

Tutorial 4.

Prediction with Near Infrared (NIR) Data

This tutorial¹ shows how a validation sample can be used to compare the performance of different models and methods. In particular, we utilize Cookies NIR data, analyzed earlier by Osbourne, et. al. (1984), Brown et. Al. (2001) and Kraemer and Boulesteix (2011), and show that improved performance can be obtained by removing predictors associated with the highest wavelengths. Use of Cross-validation to determine the number of components based on 700 predictors yields good performance regardless of the particular regression method used. However, we will see that the Correlated Component Regression method (CCR.LM) provides an improvement over the PLS regression method (regardless whether the predictors are standardized) due to the failure of PLS to accurately assess the unreliability of predictors associated with the highest wavelengths.

Cookie Dough Data

These data arise from an experiment designed to test the feasibility of NIR spectroscopy to obtain accurate measurements of 4 dependent variables -- calculated percentages of the ingredients Fat (Y1), sucrose (Y2), dry flour (Y3), and water (Y4) of biscuit dough pieces (formed but unbaked biscuits). The 700 predictor variables (wavelengths) were obtained with quantitative NIR spectroscopy from 700 different wave-lengths measured from 1100 to 2498 nanometers (nm) in steps of 2 nm. The calibration (training) set consists 40 samples and a further 32 samples were used as a separate validation set.

Osbourne, B., T. Fearn, A. Miller, and S. Douglas (1984). Application of near infrared reflectance spectroscopy to compositional analysis of biscuits and biscuit dough. *Journal of Science and Food Agriculture*, 35:99-105.

Brown, Fearn and Vannucci, *JASA* (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. *Journal of the American Statistical Association*, 96(454):398-408.

¹ To reproduce the results shown in this tutorial exactly, you will need to fix the seed to '123456789'. To fix the seed in XLSTAT, go to Options, then click on the Advanced tab. Check the box to activate the option 'Fix the seed to:', and change the seed to 123456789.

Kraemer, N. and Boulesteix, A. (2011). Penalized Partial Least Squares (PPLS). R Package. Data source: B.G. Osborne, T. Fearn, A.R. Miller, and S. Douglas (1984). Application of Near-Infrared Reflectance Spectroscopy to Compositional Analysis of Biscuits and Biscuit Dough. *Journal of the Science of Food and Agriculture*, 35, pp. 99 - 105.

An Excel sheet containing the data for use in this tutorial can be downloaded by clicking [here](#).

Goal of the Correlated Component Regression in this example

This tutorial focuses on prediction of fat content (Y_1) from the $P = 700$ wavelengths, with only $N=40$ samples. Since $P \gg N$, these constitute high dimensional data which require special methods to avoid overfitting. The goal of this tutorial is to show how to use XLSTAT-CCR to compare the performance of 3 methods in obtaining reliable predictions. We utilize cross-validation (CV) techniques to determine the tuning parameter K (reflecting the proper amount of regularization), and utilize 32 separate (validation) samples, to evaluate the different methods.

The 3 methods being compared are CCR.LM, PLS with unstandardized predictors, and PLS with standardized predictors. Unlike PLS regression, CCR.LM is scale invariant and hence yields the same predictions with standardized or unstandardized predictors. The results of the comparison show that CCR.LM performs best.

In addition to the high-dimensional aspect, these data present another challenge in that the standard deviations of the wavelengths are much higher ($> .1$) for wavelengths 659-700 than from the lower wavelengths (Figure 1). These wavelengths were “thought to contain little useful information” (Brown et. al., 2001) and hence excluded from that analysis. Our analyses here agrees with that conclusion suggesting that the large standard deviations reflect larger amounts of irrelevant variation in these higher frequencies.

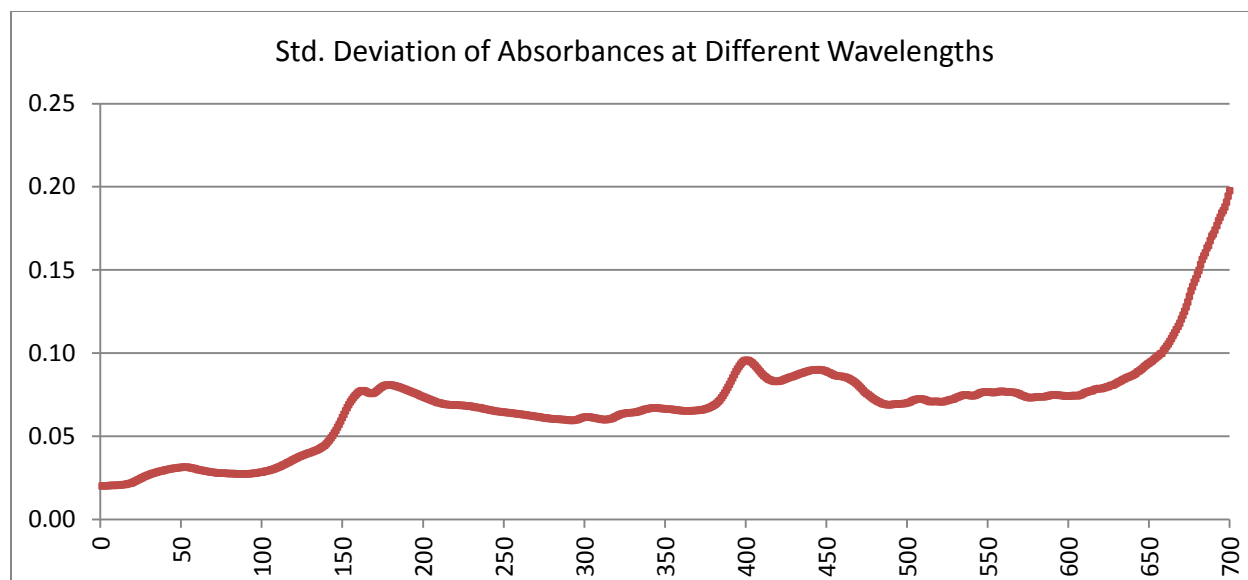


Figure 1. Plot of standard deviation for each of the 700 predictor variables (wavelengths)

The results (Tables 1A, 1B and 1C) suggest that CCR.LM outperforms PLS regression on these data, providing higher values on both the cross-validated R^2 and the R^2 obtained from the validation data. In addition, the model obtained from CCR.LM is simpler, requiring only 9 components compared to 13 for PLS regression.

Goodness of fit statistics:				
	Value	Validation	Cross-validation	Std. dev.(CV)
Number of observations	40	32		
Sum of weights	40	32		
R^2	0.991	0.977	0.964	0.005

Table 1A: Goodness of fit statistics: CCR.LM with P=700 ($K^*=9$ components)


Goodness of fit statistics:				
	Value	Validation	Cross-validation	Std. dev.(CV)
Number of observations	40	32		
Sum of weights	40	32		
R^2	0.997	0.975	0.952	0.007

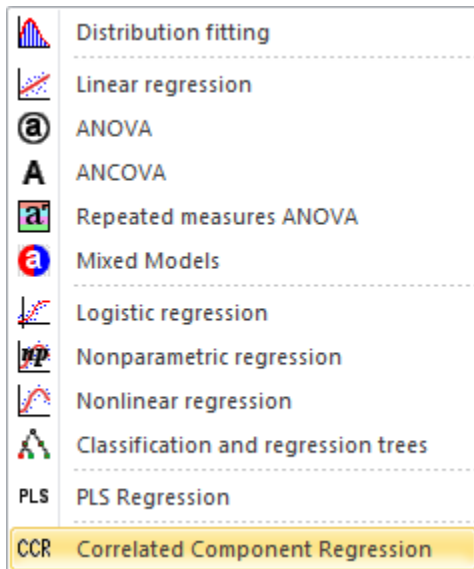
Table 1B: Goodness of fit statistics: PLS with P=700 (unstandardized) predictors, $K^*=13$ components

Goodness of fit statistics:				
	Value	Validation	Cross-validation	Std. dev.(CV)
Number of observations	40	32		
Sum of weights	40	32		
R ²	0.998	0.970	0.958	0.006

Table 1C: Goodness of fit statistics: PLS with P=700 (standardized) predictors, K*=14 components

Setting up a Correlated Component Regression (CCR) model

To activate the Correlated Component Regression dialog box, start XLSTAT by clicking on the  button in the Excel toolbar, and then select the **XLSTAT / Modeling data / Correlated Component Regression** command in the Excel menu or click the corresponding button on the **Modeling data** toolbar.



Once you have selected CCR, the **Correlated Component Regression** dialog box is displayed.

To setup the CCR runs, in the **Y / Dependent variable(s)** field, select (see the tutorial on [Selecting data](#) for more information on this topic) Column B (*Fat*). The fat content

represents the "Ys" of the model as we want to explain these as a function of the 700 wavelengths.

In the **X / Predictors** field, select columns I through AAF corresponding to the variable.

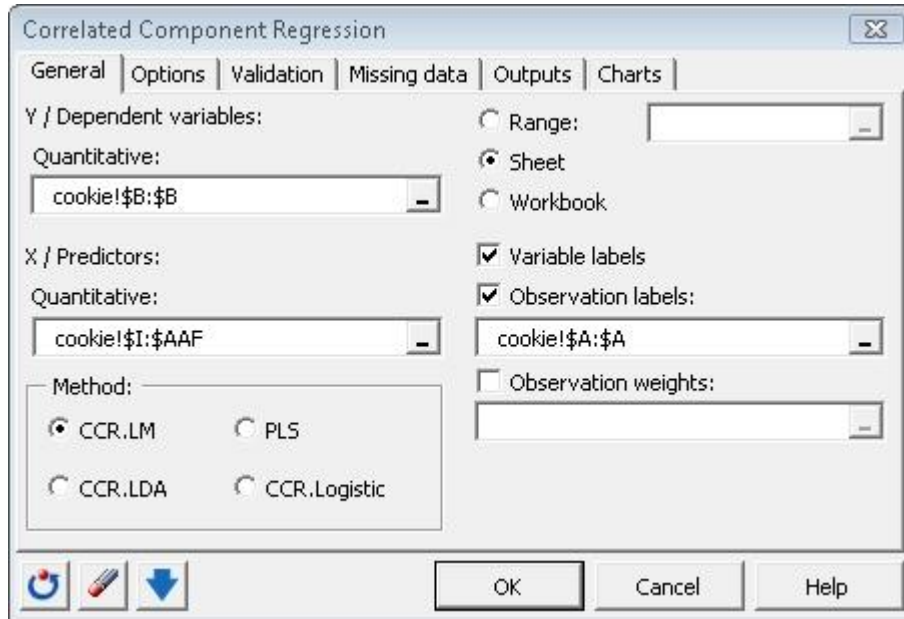


Figure 2. General Tab

To determine the number of components, in the **Options** tab of the dialog box, activate the 'Automatic' option and enter '20' in the 'Max components' field.

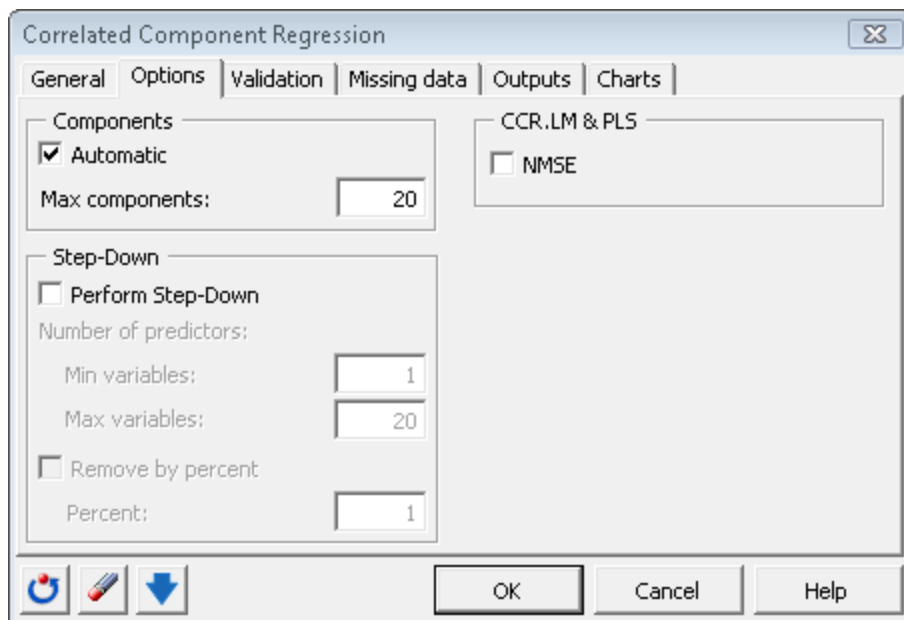


Figure 3. Options Tab

In the **Validation** tab of the dialog box, activate the **Validation** option and select 'N last rows' from the **Validation set** drop down menu. In the **Number of observations** field, type '32'. We have now specified the 'Training set' as the first 40 rows of the data file and the last 32 rows of the data file will be used as the validation set. Note that the Cross-validation option in the **Validation** tab is automatically activated with the default parameters (1 round of 10-folds). Change the default number of rounds from '1' to '10'. Change the default number of folds from '10' to '5'.

Make sure that the **settings** are **as shown below**.

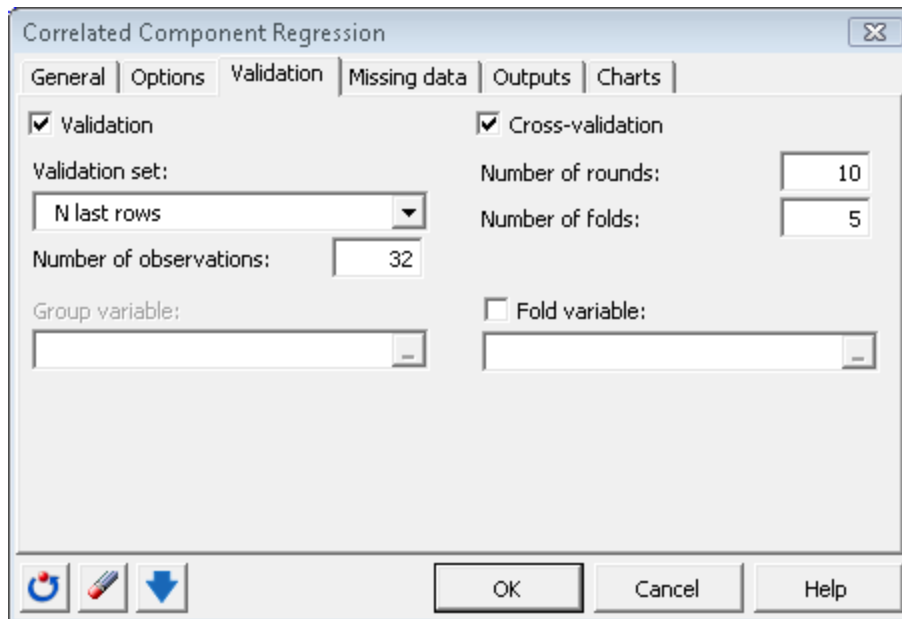


Figure 4. Validation Tab

The fast computations start when you click on **OK**.

Interpreting CCR results

From the results drop down menu, select "Cross-validation results".

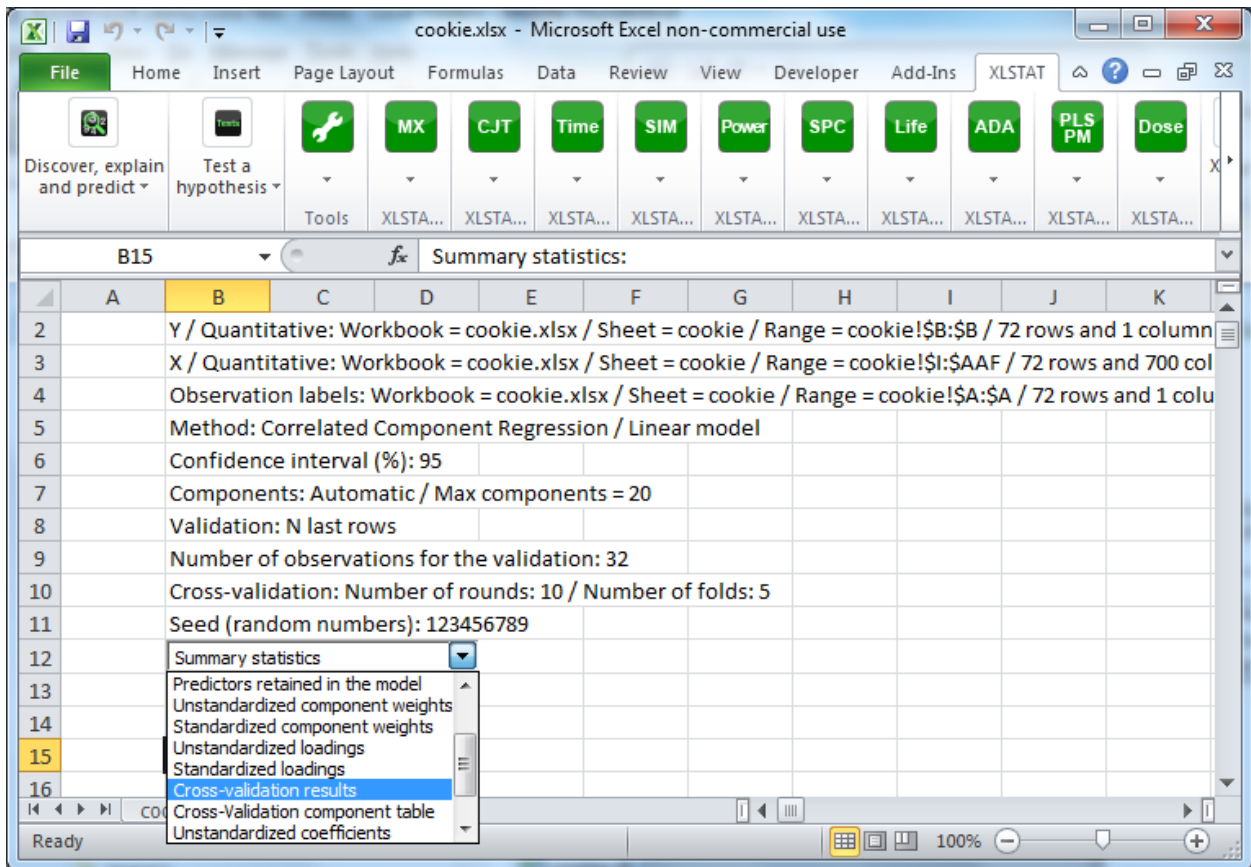


Figure 5. Navigating to "Cross-validation results" output

Cross-validation results:

Cross-Validation component table:

Components	R ²	Std. dev.(R ²)
1	0.237	0.040
2	0.506	0.063
3	0.759	0.028
4	0.914	0.014
5	0.948	0.006
6	0.948	0.007
7	0.945	0.008
8	0.955	0.007
9	0.962	0.006
10	0.960	0.004
11	0.957	0.007
12	0.958	0.008
13	0.958	0.008
14	0.958	0.007
15	0.958	0.007
16	0.957	0.008
17	0.957	0.008
18	0.956	0.008
19	0.956	0.008
20	0.956	0.008

Figure 6. Cross-validation Component Table

Goodness of fit statistics:

	Value	Validation	Cross-validation	Std. dev.(CV)
Number of observations	40	32		
Sum of weights	40	32		
R ²	0.991	0.977	0.964	0.005

Figure 7. Goodness of Fit Statistics for CCR Model

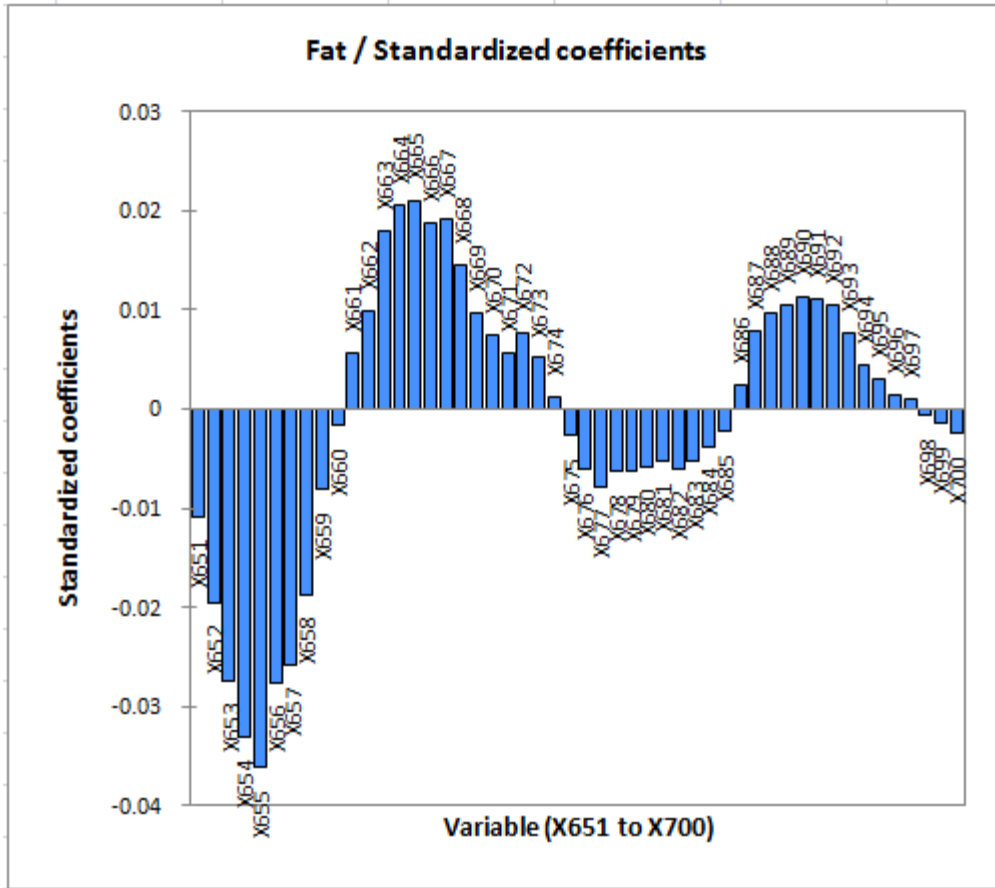


Figure 8. CCR.LM results showing standardized coefficients at highest wavelengths close to 0

Setting up a PLS (standardized) model

Re-open the CCR dialog box by click on **Modeling data / Correlated Component Regression** command in the Excel menu or click the corresponding button on the **Modeling data** toolbar.

The previous model specifications are currently displayed. In **General** tab, change the Method from CCR.LM to PLS. Then, in **Options** tab activate the '**Standardize**' option **for X / Predictors**

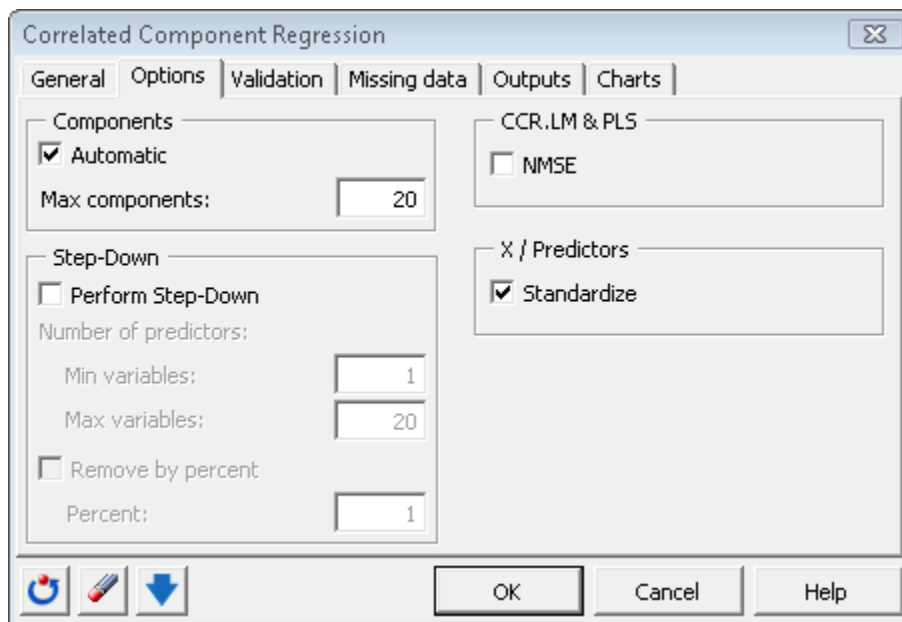


Figure 9. Options Tab

Click OK to estimate.

Interpreting PLS (standardized) Results

Cross-validation results:		
Cross-Validation component table:		
Components	R ²	Std. dev.(R ²)
1	0.237	0.040
2	0.328	0.100
3	0.670	0.027
4	0.891	0.009
5	0.920	0.013
6	0.916	0.014
7	0.934	0.011
8	0.947	0.004
9	0.945	0.006
10	0.945	0.005
11	0.950	0.005
12	0.954	0.006
13	0.954	0.006
14	0.955	0.007
15	0.954	0.006
16	0.954	0.006
17	0.954	0.006
18	0.953	0.006
19	0.953	0.006
20	0.953	0.006

Figure 10. Cross-Validation Component Table for PLS with standardized predictors

Goodness of fit statistics:				
	Value	Validation	Cross-validation	Std. dev.(CV)
Number of observations	40	32		
Sum of weights	40	32		
R ²	0.998	0.970	0.958	0.006

Figure 11. Goodness of Fit Statistics from PLS-regression with 700 standardized predictors

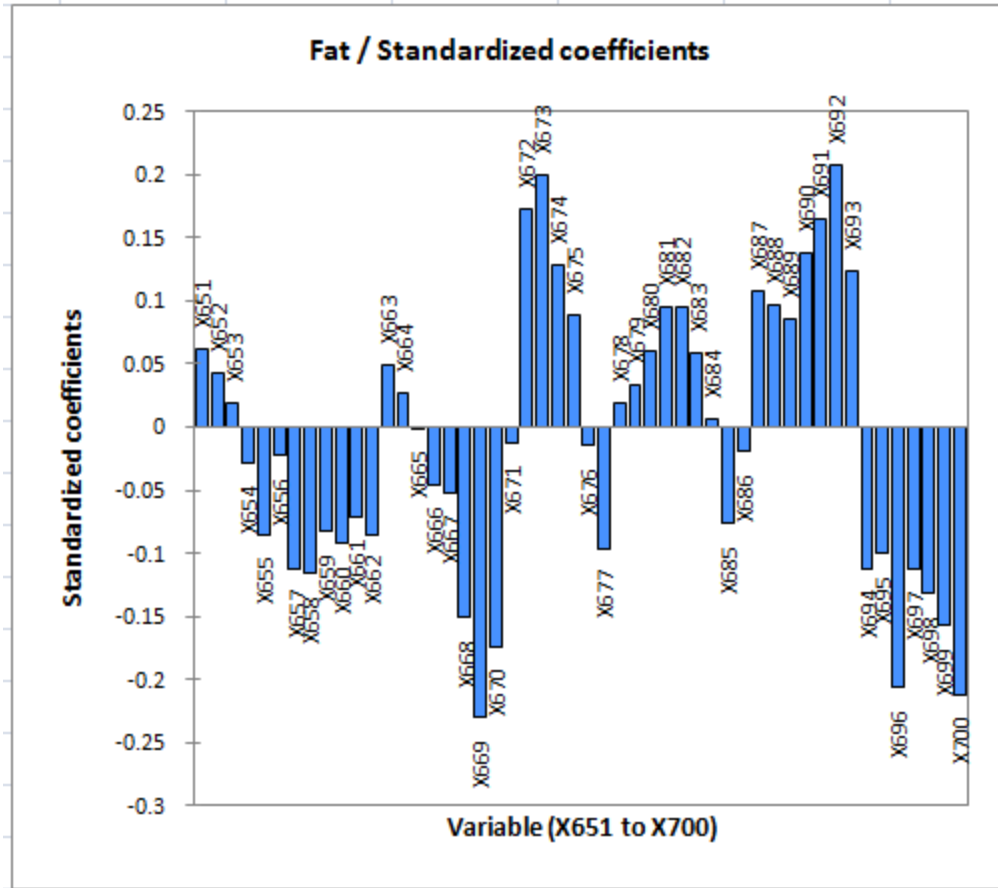


Figure 12. PLS results with 700 standardized predictors showing standardized coefficients at highest wavelengths far from 0.

Estimating and Interpreting PLS (unstandardized)

Re-open the CCR dialog box by click on **Modeling data / Correlated Component Regression** command. In **Options** tab de-activate the '**Standardize**' option **for X / Predictors**.

Click OK to estimate the model.

Cross-validation results:		
Cross-Validation component table:		
Components	R ²	Std. dev.(R ²)
1	0.260	0.041
2	0.345	0.103
3	0.736	0.027
4	0.906	0.019
5	0.916	0.016
6	0.919	0.012
7	0.930	0.010
8	0.936	0.008
9	0.932	0.008
10	0.939	0.008
11	0.942	0.008
12	0.949	0.007
13	0.950	0.007
14	0.947	0.009
15	0.946	0.009
16	0.947	0.009
17	0.948	0.009
18	0.948	0.009
19	0.948	0.009
20	0.948	0.009

Figure 13. Cross-validation Component Table for PLS regression with unstandardized predictors

Goodness of fit statistics:				
	Value	Validation	Cross-validation	Std. dev.(CV)
Number of observations	40	32		
Sum of weights	40	32		
R ²	0.997	0.975	0.952	0.007

Figure 14. Results from PLS-regression with 700 unstandardized predictors

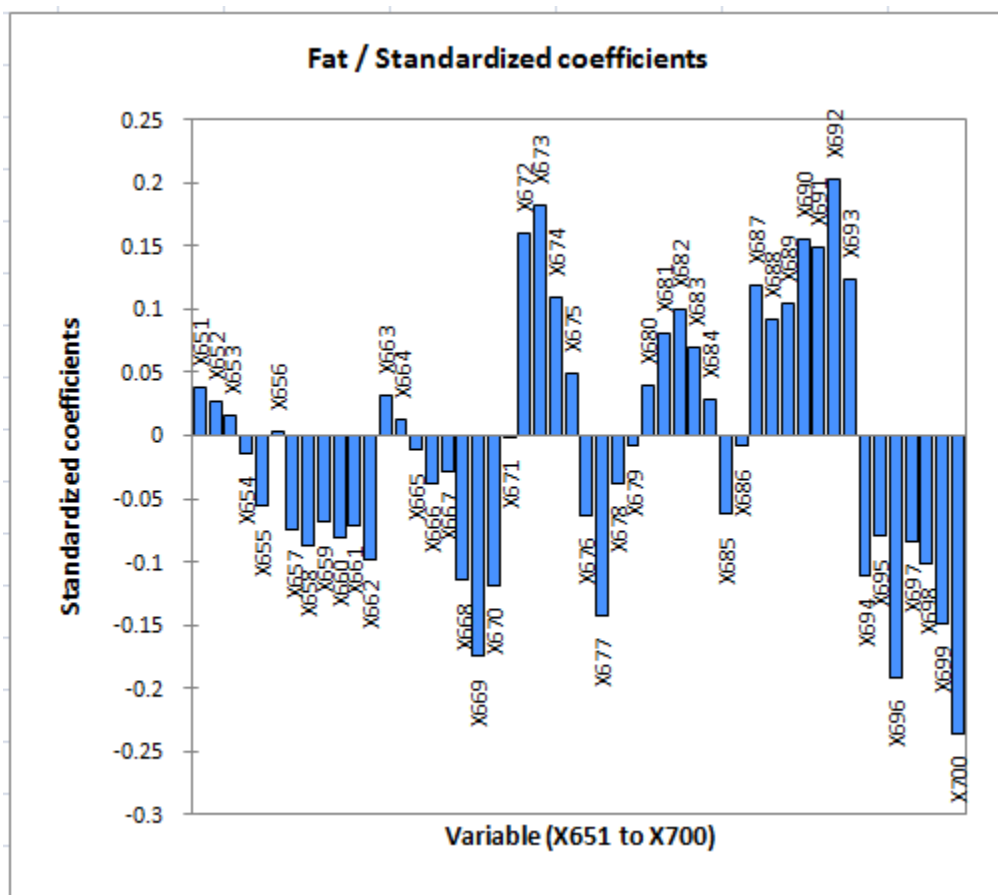


Figure 15. PLS results with 700 unstandardized predictors showing standardized coefficients at highest wavelengths far from 0.

Excluding Higher Wavelength Predictors

Unlike CCR.LM, which downplays the effects of the highest wavelengths, the PLS results (with or without standardized predictors) place relatively *large* weights on the highest wavelengths. To see that these higher wavelengths mostly contain irrelevant variation, we can re-estimate the models after excluding these higher wavelength predictors and verify that the CV and Validation performance improves.

Re-open the CCR dialog box by clicking on **Modeling data / Correlated Component Regression** command. To exclude the higher wavelengths associated with standard deviations $> .1$, in the **General** Tab change the **X / Predictors** field to include columns I through YH (Fig. 16).

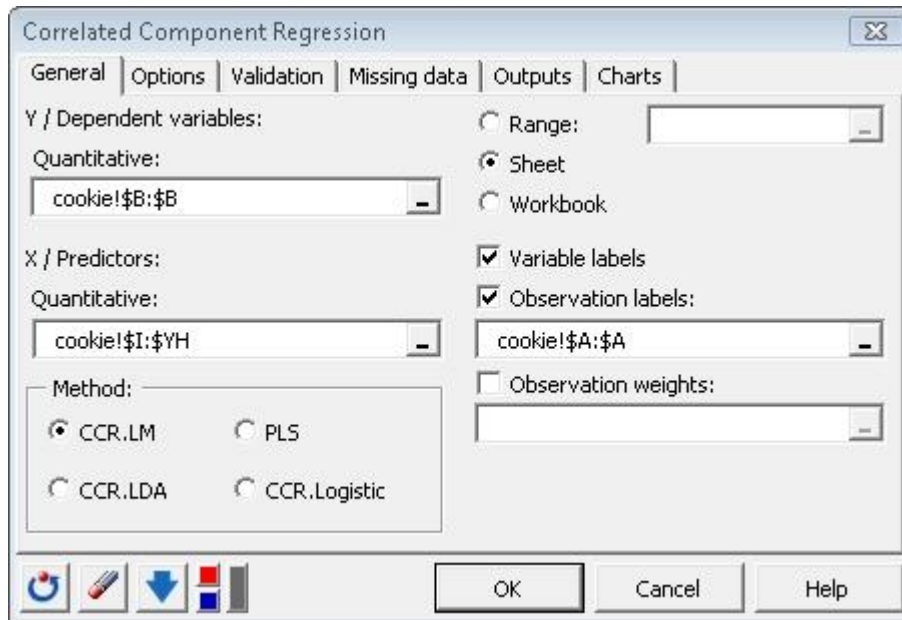


Figure 16. Excluding wavelengths 651-700

Examining the results from these models reveals that the CV and Validation performance improves after eliminating these higher wavelengths.