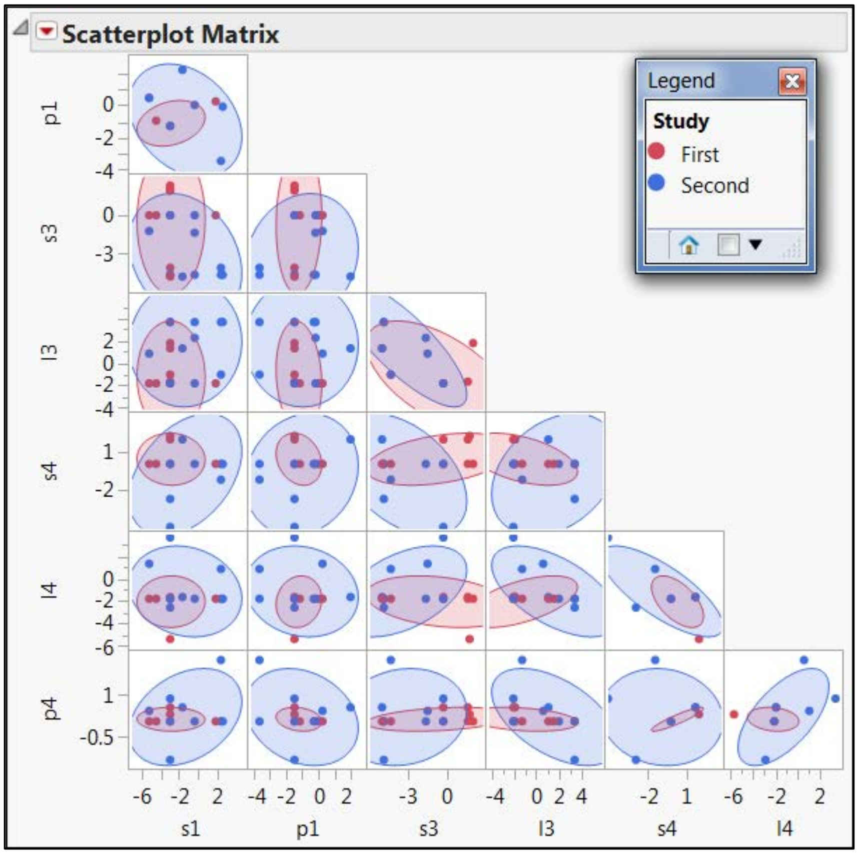
A small portable legend window appears, showing that the color red is associated with the training set and blue with the test set.

The following steps explain how to construct our scatterplot matrix in JMP 11. (The script is Scatterplot Matrix.) The steps in JMP 10 differ slightly.

1. Select **Graph > Scatterplot Matrix**.
2. From the Predictors column group, select s1, p1, s3, l3, p3, s4, l4, and p4, and enter these as **Y, Columns**.
3. Click **OK**.
4. From the red triangle menu in the resulting report, select **Group By**. Click **Grouped by Column** and select Study. These choices cause subsequent analyses to be run separately for each specified group.
5. Click **OK**. This brings you back to the **Scatterplot Matrix** report.
6. Select **Density Ellipses > Density Ellipses** and **Density Ellipses > Shaded Ellipses** from the red triangle menu options.

The scatterplot matrix is shown in Figure 5.23. You can click on the study names in the legend window to highlight the corresponding points in the scatterplot matrix. Note that the observations from the test set (second study) have measurements on the amino acid properties that vary more extensively than the observations for the training set (first study). This plot confirms our suspicions that the two sets of observations are different in some intrinsic way.

Figure 5.23: Scatterplot Matrix for First and Second Study Data



Now recall that the first 15 observations were from the study reported in Ufkes et al. (1978), while the last 15 were from the study reported in Ufkes et al. (1982). It appears that the second study used a broader range of amino acids than did the first. In addition, the second study integrated more variation in the positioning of amino acids in the peptide chain than did the first study. Were we to use our training set model to make predictions relative to the test set, we would be guilty of extrapolation.

In addition to noting differences in the peptides used in the study design, the authors indicate that the Bradykinin used in the two studies came from different sources. This fact might have had an impact on the response measure. Also, when studies conducted at two distinct time points are compared, so-called *lurking variables* (differences in setup, control variables, or measurement procedures) can come into play. So the fact that the model based on the training data does not generalize well to the test data in this case is perhaps not surprising.

At this point, we encourage you to derive a model using all 30 observations. You might want to use cross validation to determine the optimal number of factors. Do you obtain a good predictive model? Note that, with more factors, the differences in the two studies are more adequately modeled.

Conclusion

We have seen how to fit a PLS model to the subset of the data set Penta.jmp reflecting the 1978 study. We used knowledge about regression coefficients, VIPs, and loadings to prune that model. We then applied that model to the data from the 1982 study and found that it did not have good predictive ability. We concluded that the lack of predictive ability might be due to the comparatively smaller predictor range of observations from the first study, or to the impact of lurking variables, or perhaps to some other fundamental difference.

These findings raise two important points relative to modeling:

- In any empirical model building, interpolation is reasonable, but extrapolation is never reasonable.
- The nature and quality of the data is of paramount importance to building sound models. Good models require valid and representative data.

We also note that the dynamic visualization capabilities of JMP are key to developing the insights that help you make sound modeling decisions.

About The Authors

Ian Cox currently works in the JMP Division of SAS. Before joining SAS in 1999, he worked for Digital, Motorola, and BBN Software Solutions Ltd. and has been a consultant for many companies on data analysis, process control, and experimental design. A Six Sigma Black Belt, he was a Visiting Fellow at Cranfield University and is a Fellow of the Royal Statistical Society in the United Kingdom. Cox holds a Ph.D. in theoretical physics. In addition to Discovering Partial Least Squares with JMP, Ian co-authored the book Visual Six Sigma: Making Data Analysis Lean.

Marie Gaudard is a consultant in the North Haven Group, a small consulting firm specializing in statistical training and consulting using JMP. She earned her Ph.D. in statistics in 1977 and was a Professor of Statistics at the University of New Hampshire from 1977 until 2004. She has been heavily involved in statistical consulting since 1981. Gaudard has worked with a variety of clients in transactional areas, including government agencies and financial departments, as well as with manufacturers, including automotive, printing, paper, plastics, precision steel, and paving, as well as shipyards. She has also been involved in the analysis of medical data.

Gaudard has extensive experience in providing consulting and training courses for business and industry in the areas of Six Sigma, Design for Six Sigma, forecasting and demand planning, and data mining. In addition to Discovering Partial Least Squares with JMP, Maria co-authored the book Visual Six Sigma: Making Data Analysis Lean.