

Tutorial 4: Prediction with Near Infrared (NIR) Data in CORExpress®

Dataset for running CCR Linear Regression (CCR.Im)

This tutorial 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.Im) 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.

Kraemer, N. and Boulesteix, A. (2011). Penalized Partial Least Squares (PPLS). R Package.

A SPSS (.sav) file of the dataset used in this tutorial can be downloaded by clicking [here](#).

Goal of CCR for 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.

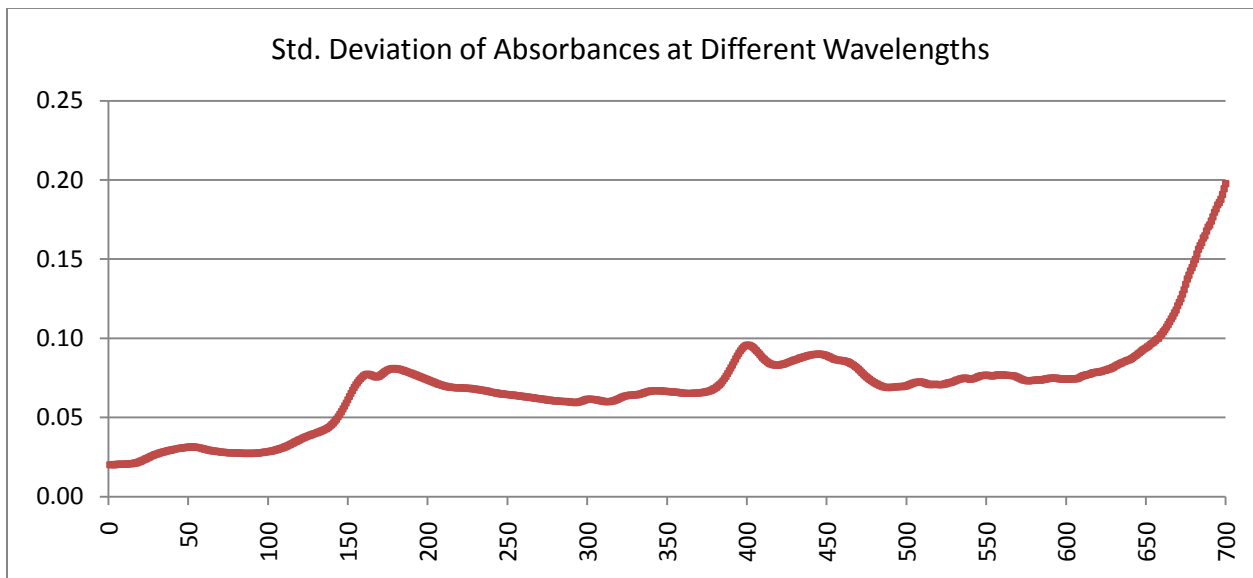


Figure 1. Plot of standard deviation for each of the 700 predictor variables (wavelengths)

The results (Table 1) 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.

Method	R^2			Std. Dev
	Training	Validation	Cross-Validation	
CCR.LM	0.9912	0.9773	0.9609	0.0059
PLS.std	0.9969	0.9764	0.9552	0.0065
PLS.unst	0.9972	0.9753	0.9477	0.0079

Table 1. Comparison of R^2 results across 3 methods

Setting up a Correlated Component Regression (CCR) model

Opening the Data File

For this example, the data file is in SPSS system file format.

To open the file, from the menus choose:

- Click File → Load Dataset...
- Select 'cookie.sav' and click Open to load the dataset

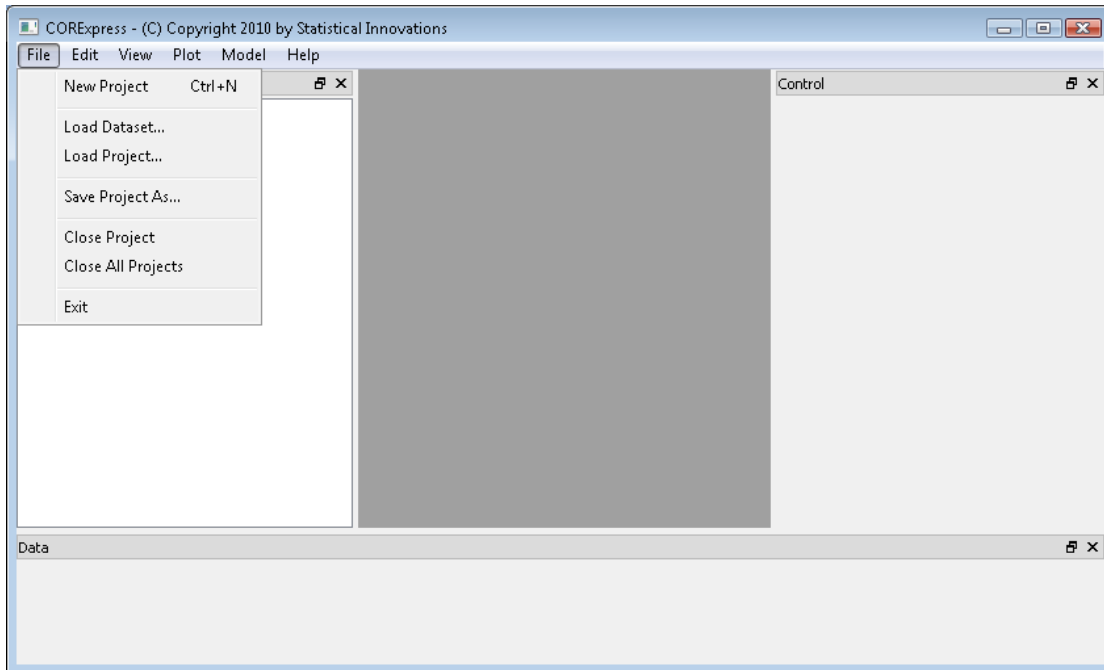


Fig. 2: File Menu

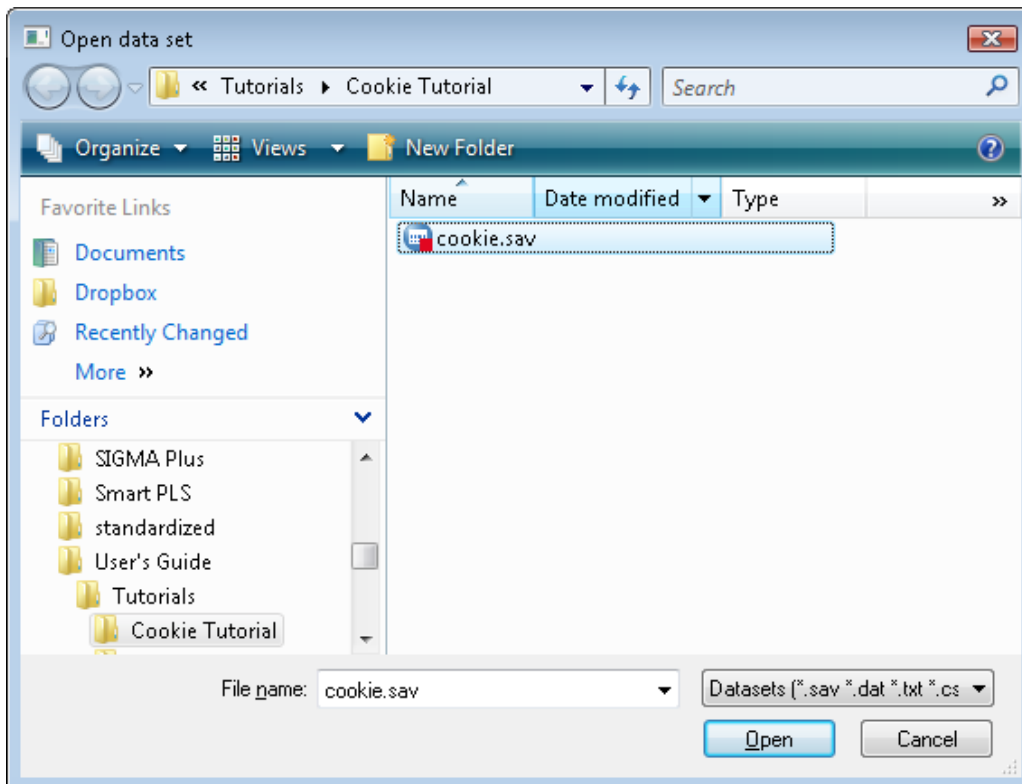


Fig. 3: Loading a Dataset

You will now see the 'cookie' dataset loaded in the 'Projects' Window on the left. In the middle (currently a dark gray box) is the workspace which will eventually show 'Model Output' windows once we have estimated CCR models. On the right is the 'Model Control' window, where models can be specified and graphs can be updated. The 'Data' Window on the bottom shows various data from the dataset.

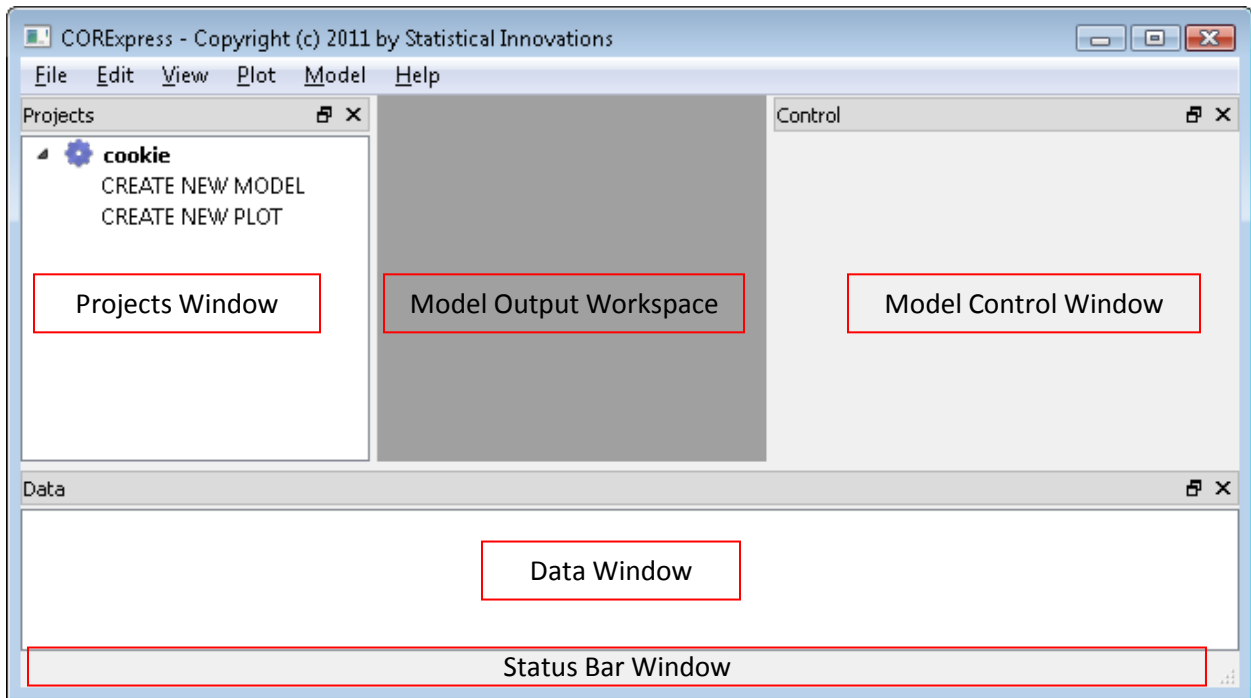


Fig. 4: CORExpress Windows

You can view the complete dataset in a new window by double clicking on 'cookie' in the Projects window. After estimating a model, the predicted scores will automatically be added to the file (and if any cases were not used to estimate the model -- validation cases -- they would also be scored).

	ID	Fat	Sucrose	Dry.Flour	Water	Validation	fold10	fold9	X001	X002	X003	X004	>
1	1	21.08	13.32	50.09	13.58	0	7	3	0.2498	0.2497	0.2494	0.249	0.
2	2	18.37	15.65	49.63	14.41	0	8	2	0.2561	0.2561	0.2564	0.2566	0.
3	3	15.35	19.06	48.63	15.04	0	8	5	0.2748	0.2747	0.2748	0.2748	0.
4	4	21.59	19.4	45.57	11.52	0	9	6	0.2437	0.2437	0.2438	0.2437	0.
5	5	17.88	15.35	50.09	14.76	0	3	9	0.2434	0.2433	0.2433	0.2434	0.
6	6	21.25	12.31	50.67	13.85	0	10	7	0.2535	0.2534	0.2534	0.2535	0.
7	7	16.19	9.95	54.61	17.32	0	1	8	0.293	0.2935	0.2938	0.2941	0.
8	8	15.85	22.77	45.86	13.59	0	2	9	0.2653	0.2654	0.2654	0.2655	0.
9	9	20.58	20.75	45.13	11.62	0	7	8	0.2441	0.2441	0.2439	0.2436	0.
10	10	18.55	15.69	49.5	14.34	0	2	4	0.2699	0.2697	0.2699	0.27	0.

Fig. 5: CORExpress Dataset View

Setting up a Correlated Component Regression (CCR) model

Opening a New Model:

- Double click on 'CREATE NEW MODEL' in the Projects window under 'cookie'

Model setup options will appear in the Control window.

Selecting the Dependent Variable:

- In the Control window below 'Dependent', click on the drop down menu and select 'Fat' as the dependent variable.

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

Selecting the Predictors:

- In the Control window below 'Predictors', click and hold on 'X001' and move the cursor down to 'X700' to highlight all 700 predictors. Click on the box next to 'X700' to select all 700 predictors.

Alternatively, you can open a Predictors Window to select the predictors:

- In the Control window below the 'Predictors' section, click the '...' button.
- The Predictors Window will open.
- Click and hold on 'X001' and move the cursor down to 'X700' to highlight all 700 predictors in the left box.
- Click on the '>>' box in the middle to select all 700 predictors and move them to the right box as candidate predictors.

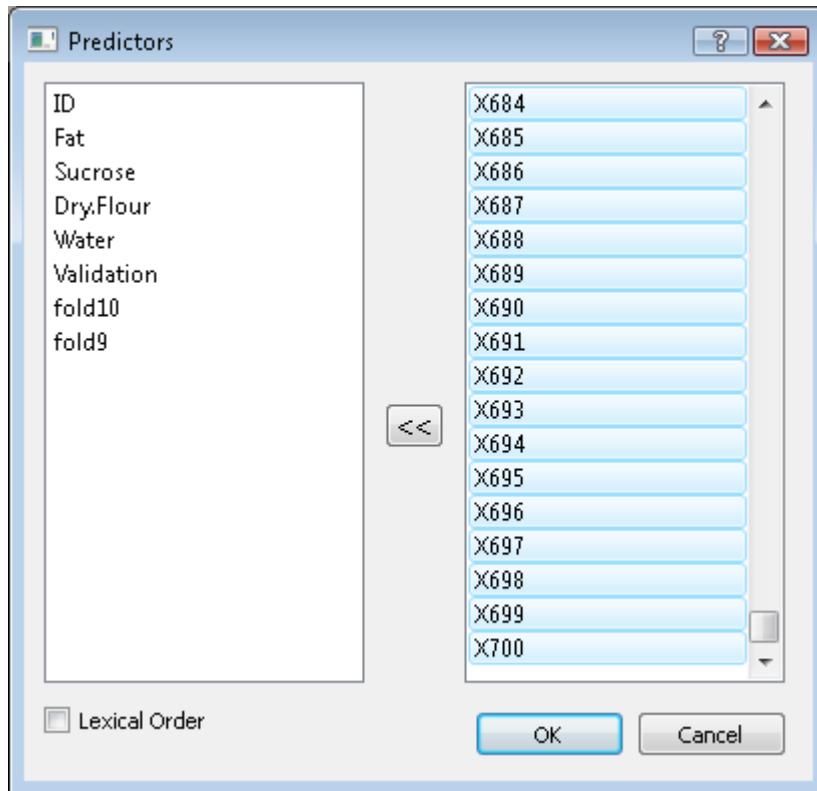


Fig. 6: Predictor Window

Selecting the Number of Components & Activating the Automatic Option:

- Under Options, click in the box to the right of '# Components', delete '4', and type '20'
- Then, check the 'Automatic' box

Selecting the Model Type:

- Click on 'CCR.lm' to select a CCR linear regression model

Your Control window should now look like this:

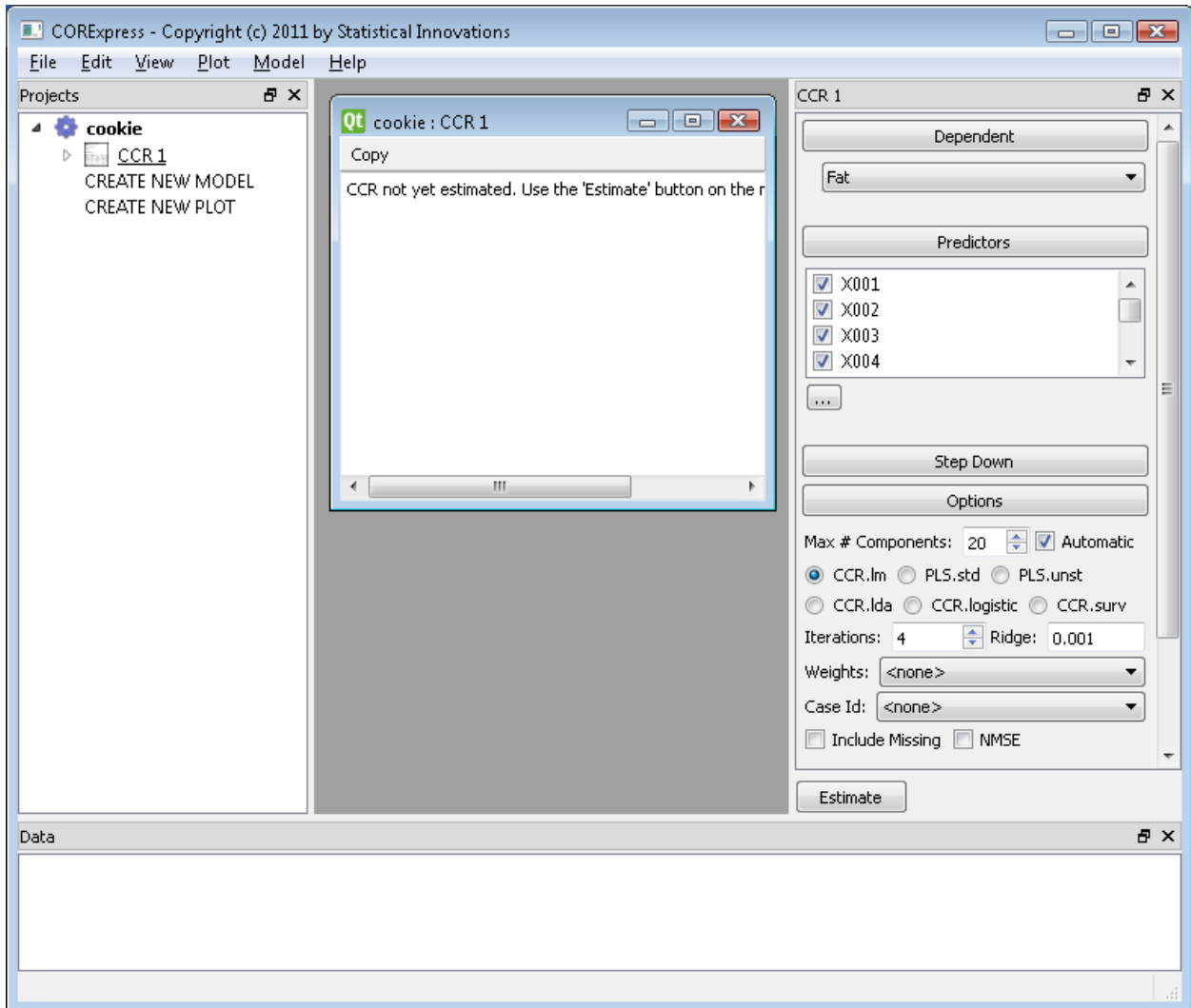


Fig. 7. Control Window

Specifying the Training Sample:

- In the Control Window, click on 'Validation' and options appear for selecting training and validation samples.

Specifying Cross Validation:

- Click on the 'Cross Validation' box and cross validation options appear.
- Click on the 'Use Cross Validation' box to enable the cross validation feature.
- In the '# rounds:' box, delete the default and type '10'
- In the '# Folds:' box, delete the default and type '5'

This divides the analysis sample into 5 subsamples (folds) that will be used to obtain values for the tuning parameters K = the number of components, and P = the number of predictors P . If a fold variable is not specified, CORExpress assigns cases randomly to each fold.

M-fold cross-validation is a common technique used in datamining. The $CV-R^2$ statistic is estimated based on model scores (predicted logits) obtained from the analysis sample after excluding a particular fold, and then applied to the fold excluded. The excluded folds are then combined and used to compute the CV statistic. Thus, the performance of the model is measured using cases not used at all in the development of the model.

Your Control window should now look like this:

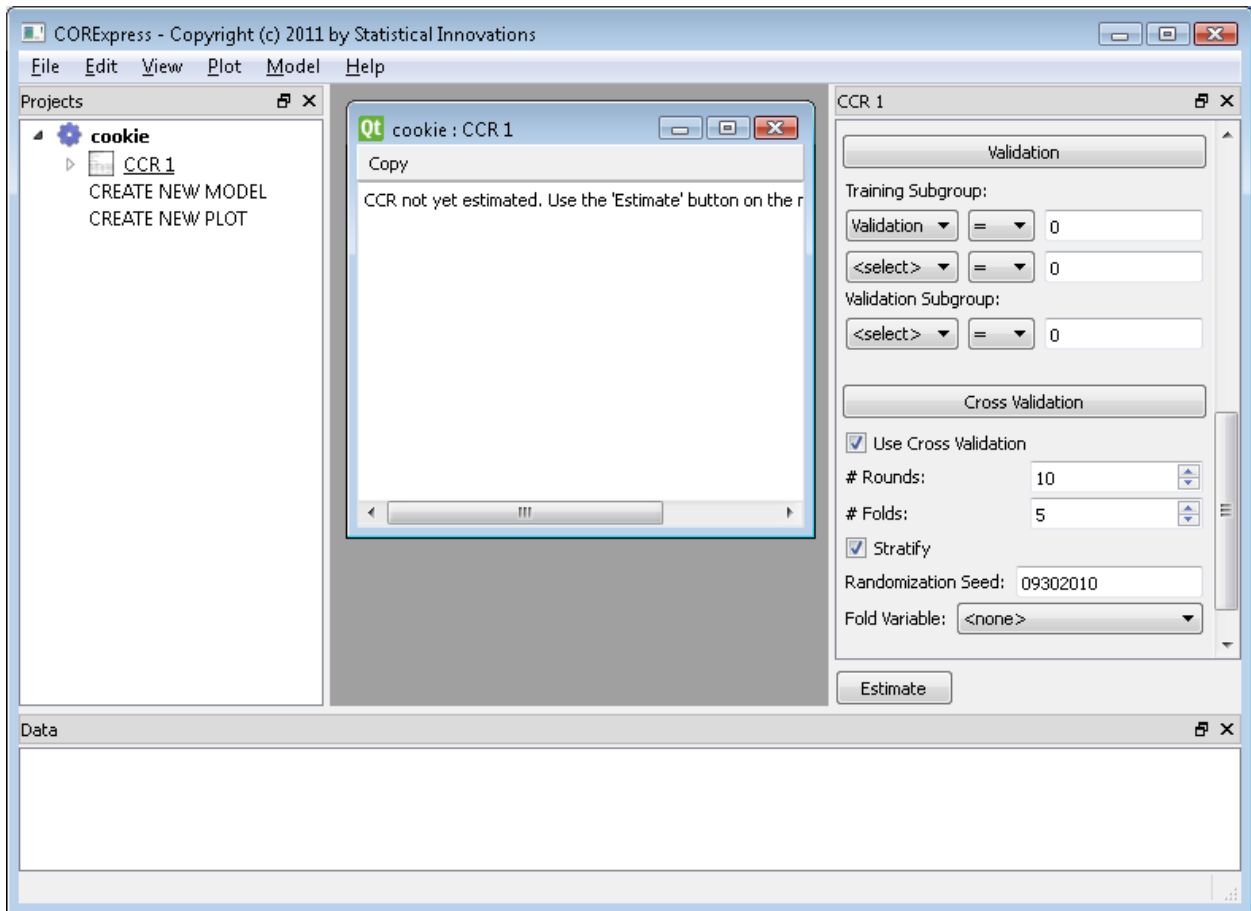


Fig. 8: Control Window

Estimate the Specified Model:

- Click on the 'Estimate' button to estimate the specified model.
- Progress of the cross-validation is reported in the status bar window at the bottom of the program window.

Note that CORExpress removed the checkmark from the Stratify CV option, which is not applicable in linear regression.

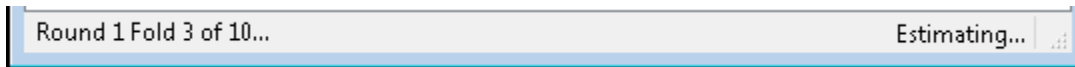


Fig. 9: CORExpress Status Bar

Interpreting CCR results

Cross-Validated Components

# Components	R²	Std. Dev
1	0.2273	0.0322
2	0.5414	0.0495
3	0.7651	0.0314
4	0.9084	0.0225
5	0.9402	0.0088
6	0.9494	0.0064
7	0.9425	0.0093
8	0.9514	0.0085
9	0.9607	0.0063
10	0.9582	0.0072
11	0.9543	0.0086
12	0.9535	0.0097
13	0.9529	0.0094
14	0.9523	0.0097
15	0.9523	0.0110
16	0.9512	0.0112
17	0.9509	0.0114
18	0.9505	0.0115
19	0.9502	0.0115
20	0.9502	0.0115

Table 2. Cross-Validation Component Table for CCR.Im

Fit

	Training	Validation	Cross-Validation	Std. Dev
R²	0.9912	0.9773	0.9609	0.0059

Table 3. Mode Fit Statistics for CCR.lm

In Excel, we plotted the standardized coefficients at the highest wavelengths.

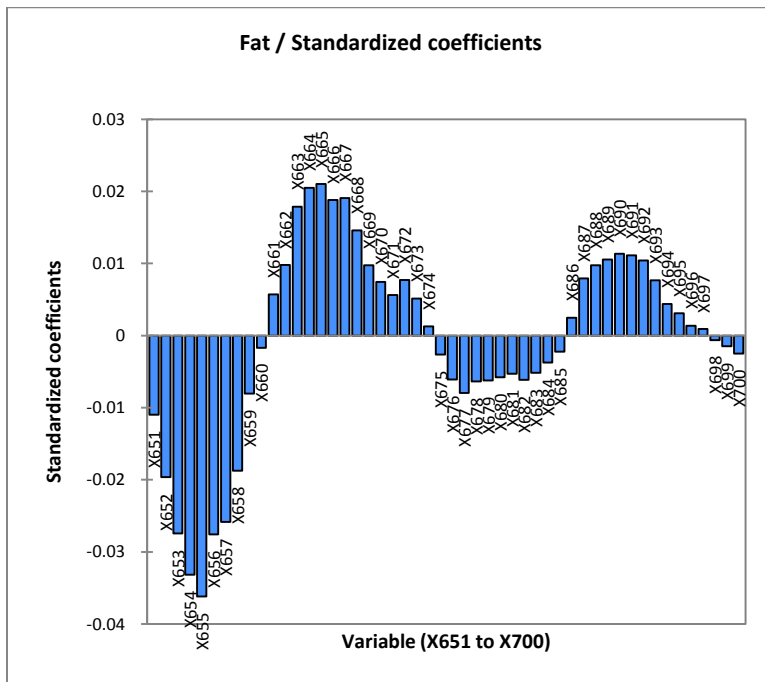
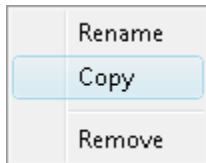


Figure 10. Plot created in Excel: CCR.LM results showing standardized coefficients at highest wavelengths close to 0

Setting up a PLS (standardized) model

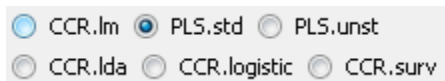
In the Projects window, right click on 'CCR 1' and select 'Copy' to create a new model with the same model specifications as 'CCR 1'.



In the Projects window, double click on 'CCR 2'.

Selecting the Model Type:

- Click on 'PLS.std' to select a PLS model with standardized predictors



Estimate the model.

Interpreting PLS (standardized) Results

Cross-Validated Components

# Components	R²	Std. Dev
1	0.2273	0.0322
2	0.3635	0.1081
3	0.7113	0.0303
4	0.8962	0.0200
5	0.9221	0.0151
6	0.9208	0.0177
7	0.9334	0.0122
8	0.9455	0.0073
9	0.9453	0.0084
10	0.9456	0.0075
11	0.9492	0.0056
12	0.9527	0.0061
13	0.9537	0.0066
14	0.9520	0.0064
15	0.9524	0.0071
16	0.9523	0.0071
17	0.9511	0.0074
18	0.9508	0.0077
19	0.9501	0.0076
20	0.9500	0.0074

Table 4. Cross-Validation Component Table for PLS with standardized predictors

Fit (PLS.std)

	Training	Validation	Cross-Validation	Std. Dev
R²	0.9969	0.9764	0.9552	0.0065

Table 5. Model Fit Statistics for PLS with standardized predictors

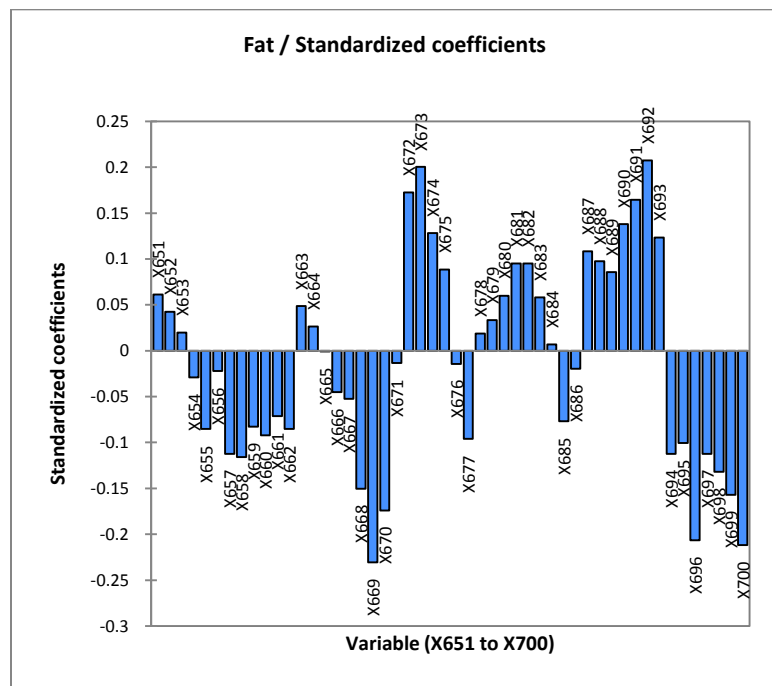
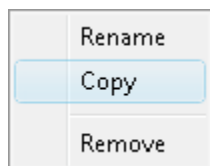


Figure 11. Plot created in Excel: PLS results with 700 standardized predictors showing standardized coefficients at highest wavelengths far from 0

Setting up a PLS (unstandardized) model

In the Projects window, right click on 'CCR 2' and select 'Copy' to create a new model with the same model specifications as 'CCR 2'.



In the Projects window, double click on 'CCR 3'.

Selecting the Model Type:

- Click on 'PLS.unst' to select a PLS model with standardized predictors

CCR.lm
 PLS.std
 PLS.unst
 CCR.lda
 CCR.logistic
 CCR.surv

Estimate the model.

Interpreting PLS (standardized) Results

Cross-Validated Components

# Components	R ²	Std. Dev
1	0.2521	0.0309
2	0.3994	0.1279
3	0.7657	0.0296
4	0.9137	0.0183
5	0.9197	0.0122
6	0.9188	0.0154
7	0.9279	0.0111
8	0.9346	0.0092
9	0.9303	0.0122
10	0.9383	0.0078
11	0.9400	0.0063
12	0.9454	0.0079
13	0.9466	0.0085
14	0.9431	0.0113
15	0.9431	0.0102
16	0.9436	0.0096
17	0.9438	0.0088
18	0.9441	0.0079
19	0.9439	0.0081
20	0.9436	0.0082

Table 6. Cross-Validation Component Table for PLS with unstandardized predictors

Fit

	Training	Validation	Cross-Validation	Std. Dev
R²	0.9972	0.9753	0.9477	0.0079

Table 7. Model Fit Statistics for PLS with unstandardized predictors

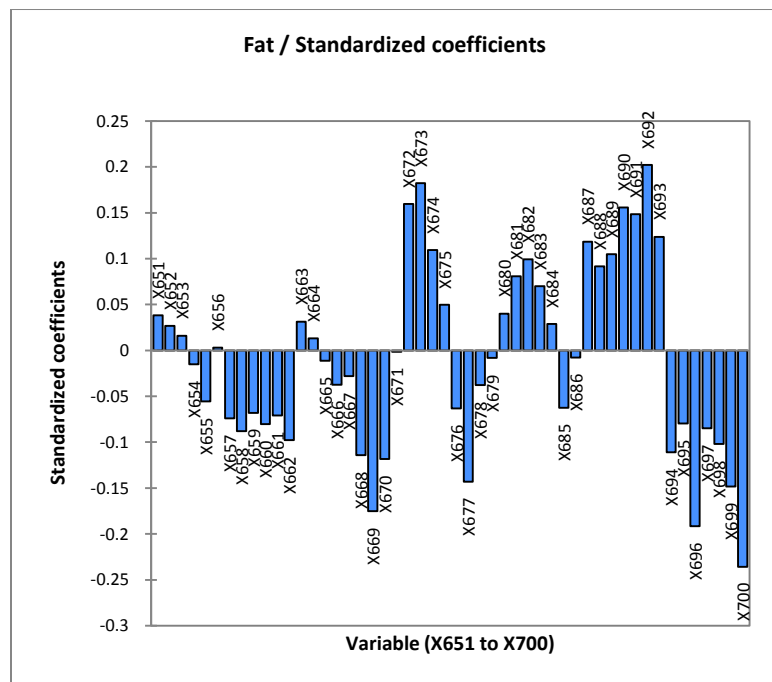
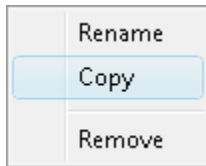


Figure 12. Plot created in Excel: PLS results with 700 unstandardized predictors showing standardized coefficients at highest wavelengths far from 0.

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.

In the Projects window, right click on 'CCR 1' and select 'Copy' to create a new model with the same model specifications as 'CCR 1'.



In the Projects window, double click on 'CCR 4'.

Selecting the Predictors:

- In the Control window below the 'Predictors' section, click the '...' button.
- The Predictors Window will open.
- Scroll down to 'X649'
- Click and hold on 'X649' and move the cursor down to 'X700' to highlight the higher wavelength predictors.
- Click on the '<<' box in the middle to remove these variables as candidate predictors.

Estimate the model.

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