Inverse calibration in multicomponent analysis

As discussed previously, the most typical approach of multiple linear regression to multicomponent analysis is the direct calibration method, in which absorbance values obtained at different wavelengths are considered as dependent variables, whereas the component concentrations represent independent variables.

The method relies on the following equation for the total absorbance obtained at the i-th wavelength, with b_{ji} representing the normalized absorptivity of component j (with j = 1, 2, …, m) at the i-th wavelength:

$$A_i = b_{0i} + b_{1i}c_1 + b_{2i}c_2 + b_{3i}c_3 + \cdots + b_{\mathrm{m}i}c_{\mathrm{m}}$$

Sometimes this simple additive model may not describe the system completely.

There are two reasons for this:

- ✓ substances of interest may interfere with each other chemically in a way that affects their spectra
- ✓ mixtures from 'real-life' sources may contain substances other than those of interest, which provide a contribution to the absorbance.

In these cases it is better to use inverse calibration and calibrate with 'real-life' mixtures.

The term "inverse calibration" means that the analyte concentration is modelled as a function of absorbances (*i.e.*, the reverse of the classical method).

The regression equation takes the following form:

$$c_i = b_{0i} + b_{1i}A_1 + b_{2i}A_2 + \cdots + b_{pi}A_p$$

thus the concentration of the i-th component is obtained by combining terms related to the mixture absorbances A_j, measured at different wavelengths (with j = 1, 2, …, p).

Multiple linear regression is one of the regression methods that can be used for inverse calibration, *i.e.*, to predict one or more concentrations from a set of absorbance values.

Application of Multiple Linear Regression (MLR) to inverse calibration

As an example of application of MLR to inverse calibration, a dataset obtained by measuring the UV absorbance at 6 different wavelengths (A_1, A_2, …, A_6) for 10 specimens (A, B, …, J) containing 3 compounds of interest at different concentrations (c_1, c_2 and c_3) , is reported in the following table (note that absorbance values are multiplied by 100):

Specimen	c_1	c_2	c_3	A_1	A_2	A_3	A_4	A_5	A_6
A	0.89	0.02	0.01	18.7	26.8	42.1	56.6	70.0	83.2
B	0.46	0.09	0.24	31.3	33.4	45.7	49.3	53.8	55.3
C	0.45	0.16	0.23	30.0	35.1	48.3	53.5	59.2	57.7
D	0.56	0.09	0.09	20.0	25.7	39.3	46.6	56.5	57.8
E	0.41	0.02	0.28	31.5	34.8	46.5	46.7	48.5	51.1
F	0.44	0.17	0.14	22.0	28.0	38.5	46.7	54.1	53.6
G	0.34	0.23	0.20	25.7	31.4	41.1	50.6	53.5	49.3
H	0.74	0.11	0.01	18.7	26.8	37.8	50.6	65.0	72.3
I	0.75	0.01	0.15	27.3	34.6	47.8	55.9	67.9	75.2
J	0.48	0.15	0.06	18.3	22.8	32.8	43.4	49.6	51.1

The MLR equation can thus be expressed, for each of the 3 compounds, in the following form:

$$c_i = b_{0i} + b_{1i}A_1 + b_{2i}A_2 + \cdots + b_{6i}A_6$$

Notably, MLR can be applied since the number of specimens is larger than the number of predictors.

MLR calculations for c_1, performed using the Minitab program, provided the following results:

Predictor	Coef	SE Coef	T	P
Constant	0.05010	0.08945	0.56	0.615
A1	0.002525	0.008376	0.30	0.783
A2	−0.009387	0.008811	−1.07	0.365
A3	0.003754	0.005852	0.64	0.567
A4	−0.009197	0.005140	−1.79	0.172
A5	−0.001056	0.005373	−0.20	0.857
A6	0.017881	0.002249	7.95	0.004

S = 0.0188690 R-Sq = 99.6% R-Sq(adj) = 98.9% *fitting*

PRESS = 0.0274584 R-Sq(pred) = 90.55% *prediction*

The resulting regression equation for c_1 was the following:

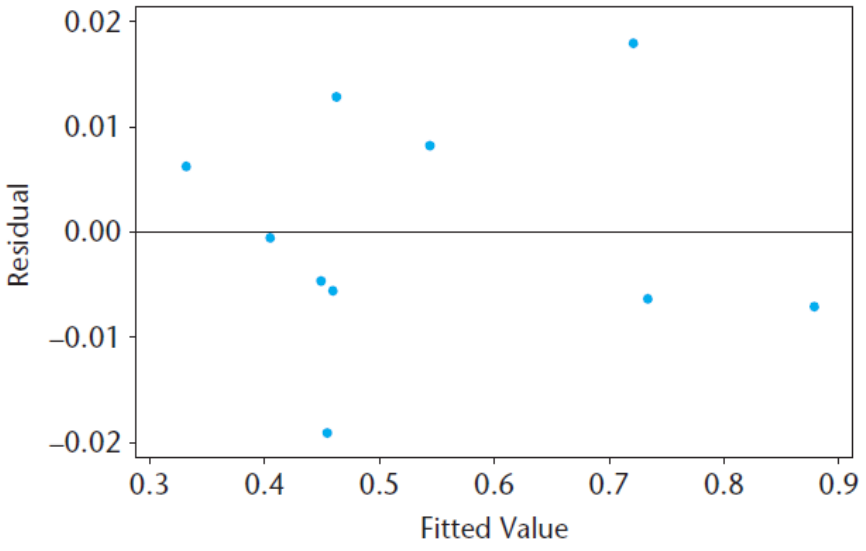
$$c_1 = 0.0501 + 0.00252A_1 - 0.00939A_2 + 0.00375A_3 - 0.00920A_4 - 0.00106A_5 + 0.0179A_6$$

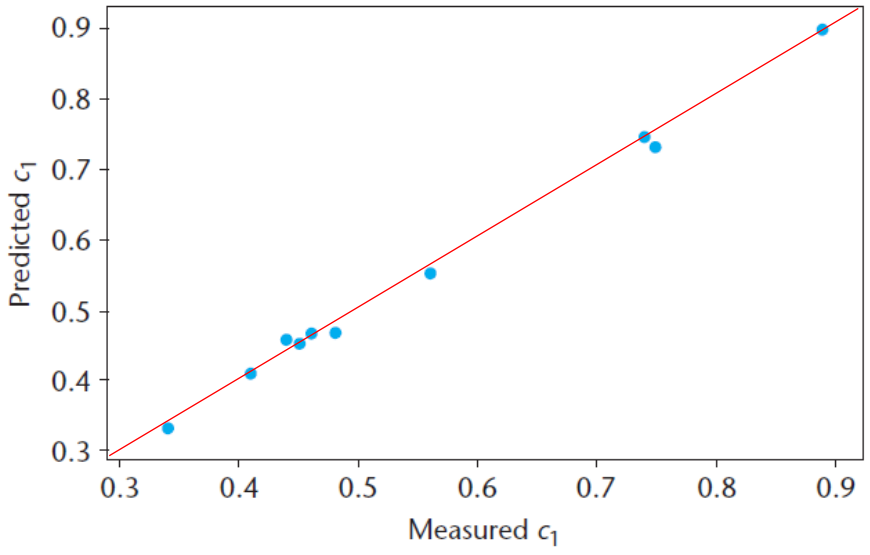
Regression equations for c_2 and c_3 were the following:

$$c_2 = 0.027 + 0.0067A_1 - 0.0007A_2 - 0.0184A_3 + 0.0141A_4 + 0.0160A_5 - 0.0152A_6$$

$$c_3 = -0.0776 + 0.00168A_1 + 0.00754A_2 + 0.00668A_3 + 0.00221A_4 - 0.00510A_5 - 0.00237A_6$$

Plots of residuals vs. fitted values and of predicted values of c_1 vs. measured ones are shown in the following figures:





As apparent, residuals do not show any particular pattern and points are reasonably close to a straight line with unitary slope in the plot comparing predicted with measured concentrations.

The prediction performance can be validated by using a cross validation method based on the "leave-one-out" principle.

Specifically, values for the first specimen (A) are omitted from the data set and those for the remaining specimens (B–J) are used to find the regression equation of, *e.g.*, c_1 on A_1, A_2, etc. The same procedure is then repeated, leaving each specimen out in turn.

The Predicted Residual Error Sum of Squares (PRESS) is then calculated in each case: the closer is the value of the PRESS to zero, the better is the predictive power of the model.

Actually, information on the relevance of predictors in terms of model quality can be inferred directly from the tabular summary of regression statistics provided by the Minitab software.
Since only the P value for the A_6 regressor is lower than 0.05, it can be concluded that any one of predictors from A_1 to A_5 could be left out of the model without reducing its effectiveness.

It is finally worth noting that many more than 6 absorbance values are usually available when an entire UV spectrum is acquired for a multicomponent mixture.

In this case other approaches need to be adopted to make a better use of data. Principal Component Regression is one of them.

Principal Components Regression (PCR)

The Ordinary Least Squares solution for Multiple Linear Regression may be ill-conditioned when:

- ✓ the predictors are highly correlated, thus leading to mathematical complications (*e.g.*, difficulties in inverting matrix $\mathbf{X}^T\mathbf{X}$ in the presence of multicollinearity) resulting in unreliable predictions;

- ✓ the number of predictors exceeds the number of training samples.

As an example, if the correlation matrix for data considered before is visualized:

	c1	c2	c3	A1	A2	A3	A4	A5
c2	-0.637							
c3	-0.717	0.088						
A1	-0.482	-0.116	0.947					
A2	-0.260	-0.194	0.832	0.941				
A3	-0.001	-0.413	0.677	0.841	0.936			
A4	0.625	-0.355	-0.096	0.148	0.422	0.598		
A5	0.899	-0.434	-0.541	-0.293	-0.002	0.227	0.857	
A6	0.977	-0.608	-0.603	-0.346	-0.089	0.161	0.771	0.960

It is apparent that a remarkable degree of correlation exists between some predictors.

A possible solution consists in projecting each measurement into a lower-dimensional subspace.

This procedure, which is the base of PCR, is equivalent to the extraction of principal components.

The first step consists in applying the already discussed matricial equation including data matrix **X**:

$$\mathbf{X} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}^{T}$$

Afterwards, matrix **T = XP** is obtained.

As evidenced before, **T** represents the score matrix, whereas **P** is the loadings (or eigenvectors) matrix.

The number of principal components is usually equal to the number of variables (p) in the initial stage of calculation, yet a low-rank approximation of X can be obtained in a second step, by keeping just the first k (with k < p) principal components, since they usually account for most of the total variance of data:

$$\mathbf{X} = (\mathbf{P}\,\mathbf{\Lambda})\,\mathbf{P}^{T} = \mathbf{T}\,\mathbf{P}^{T} \Rightarrow \mathbf{X} \cong \mathbf{T}_k\,\mathbf{P}_k^{T} \Rightarrow \mathbf{T}_k = \mathbf{X}\,\mathbf{P}_k$$

In the final step the regression problem is treated in a lower-dimensional predictor space by using principal components as the new predictors:

$$\mathbf{y} = \mathbf{T}_k\,\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \quad \Longleftrightarrow \quad \mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\alpha}$ is the vector of regression coefficients in the k-dimensional space of principal components.

Based on the analogy shown above, the following equations can be written:

$$\mathbf{y} = \mathbf{T}_k\,\boldsymbol{\alpha} + \boldsymbol{\varepsilon} = \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}\,\mathbf{P}_k\;\mathbf{P}_k^{\mathrm{T}}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Since $\mathbf{T}_k = \mathbf{X}\,\mathbf{P}_k$ vector $\boldsymbol{\alpha}$ can be expressed as:

$$\boldsymbol{\alpha} = \mathbf{P}_k^{\mathrm{T}}\,\boldsymbol{\beta}$$

Considering the properties of the $\mathbf{P}_k$ matrix, this equation is equivalent to the following one:

$$\boldsymbol{\beta} = \mathbf{P}_k\,\boldsymbol{\alpha}$$

Starting from the equation considered in multiple linear regression for vector b:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{y}$$

vector $\mathbf{a}_{PCR}$, representing estimates of regression coefficients in the principal component space, can be obtained using the following equations:

$$\mathbf{a}_{PCR} = \mathbf{P}_k^T\,\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{P}_k^T\,\mathbf{y} = (\mathbf{X}^T\,\mathbf{P}_k^T\;\mathbf{X}\,\mathbf{P}_k)^{-1}\,\mathbf{T}_k^T\,\mathbf{y} = (\mathbf{T}_k^T\mathbf{T}_k)^{-1}\,\mathbf{T}_k^T\,\mathbf{y}$$

Since: $\beta = \mathbf{P}_k\,\alpha$

vector $\mathbf{b}_{PCR}$, representing estimates of regression coefficients in the space of original variables can be calculated from vector $\mathbf{a}_{PCR}$:

$$\mathbf{b}_{PCR} = \mathbf{P}_k\;\mathbf{a}_{PCR} = \mathbf{P}_k\,(\mathbf{T}_k^T\mathbf{T}_k)^{-1}\,\mathbf{T}_k^T\,\mathbf{y}$$

Once this vector is known, the vector of predicted values corresponding to a new set of variable measurements, $\mathbf{X}_{new}$, can be readily obtained:

$$\hat{\mathbf{y}} = \mathbf{X}_{\mathbf{new}}\,\mathbf{b}_{PCR} = \mathbf{X}_{\mathbf{new}}\,\mathbf{P}_k\mathbf{a}_{PCR} = \mathbf{T}_{k\,\mathbf{new}}\,\mathbf{a}_{PCR}$$

A numerical example of Principal Component Regression

Let us re-consider the dataset previously shown for the MLR application to inverse calibration:

Specimen	c_1	c_2	c_3	A_1	A_2	A_3	A_4	A_5	A_6
A	0.89	0.02	0.01	18.7	26.8	42.1	56.6	70.0	83.2
B	0.46	0.09	0.24	31.3	33.4	45.7	49.3	53.8	55.3
C	0.45	0.16	0.23	30.0	35.1	48.3	53.5	59.2	57.7
D	0.56	0.09	0.09	20.0	25.7	39.3	46.6	56.5	57.8
E	0.41	0.02	0.28	31.5	34.8	46.5	46.7	48.5	51.1
F	0.44	0.17	0.14	22.0	28.0	38.5	46.7	54.1	53.6
G	0.34	0.23	0.20	25.7	31.4	41.1	50.6	53.5	49.3
H	0.74	0.11	0.01	18.7	26.8	37.8	50.6	65.0	72.3
I	0.75	0.01	0.15	27.3	34.6	47.8	55.9	67.9	75.2
J	0.48	0.15	0.06	18.3	22.8	32.8	43.4	49.6	51.1

A Principal Component Regression (PCR) of these data can be performed using the Minitab 18 program, *i.e.*, by performing a multivariate linear regression using principal components obtained preliminarily by PCA.

In the upper part of the following table eigenvalues corresponding to principal components, with their specific and cumulative contributions to total variance, as obtained after PCA, are reported.

Principal Component Analysis: A1, A2, A3, A4, A5, A6

Eigenanalysis of the Covariance Matrix

Eigenvalue	210.01	73.86	4.62	0.93	0.79	0.28
Proportion	0.723	0.254	0.016	0.003	0.003	0.001
Cumulative	0.723	0.977	0.993	0.996	0.999	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
A1	−0.124	−0.592	−0.253	−0.048	0.340	0.672
A2	−0.017	−0.513	0.048	0.196	0.493	−0.673
A3	0.066	−0.571	−0.102	0.128	−0.793	−0.118
A4	0.244	−0.239	0.575	−0.743	−0.002	−0.002
A5	0.510	−0.042	0.545	0.602	0.059	0.276
A6	0.813	0.043	−0.544	−0.168	0.091	−0.075

In the lower part of the table values of loadings of each variable in the six PCs are reported. Notably, variable A6, the only appearing significant in the MLR model developed before, is the one having the highest loading on the first PC, that explains more than 72% of total variance.

Moreover, the first three principal components already account for the 99.3% of the total variance, thus a $\mathbf{T}_k$ matrix, with k = 3, can be calculated.

The corresponding values, corresponding to scores, are reported in the following table, with the notation adopted by Minitab, which uses Z instead of T.

Specimen	Z_1	Z_2	Z_3
A	117.1	−61.7	17.7
B	83.0	−73.4	16.6
C	89.0	−76.1	20.8
D	86.8	−58.4	18.3
E	76.2	−74.0	14.5
F	81.9	−60.5	19.0
G	78.7	−67.0	22.3
H	104.0	−58.1	17.9
I	108.6	−74.1	18.1
J	76.9	−51.5	17.3

These scores can thus be employed for a multivariate linear regression of concentration c_1 based on Z_1, Z_2 and Z_3:

Regression Analysis: c1 versus z1, z2, z3

The regression equation is

$c1 = 0.0685 + 0.0119Z1 + 0.00419Z2 - 0.0171Z3$

Predictor	Coef	SE Coef	T	P
Constant	0.06849	0.06571	1.04	0.337
Z1	0.0118502	0.0003480	34.05	0.000
Z2	0.0041884	0.0005868	7.14	0.000
Z3	−0.017058	0.002345	−7.27	0.000

S = 0.0151299 R-Sq = 99.5% R-Sq(adj) = 99.3%

PRESS = 0.00301908 R-Sq(pred) = 98.96%

It is worth noting that the PRESS statistic is lower than the one obtained using MLR.

Moreover, based on P values, all the regression coefficients other than the constant term are significantly different from zero, thus the possibility of fitting a model with zero intercept could be explored.

The regression equation can be expressed as:

$$c_1 = 0.0685 - 0.0119Z_1 + 0.00419Z_2 - 0.171Z_3$$

but it can also be transformed in terms of A_i variables, considering the equations relating each principal component Z_i to A_i variables. As an example:

$$Z_1 = -0.124A_1 - 0.017A_2 + 0.066A_3 + 0.244A_4 + 0.510A_5 + 0.813A_6$$

A similar calculation can be made also for Z_2 and Z_3, thus the regression equation can be finally be expressed as:

$$c_1 = 0.06849 + 0.00037A_1 - 0.00317A_2 + 0.00014A_3 - 0.00792A_4 - 0.00343A_5 + 0.01909A_6$$

Interestingly, A6 is still the numerically most relevant variable in the regression model.

Partial Least Squares (PLS) regression

Partial Least Squares regression is a biased regression method. As already discussed when describing the validation of methods, a biased method looks for a bias-variance trade-off, *i.e.*, for a compromise between the model complexity and its variability.

PLS regression can be used when:

- ✓ the samples/variables ratio is lower than 1
- ✓ highly correlated variables are present
- ✓ more than one response is present (*i.e.*, a matrix of responses, **Y**, replaces the vector of responses **y**).

Using PLS a unique model, explaining all responses, rather than as many models as the available responses (like for PCR), can be obtained.

This is very useful when responses are correlated between each other.

Supposing that **X** and **Y** represent matrices of variables and responses, respectively, the first step of a PLS method is the decomposition of both matrices **X** and **Y**:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \quad \mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}}$$

By analogy with procedures described for PCA and PCR, this operation corresponds to projecting the two matrices into spaces of latent variables **T** and **U**, respectively, and then a regression between **T** and **U** is performed.

The decomposition can be made independently, so that couples of component vectors **t** and **u** are iteratively selected until the couple leading to the maximum covariance between matrices **X** and **Y** is found.

Peculiar features of the PLS method are:

1) PLS components are selected to generate the maximum reduction of the covariance matrix ($\mathbf{XY}^T$), so that the method provides the minimum required number of variables

2) One component at a time is provided (for each iteration), thus PLS can be considered as a step-wise procedure

3) The procedure goes on until significant components, able to improve the predictive power of the model, exist

4) The iterative process stops when there are no further couples of components in **X** and **Y** correlated enough or when there is no further useful information that can be extracted from **X** and used to predict **Y**

5) The number M of optimal components in PLS is determined by validation, *i.e.*, it is the one leading to maximum R^2_{cv}.

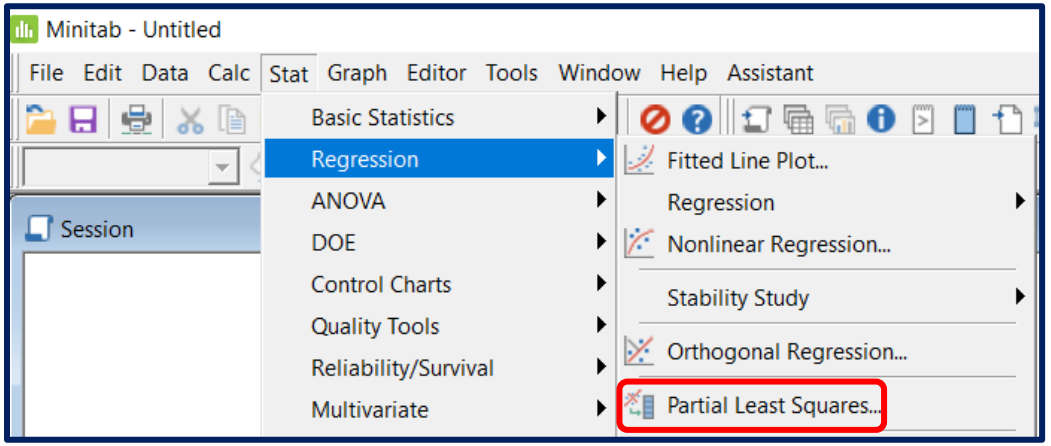
A numerical example of PLS

Let us re-consider the dataset previously used for inverse calibration based on MLR and for PCR with matrices **X** and **Y** emphasised:

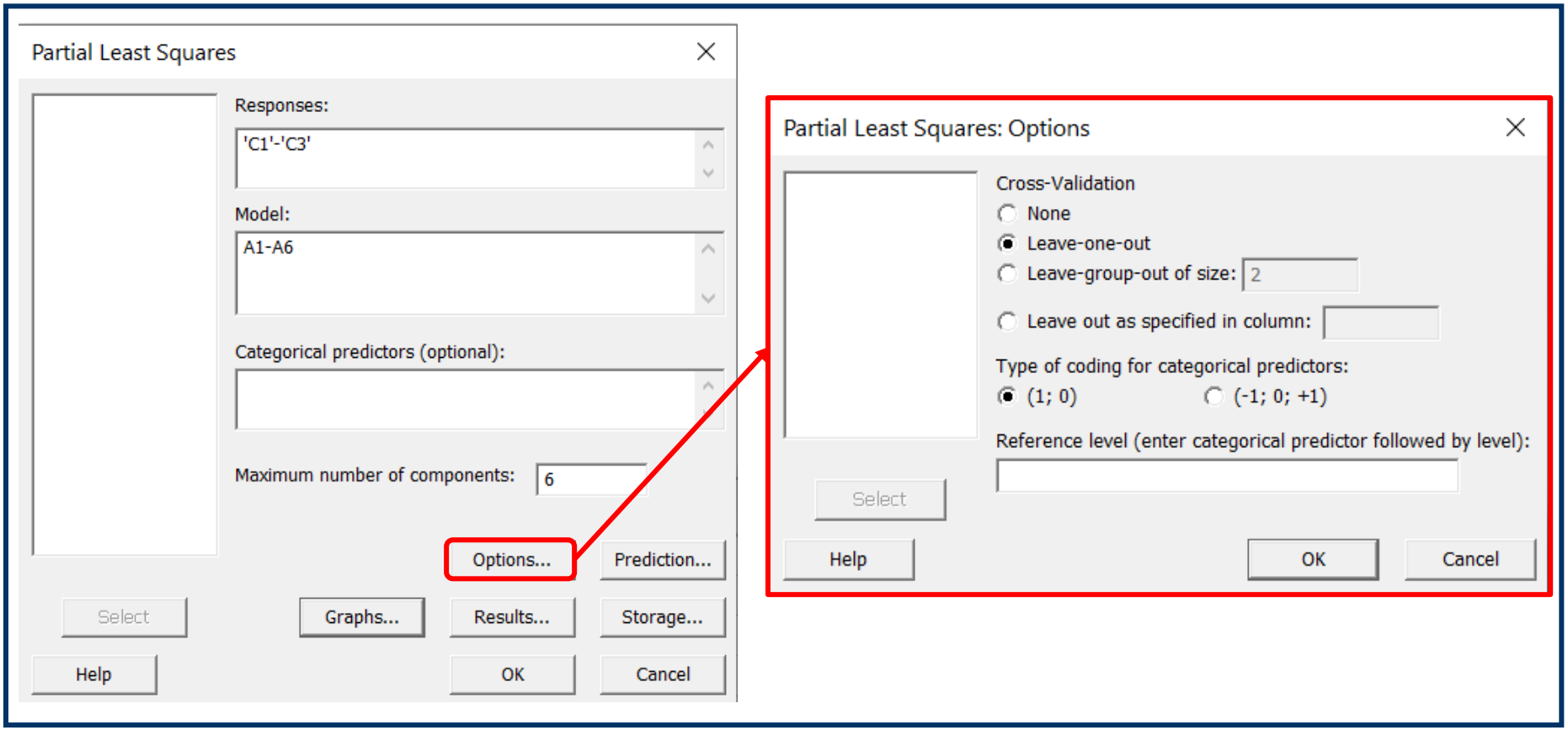
Specimen	c_1	c_2	c_3	A_1	A_2	A_3	A_4	A_5	A_6
A	0.89	0.02	0.01	18.7	26.8	42.1	56.6	70.0	83.2
B	0.46	0.09	0.24	31.3	33.4	45.7	49.3	53.8	55.3
C	0.45	0.16	0.23	30.0	35.1	48.3	53.5	59.2	57.7
D	0.56	0.09	0.09	20.0	25.7	39.3	46.6	56.5	57.8
E	0.41	0.02	0.28	31.5	34.8	46.5	46.7	48.5	51.1
F	0.44	0.17	0.14	22.0	28.0	38.5	46.7	54.1	53.6
G	0.34	0.23	0.20	25.7	31.4	41.1	50.6	53.5	49.3
H	0.74	0.11	0.01	18.7	26.8	37.8	50.6	65.0	72.3
I	0.75	0.01	0.15	27.3	34.6	47.8	55.9	67.9	75.2
J	0.48	0.15	0.06	18.3	22.8	32.8	43.4	49.6	51.1

$\mathbf{Y}_{(10x3)}$ $\mathbf{X}_{(10x6)}$

The PLS regression on c_1 can be performed using the Minitab 18 software, using the Stat > Regression > Partial Least Squares… pathway.

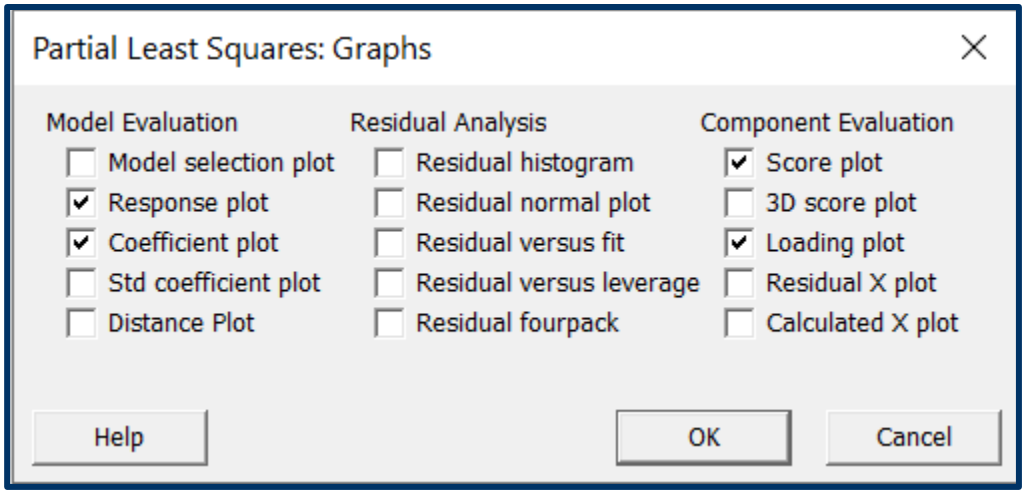


Worksheet columns including data for responses and for predictors are selected appropriately in the Partial Least Squares window (the Model box is used for predictor columns):

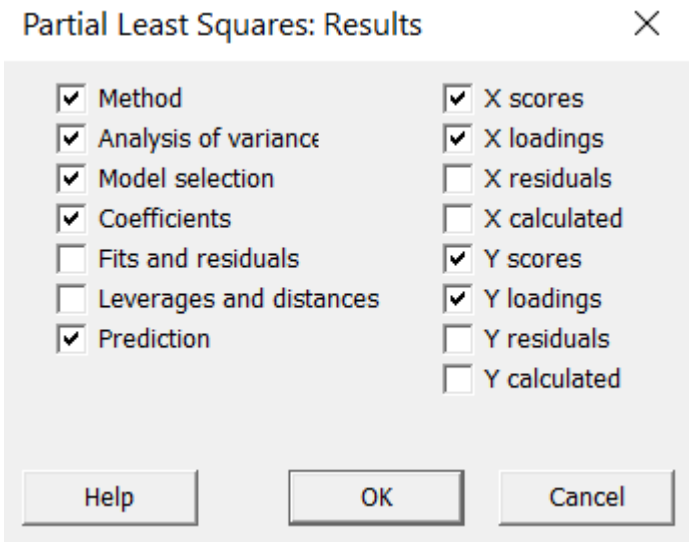


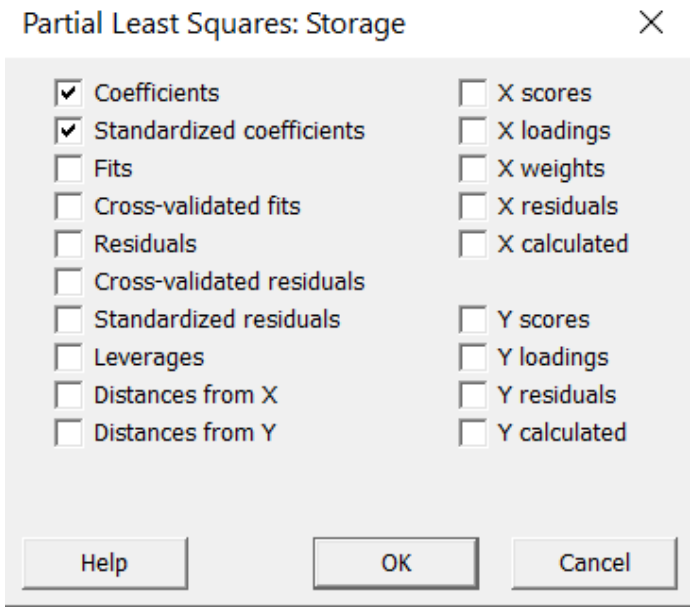
The type of Cross-Validation (if any) can be selected in the Options… window. The Prediction… window can be used to select further columns, where new values of variables and the corresponding values of responses can be stored.

Several types of Graphs, concerning model and component evaluation and residual analysis, can be selected in the Graphs… window:



Several types of results can be selected in the Results…window, so that they can be displayed in the Session window after calculations. Many output data can also be selected in the Storage… window for storage in the Worksheet after calculations.





In the following panels the results obtained by choosing the "leave-one-out" method for cross-validation are reported for response c_1.

PLS Regression: c1 versus A1, A2, A3, A4, A5, A6

Number of components selected by cross-validation: 4
Number of observations left out per group: 1
Number of components cross-validated: 6

Analysis of Variance for c1

Source	DF	SS	MS	F	P
Regression	4	0.289476	0.0723690	333.84	0.000
Residual Error	5	0.001084	0.0002168		
Total	9	0.290560			

Model Selection and Validation for c1

Components	X Variance	Error SS	R-Sq	PRESS	R-Sq (pred)
1	0.457325	0.0287984	0.900887	0.0469069	0.838564
2	0.957200	0.0255230	0.912159	0.0511899	0.823823
3	0.988793	0.0021123	0.992730	0.0078758	0.972894
4	0.992990	0.0010839	0.996270	0.0052733	0.981851
5		0.0010724	0.996309	0.0186933	0.935664
6		0.0010681	0.996324	0.0274584	0.905498

Four components have been found to provide the best model for c_1 and an explanation for this choice is provided by PRESS values reported in the Minitab's output.

Indeed, the lowest PRESS value (0.0052733), slightly lower than the one found after PCR using 3 components, is obtained for a PLS model with 4 components and it is increased when further components are added. Equivalently, the R^2 value in prediction is highest for a 4-component model, then it is decreased at the increase of the number of components.

Coefficients obtained for the PLS model are reported in the table on the right:

	c1	c1 standardized
Constant	0.0426293	0.00000
A1	0.0039542	0.11981
A2	−0.0111737	−0.27695
A3	0.0038227	0.10753
A4	−0.0092380	−0.22261
A5	−0.0003408	−0.01425
A6	0.0176165	1.16114

Consequently, the regression equation is:

$$c_1 = 0.0426 + 0.0040A_1 - 0.0112A_2 + 0.0038A_3 - 0.0092A_4 - 0.0003A_5 + 0.0176A_6$$

Comparison between MLR, PCR and PLS approaches

The following equations were obtained for c_1 according to the approach adopted:

$$\text{MLR: } c_1 = 0.0501 + 0.00252A_1 - 0.00939A_2 + 0.00375A_3 - 0.00920A_4 - 0.00106A_5 + 0.0179A_6$$

$$\text{PCR: } c_1 = 0.06849 + 0.00037A_1 - 0.00317A_2 + 0.00014A_3 - 0.00792A_4 - 0.00343A_5 + 0.01909A_6$$

$$\text{PLS: } c_1 = 0.0426 + 0.0040A_1 - 0.0112A_2 + 0.0038A_3 - 0.0092A_4 - 0.0003A_5 + 0.0176A_6$$

Although the coefficients differ from one model to another, they have the same sign in each equation and in all three equations the coefficient for A6 dominates.

It is finally worth recalling that MLR cannot be carried out when the number of variables is greater than the number of specimens, which is not a rare circumstance.
In this case, rather than selecting only a few variables, it is better to reduce their number by using PCR or PLS.

Many recent applications of PCR and PLS have arisen in molecular spectroscopy, where strongly overlapping absorption or emission spectra are often observed, even in simple mixtures.