Tutorial 1

Getting Started with Correlated Component Regression (CCR) in XLSTAT-CCR

Dataset for running Correlated Component Regression

This tutorial[1] is based on data provided by Michel Tenenhaus and used in Magidson (2011), "Correlated Component Regression: A Sparse Alternative to PLS Regression", 5th ESSEC-SUPELEC Statistical Workshop on PLS (Partial Least Squares) Developments.

The data consists of N=24 car models, the dependent variable PRICE = price of a car, and P=6 explanatory variables (predictors), each of which has a positive correlation with PRICE

Explanatory Variable	Correlation with PRICE
CYLINDER (engine measured in cubic centimeters)	.85
POWER (horsepower)	.89
SPEED (top speed in kilometers/hour)	.72
WEIGHT (kilograms)	.81
LENGTH (centimeters)	.75
WIDTH (centimeters)	.61

but each predictor also has a moderate correlation with the other predictor variables

Predictor	CYLINDER	POWER	SPEED	WEIGHT	LENGTH
CYLINDER	1				
POWER	.86	1			
SPEED	.69	.89	1		
WEIGHT	.90	.75	.49	1	
LENGTH	.86	.69	.53	.92	1
WIDTH	.71	.55	.36	.79	.86

An Excel sheet containing both the data and the results for use in this tutorial can be downloaded by clicking here.

[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.

Goal of CCR for this example

CCR will apply the proper amount of regularization to reduce confounding effects of high predictor correlation, thus allowing us to obtain more interpretable regression coefficients, better predictions, and include more significant predictors in a model than traditional OLS regression.

The OLS regression solution maximizes R^2 in the training sample, yielding R^2= .85. However, since this solution is based on a relatively small sample (N=24) and correlated predictors, it is likely that this model overfits the data and that .85 is an overly optimistic estimate of the true population R^2. Consistent with an overfit model, Table 1 shows that the OLS solution yields large standard errors and unrealistic ***negative*** coefficient estimates for the predictors CYLINDER, SPEED, and WIDTH.

OLS Regression	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
(Constant)	12070.41	194786.56		.06	.95
CYLINDER	**-1.94**	33.62	**-.02**	-.06	.95
POWER	1315.91	613.51	.89	2.14	.05
SPEED	**-472.51**	740.32	**-.21**	-.64	.53
WEIGHT	45.92	100.05	.18	.46	.65
LENGTH	209.65	504.15	.15	.42	.68
WIDTH	**-505.43**	1501.59	**-.07**	-.34	.74

Table 1: Results from traditional OLS regression: CV-R^2 = 0.63

Moreover, POWER is the only predictor that achieves statistical significance (p=.05) according to the traditional t-test.

CCR utilizes the cross-validated R^2 as its criterion for determining the proper amount of regularization (K) to use in a regression model. Fig.1 shows that substantial decay in CV-R^2 occurs for K>2. Thus, a substantial amount of regularization is required (K<3) to obtain a reliable result. Since OLS regression applies no regularization at all (K=P=6), this plot indicates that the OLS model definitely is overfit, and the CCR model (with K=2) should predict PRICE better than OLS regression when applied out-of-sample to new data. The results based on all 6 predictors: CV- R^2 = .75 for CCR compared to .63 for OLS regression.

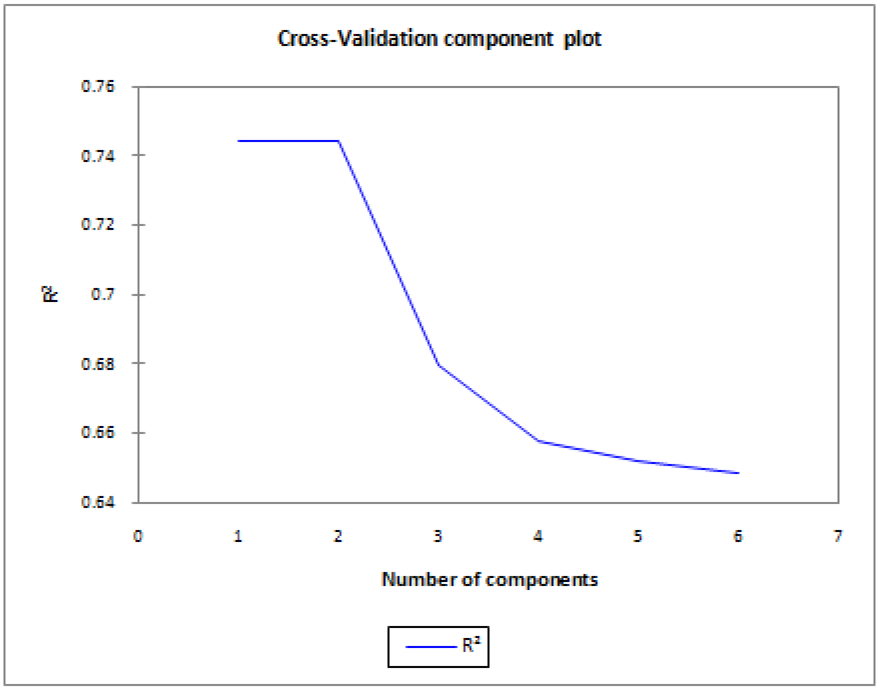


Fig. 1. Cross-Validation Component (CV-R^2) Plot showing deterioration for K>2

Also, in contrast to OLS regression which yields some negative coefficient estimates, CCR yields more reasonable ***positive*** coefficients for all 6 predictors as shown below.

Predictor	B	Beta
CYLINDER	20.9	0.19
POWER	545.5	0.37
SPEED	445.7	0.20
WEIGHT	43.4	0.17
LENGTH	32.6	0.02
WIDTH	343.6	0.05
(Constant)	-177941	

Table 2. CCR solution with K=2 components.

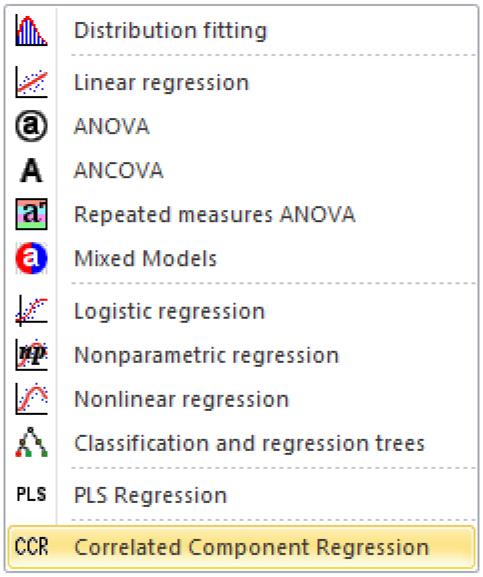
The first part of this tutorial shows how to use XLSTAT-CCR to obtain these results. The second part (see 'Activating the Step-down Algorithm') shows how to activate the CCR step-down procedure to eliminate extraneous predictors and obtain even better results (CV-R^2 = .77) as indicated in the following table.

CV- R^2 =	0.77	
Predictor	B	Beta
POWER	673.3	0.45
SPEED	222.9	0.10
WEIGHT	110.9	0.44
(Constant)	-115044	

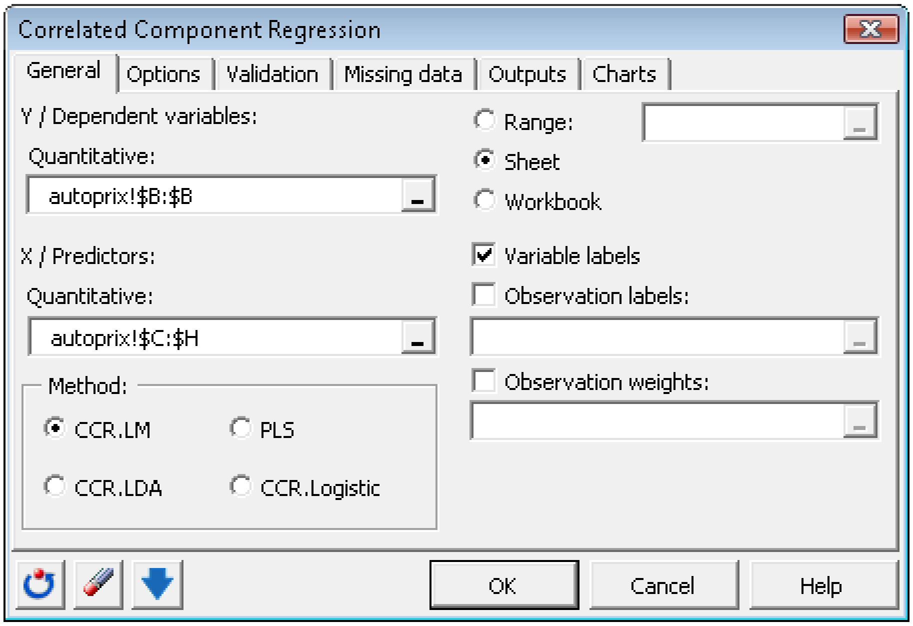
Table 3. Results from CCR with step-down algorithm

Setting up a Correlated Component Regression

To activate the Correlated Component Regression dialog box, first start XLSTAT by clicking on the button in the Excel toolbar, 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 clicked the button, the **Correlated Component Regression** dialog box is displayed with the Method=CCR.LM (linear regression model) selected by default.



Correlated Component Regression

Y / Dependent variables:

Quantitative: autoprix!$B:$B

X / Predictors:

Quantitative: autoprix!$C:$H

Method:

● CCR.LM ○ PLS

○ CCR.LDA ○ CCR.Logistic

○ Range:

● Sheet

○ Workbook

☑ Variable labels

□ Observation labels:

□ Observation weights:

Fig. 2: General Tab

In the **Y/ Dependent variables** field, use your mouse to select the variable PRICE (see the tutorial on Selecting data for more information on this topic).

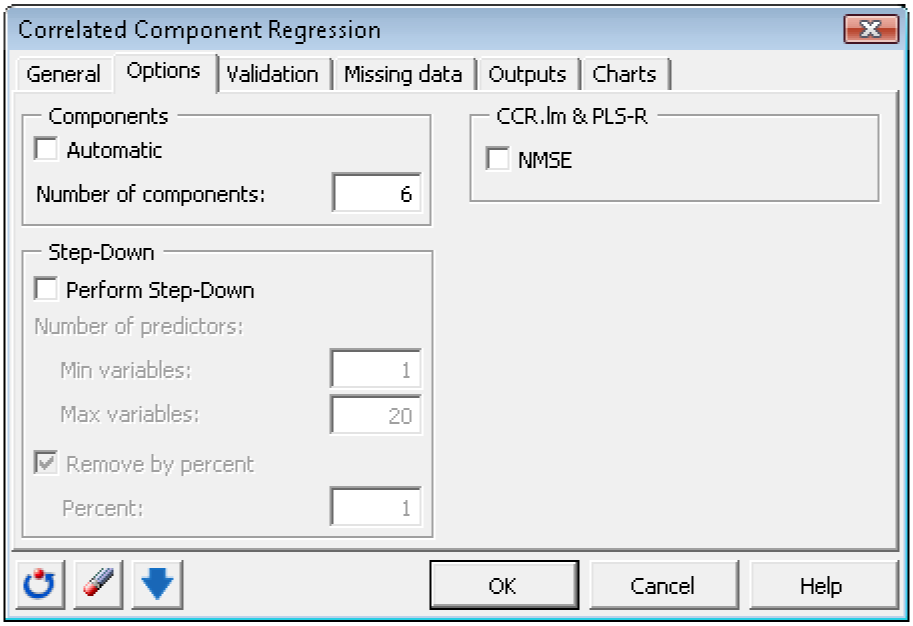
The prices are the "Ys" of the model as we want to predict these prices as a linear function of the other car attributes.

In **X/ Predictors** field, select the other 6 car attributes.

The name of the car models (MODEL) has also been selected as **Observation labels**.

To obtain the OLS regression solution, fix the number of components at 6, so it equals the number of predictors. To accomplish this, in the Options tab set Number of components to ‘6’ and uncheck ‘Automatic’.

In the **Options** tab of the dialog box, make sure that the **settings** are **as shown below.**



Correlated Component Regression

General | Options | Validation | Missing data | Outputs | Charts

Components

□ Automatic

Number of components: 6

Step-Down

□ Perform Step-Down

Number of predictors:

Min variables: 1

Max variables: 20

☑ Remove by percent

Percent: 1

CCR.lm & PLS-R

□ NMSE

Fig. 3: Options Tab

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

Interpreting CCR Model Output

Following the basic statistics output section, the coefficients (unstandardized and standardized) are presented. For example, Table 3A presents the unstandardized coefficients. Comparing Table 3A to Table 1, we see that the results match the OLS regression coefficients.

Unstandardized coefficients:

Variable	Coefficient
Intercept	12070.645
CYLINDER	-1.936
POWER	1315.907
SPEED	-472.509
WEIGHT	45.923
LENGTH	209.654
WIDTH	-505.431

Table 3A. Unstandardized coefficient estimates obtained from the 6-component (saturated) CCR model

These coefficients can be decomposed into parts associated with each of the 6 components using the component weights provided in Table 3B and the component coefficients (loadings) provided in Table 3C.

Number of components: 6

Unstandardized component Weights:

Component	Value
1	0.006
2	0.124
3	0.804
4	0.627
5	0.422
6	0.167

Table 3B. Unstandardized component weights

Unstandardized loadings:

Variable \ Component	1	2	3	4	5	6
CYLINDER	92.774	1.381	-3.728	-11.016	15.190	5.053
POWER	1320.804	728.560	1228.528	472.999	-198.632	107.935
SPEED	1642.054	239.324	-818.476	512.133	-367.926	-119.735
WEIGHT	203.058	-4.066	88.336	-37.350	-3.215	-5.772
LENGTH	1038.776	-563.277	52.095	283.260	154.288	-67.762
WIDTH	4588.211	-1915.940	-191.909	-29.443	-343.730	136.449

Table 3C. Unstandardized loadings

For example, the coefficient -1.94 for CYLINDER, can be decomposed as follows:

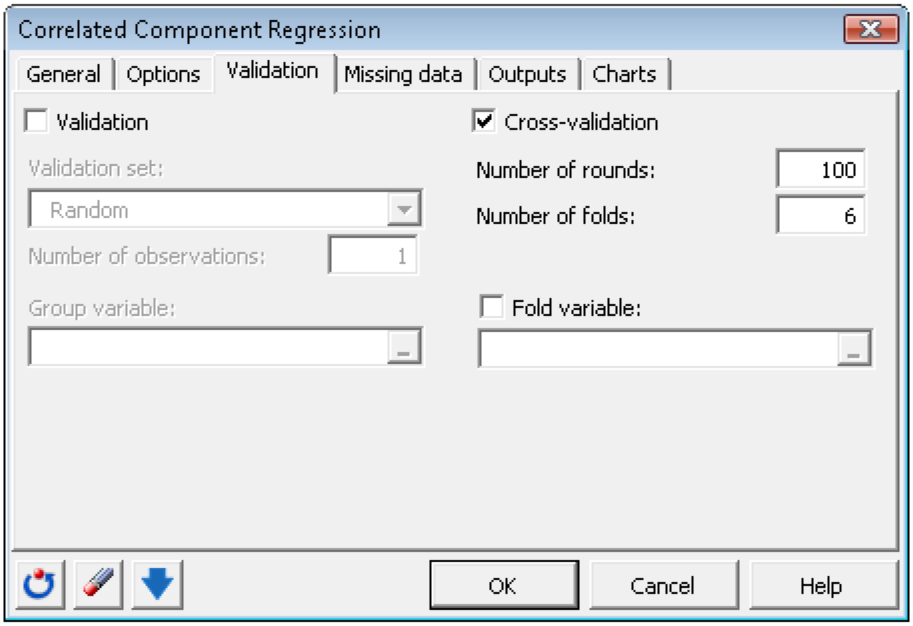
-1.94 = .006*(92.774) + .124*(1.381) + .804*(-3.728) + .627*(-11.016) + .422*(15.190) + .167*(5.053)

Activating the Automatic and M-fold Cross-validation options

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

Since N is relatively small (N=24) and the correlation between the predictors is fairly high, this *saturated* regression model overfits these data. We will now show how to activate the M-fold cross-validation (CV) option and *show* that this model is overfit, and that eliminating CCR components 3-6 provides the proper amount of regularization to produce more reliable results. To allow CV to assess all possible degrees of regularization, we will estimate all 6 CCR models (K≤6). We do this by activating the **Automatic** option in the **Options** tab.

The number of folds M is generally taken to be between 5 and 10, so we select M=6, the only integer between 5 and 10 that divides evenly into 24. In the **Validation** tab we activate 'Cross-validation' and request 100 rounds of 6-folds. By requesting more than 1 round, we obtain a standard error for the CV-R^2.



☐ Validation

Validation set: Random

Number of observations: 1

Group variable:

☑ Cross-validation

Number of rounds: 100

Number of folds: 6

☐ Fold variable:

Fig. 4: Validation Tab

Note that activating the ‘Automatic’ option also requests the Cross-Validation Component Plot to be generated (this is checked in the **Charts** tab) shown earlier in Fig. 1.

Click OK to perform these analyses. The Goodness of Fit Statistics show that the resulting model has K=2 components. For this model, the CV-R^2 increases to .750 with a standard error of only .014, providing a significant improvement over the OLS regression CV-R^2 =.64.

Unstandardized coefficients:

Variable	Coefficient
Intercept	-177941.121
CYLINDER	20.944
POWER	545.463
SPEED	445.654
WEIGHT	43.368
LENGTH	32.618
WIDTH	343.616

Table 4A. Coefficients obtained from the 2-component model

Number of components: 2

Unstandardized component Weights:

Component	Value
1	0.221
2	0.349

Table 4B. Component weights obtained from the 2-component model

Unstandardized loadings:

Variable \ Component	1	2
CYLINDER	92.774	1.381
POWER	1320.804	728.560
SPEED	1642.054	239.324
WEIGHT	203.058	-4.066
LENGTH	1038.776	-563.277
WIDTH	4588.211	-1915.940

Table 4C. Loadings obtained from the 2-component model

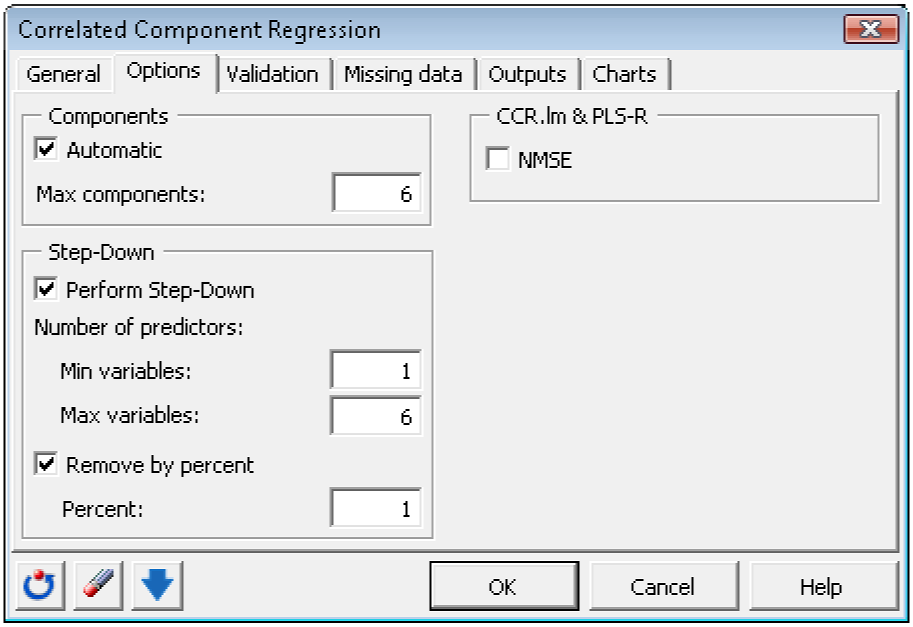
From the Coefficients Output in Tables 4A, 4B and 4C we see how the coefficients are now constructed based on only 2 components. For example, the coefficient for CYLINDER can be decomposed as follows:

20.944 = .221*92.774 + .349*1.381

Activating the Step-down Algorithm

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

To eliminate extraneous and weak predictors, in the options tab we will now activate the step-down algorithm as shown below:



Correlated Component Regression

Components

☑ Automatic

Max components: 6

Step-Down

☑ Perform Step-Down

Number of predictors:

Min variables: 1

Max variables: 6

☑ Remove by percent

Percent: 1

CCR.lm & PLS-R

□ NMSE

Figure 5. Options Tab

Activation of the step-down option automatically requests the step-down predictor selection plot in the Charts tab and the Predictor Count table from the Output tab.

Click on **OK** to estimate.

The predictor selection plot suggests that inclusion of 3 predictors in the model is optimal.

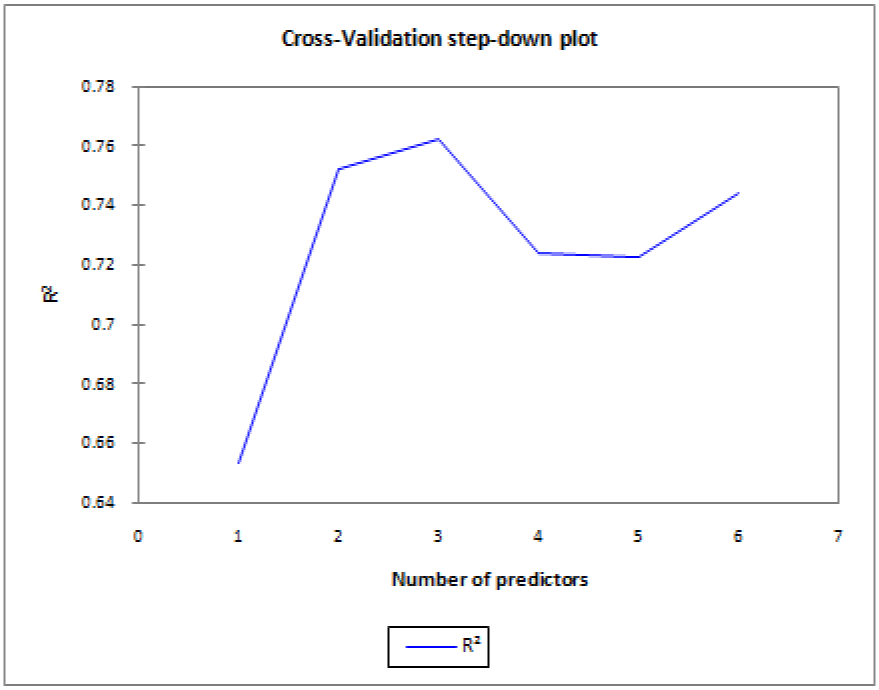


Figure 6. Cross-validation Step-down Plot

The Cross-validation Predictor Count table suggests that POWER and WEIGHT are the most important predictors, being included in 600 and 584 of the 1800 cross-validated regressions.

Cross-validation predictor count table:

Predictor	Round_1	Round_2	Round_3	Round_4	Round_5	…	Round_100	Total
POWER	6	6	6	6	6	…	6	600
WEIGHT	6	6	6	6	6	…	6	584
SPEED	1	2	2	3	0	…	4	260
CYLINDER	4	4	3	3	0	…	2	242
LENGTH	0	0	1	0	0	…	0	69
WIDTH	1	0	0	0	0	…	0	45
Total	18	18	18	18	12	…	18	1800

The final model has CV-R^2 = .77 and includes the predictors POWER, SPEED and WEIGHT:

Goodness of fit statistics:

	Value	Cross-validation	Std. dev.(CV)
Number of observations	24		
Sum of weights	24		
R^2	0.836	0.766	0.021

Predictors retained in the model:

POWER
SPEED
WEIGHT

General Discussion and Additional Tutorials

Key driver regression attempts to ascertain the importance of several key explanatory variables (predictors) X_1, X_2, … , X_P that influence a dependent variable. For example, a typical dependent variable in key driver regression is "Customer Satisfaction". Traditional OLS regression methods have difficulty with such *derived importance* tasks because the predictors usually have moderate to high correlation with each other, resulting in problems of confounding, making parameter estimates unstable and thus unusable as measures of importance.

Correlated Component Regression (CCR) is designed to handle such problems, and as shown in Tutorial 2 it even works with high-dimensional data where there are more predictors than cases! Parameter estimates become more interpretable and cross-validation is used to avoid over-fitting, thus producing better out-of-sample predictions.

For CCR **Tutorial 2** click here: **Using Correlated Component Regression with a Dichotomous Y and Many Correlated Predictors**

For CCR **Tutorial 3** click here: **Obtaining Predictions from a 2-class Regression**

For other XLSTAT Tutorials click here

Copyright ©2011 Statistical Innovations Inc. All rights reserved.