Getting Started with Correlated Component Regression (CCR) in COREXpress®

Dataset for running CCR Linear Regression (CCR.lm)

This tutorial 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 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

Table 1.

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

Table 2.

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

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.

As shown in Table 3 below, traditional OLS regression 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 3: 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's Cross-Validation Component (CV-R^2) Plot shows that substantial decay in the cross-validated 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=6), this plot indicates that the CCR model (with K=2) should predict PRICE better than traditional OLS regression when applied out-of-sample to new data (results based on all 6 predictors: CV- R^2 = .75 for CCR vs. CV- R^2 = .63 for OLS regression).

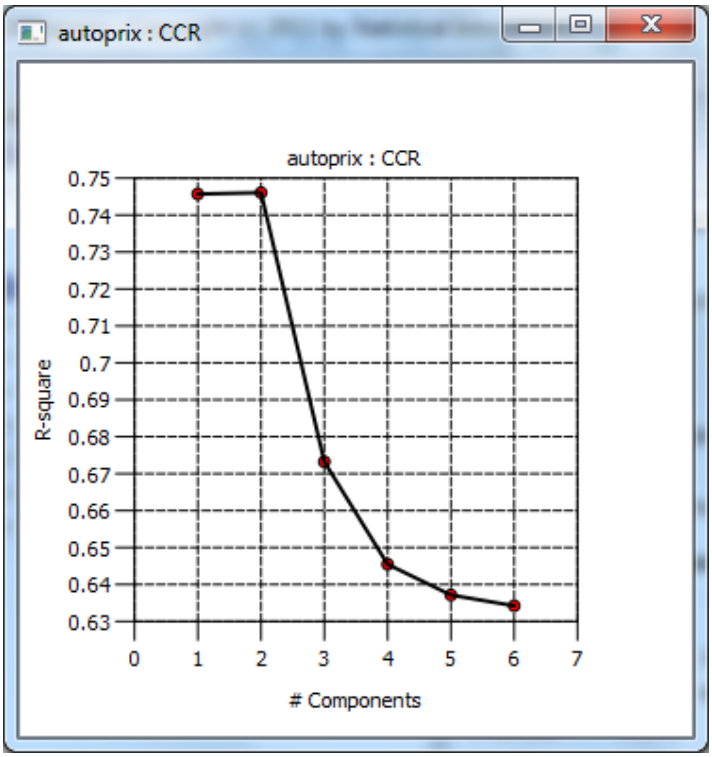


Fig. 1. Cross-Validation Component (CV-R^2) Plot

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 4. CCR solution with K=2 components.

Part A of this tutorial shows how to use COREXpress to obtain these results. Part B shows how to activate the CCR step-down procedure to eliminate extraneous predictors and obtain even better results 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 5. Results from CCR with step-down algorithm

Part A: Setting up a Correlated Component Regression

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 'autoprice.sav' and click Open to load the dataset

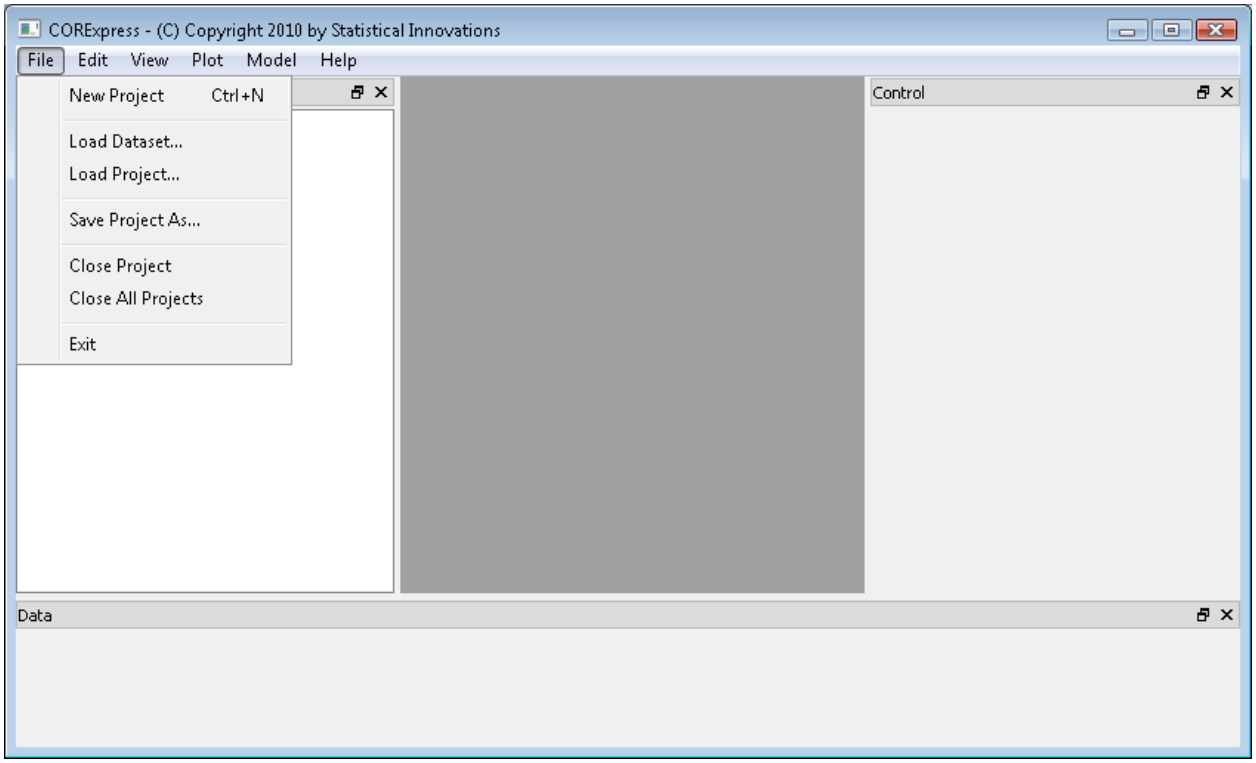


Fig. 2: File Menu

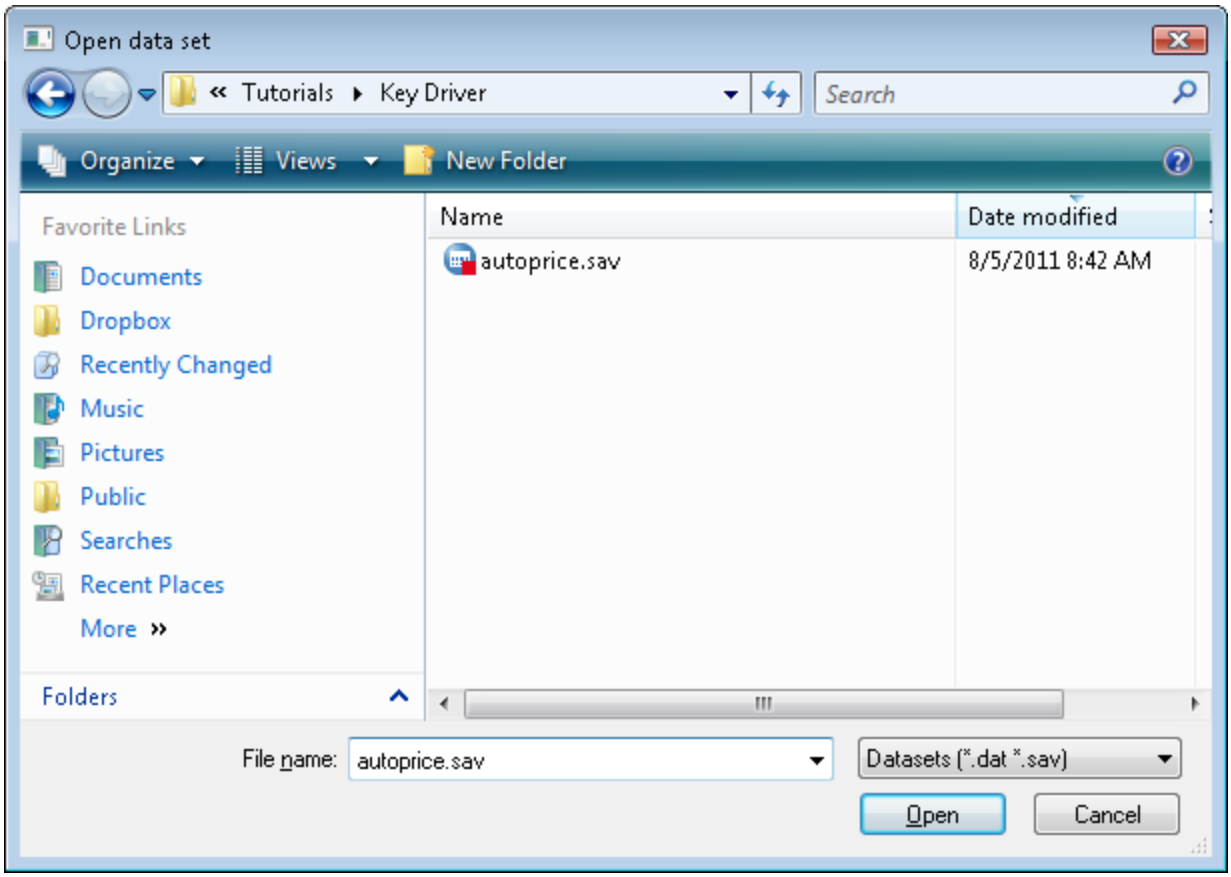


Fig. 3: Loading a Dataset

You will now see the “autoprice” 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.

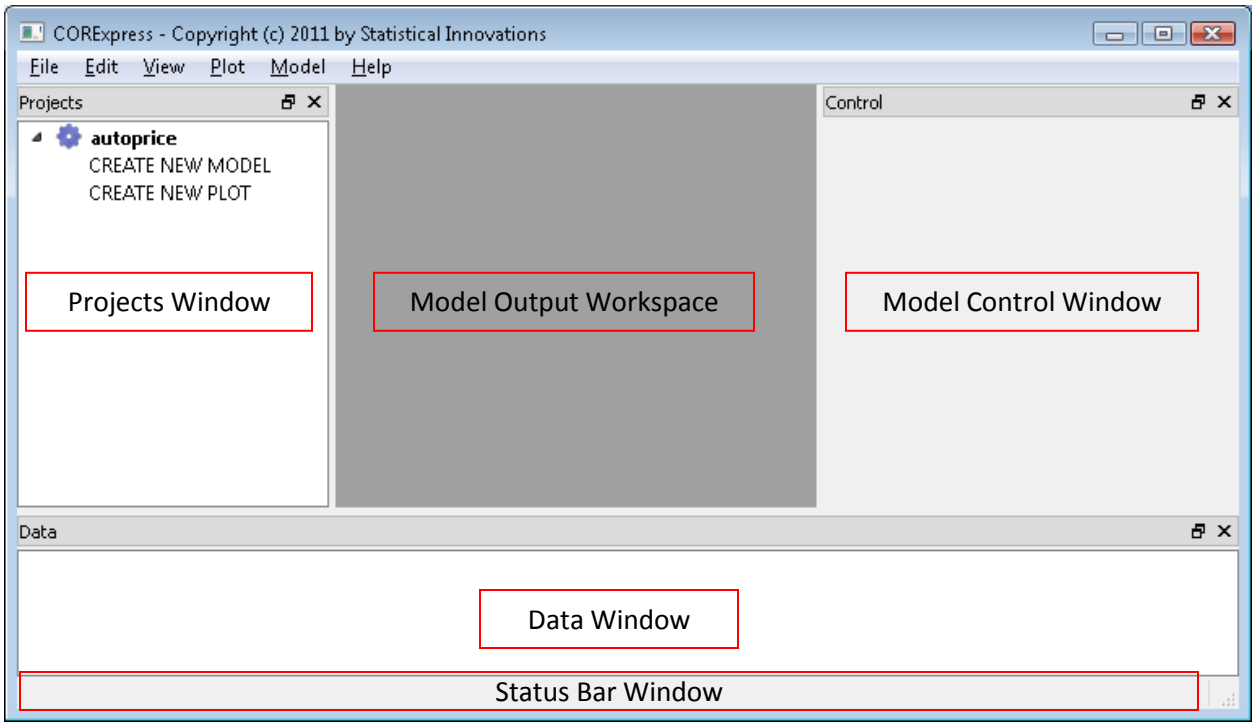
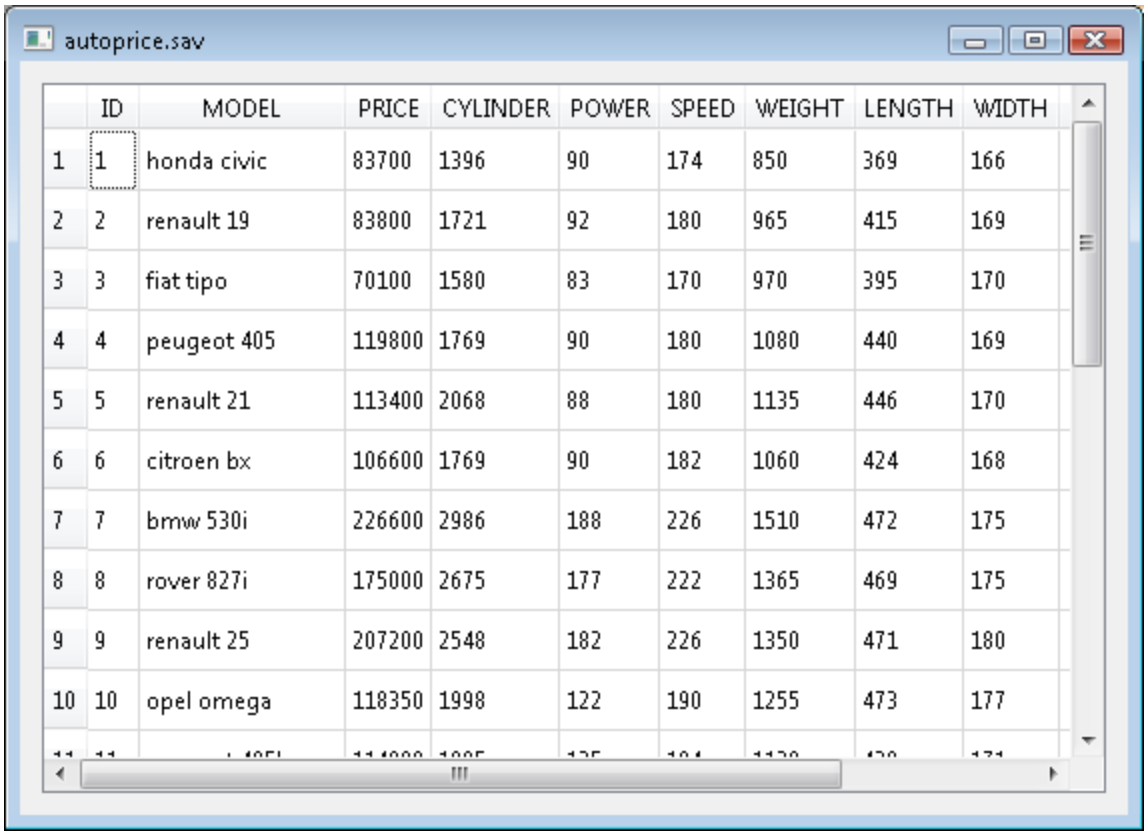


Fig. 4: COREXpress Windows

You can view the complete dataset in a new window by double clicking on “autoprice” 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	MODEL	PRICE	CYLINDER	POWER	SPEED	WEIGHT	LENGTH	WIDTH
1	1	honda civic	83700	1396	90	174	850	369	166
2	2	renault 19	83800	1721	92	180	965	415	169
3	3	fiat tipo	70100	1580	83	170	970	395	170
4	4	peugeot 405	119800	1769	90	180	1080	440	169
5	5	renault 21	113400	2068	88	180	1135	446	170
6	6	citroen bx	106600	1769	90	182	1060	424	168
7	7	bmw 530i	226600	2986	188	226	1510	472	175
8	8	rover 827i	175000	2675	177	222	1365	469	175
9	9	renault 25	207200	2548	182	226	1350	471	180
10	10	opel omega	118350	1998	122	190	1255	473	177

Fig. 5: COExpress Dataset View

Step 1: Determining the Optimal Number of Components

Selecting the Type of Model:

- Double click on “CREATE NEW MODEL” in the Workspace window under “autoprice”

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 “PRICE” as the dependent variable.

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

Selecting the Predictors:

- In the Control window below “Predictors”, click and hold on “CYLINDER” and move the cursor down to “WIDTH” to highlight all 6 predictors. Click on the box next to “WIDTH” to select all 6 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 “CYLINDER” and move the cursor down to “WIDTH” to highlight all 6 predictors in the left box.
- Click on the “>>” box in the middle to select all 6 predictors and move them to the right box as candidate predictors.

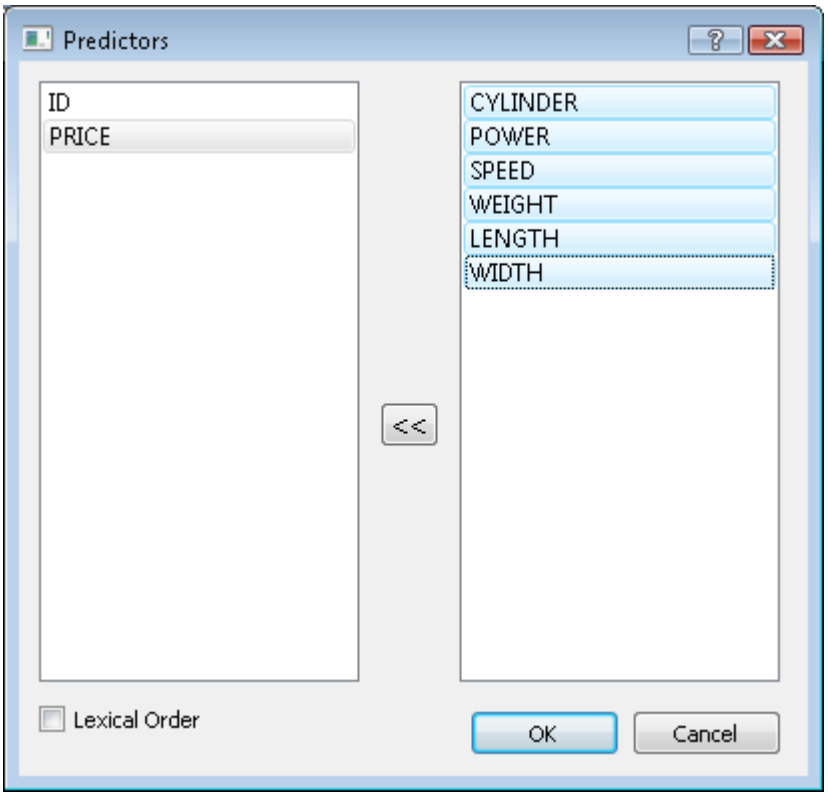


Fig. 6: Predictor Window

To obtain the OLS regression solution, fix the number of components at 6, so it equals the number of predictors.

Selecting the Number of Components:

- Under Options, click in the box to the right of “# Components”, delete “4”, and type “6”

Selecting the Model Type:

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

Selecting the Case ID:

- Click on the Case ID drop down menu and select “ID” to select the name of the car models as case ids.

Your Control window should now look like this:

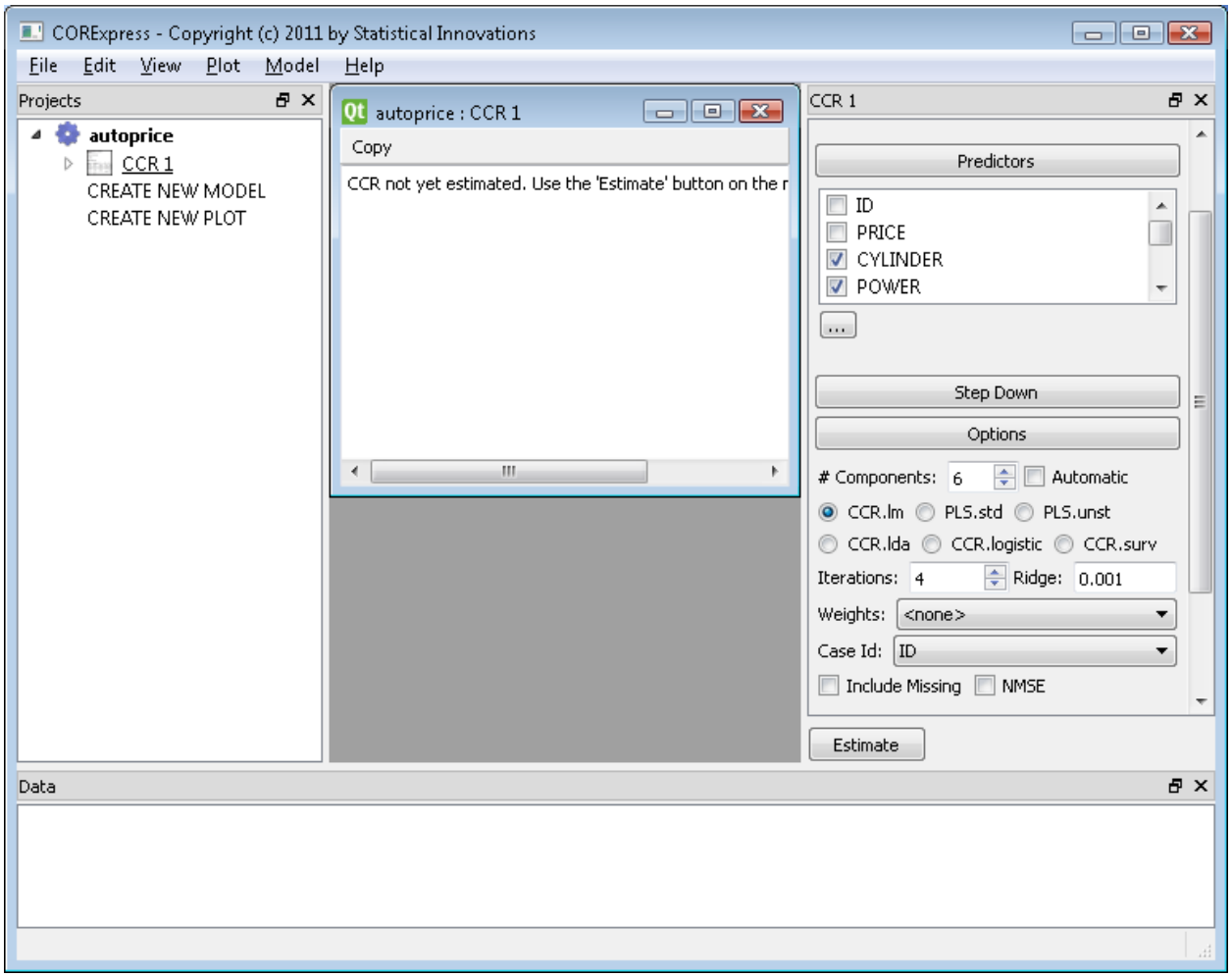


Fig. 7: Control Window

Estimate the Specified Model:

- Click on the “Estimate” button to estimate the specified model.

Interpreting CCR Model Output

Following the basic statistics output section, the coefficients (unstandardized and standardized) are presented. In addition to the standard OLS regression coefficients, the right-most columns of the output contain loadings for each predictor on each of the K=6 components (CC1, CC2, … , CC6) as well as the component weights for the components.

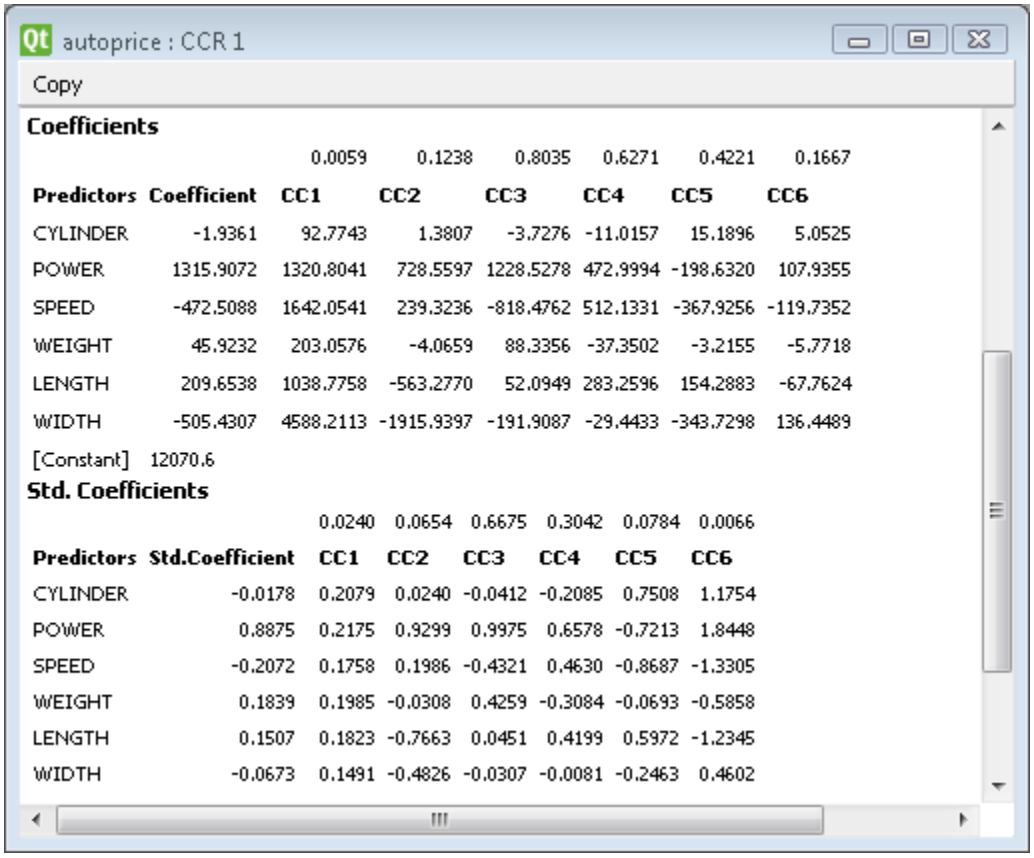


Fig. 8. Coefficient estimates obtained from the 6-component (saturated) CCR model

Comparing Figure 8 to Table 3, we see that the results match the OLS regression coefficients. These coefficients can be decomposed into parts associated with each of the 6 components using the component weights provided (numbers above CC1, CC2, …, CC6) and the component coefficients (loadings) provided below CC1, CC2, …, CC6.

For example, the coefficient -1.9361 for CYLINDER, is decomposed as follows:

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

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 Model Control Window.

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

Activating the Automatic Option:

- Under Options, check the "Automatic" box

Note that activating the 'Automatic' option also requests the Cross-Validation Component Plot to be generated shown earlier in Fig. 1.

Specifying Cross Validation:

- In the Model Control Window, click on the "Cross Validation" box and cross validation options will appear.
- Click on the "Use Cross Validation" box to enable the cross validation feature.
- In the "# Rounds:" box, type "10"
- In the "# Folds:" box, type "6"
- Keep the "<none>" in the Fold Variable drop down drop down menu

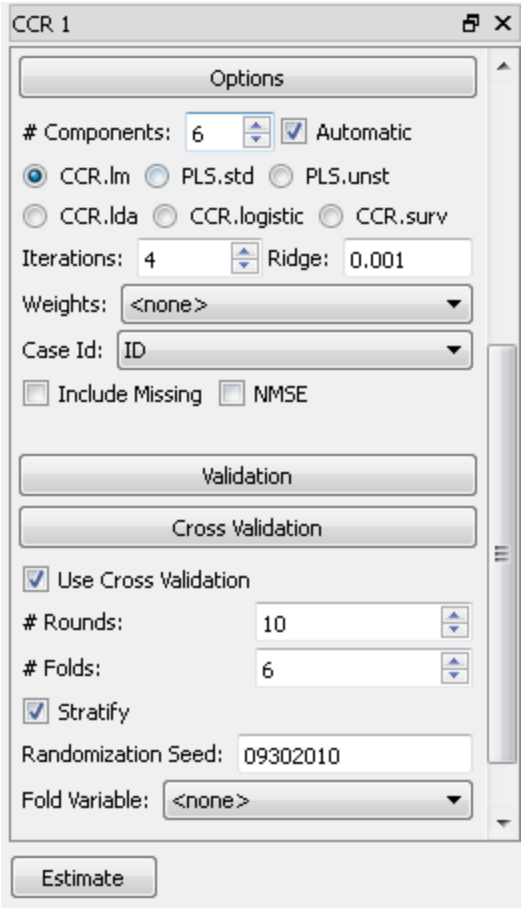


Fig. 9. Model Control Window with Automatic activated and Cross-validation specified

Estimate the Specified Model:

- Click on the "Estimate" button to estimate the specified model.

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

The Summary Statistics show that the resulting model has K=2 components. For this model, the CV-R^2 increases to .75 with a standard error of only .02, providing a significant improvement over the OLS regression CV-R^2 =.63.

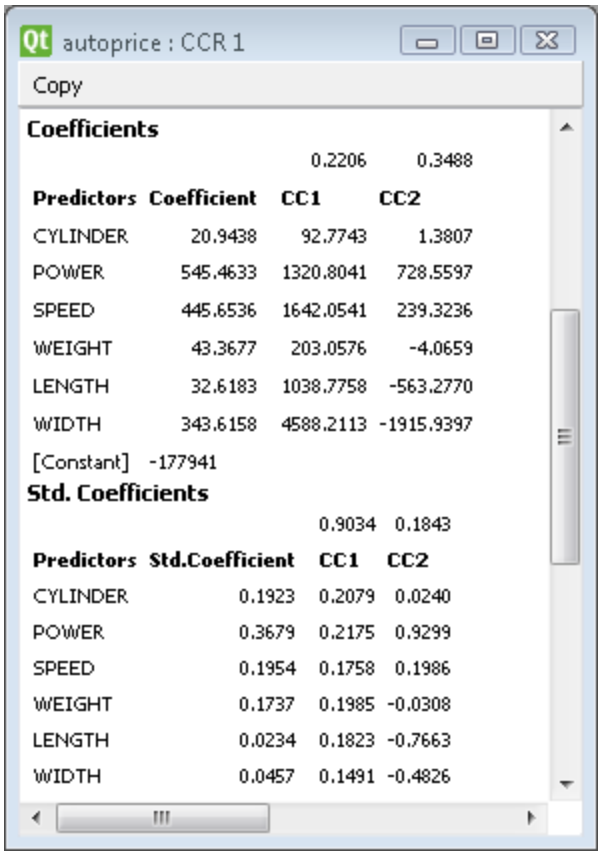


Fig. 10. Model Output

From the Coefficients Output in Figure 10 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

Part B: Activating the Step-down Algorithm

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

Specifying the Number of Predictors to Step Down:

- In the Model Control Window, click on the "Step Down" box and step down options will appear.
- Click on the "Perform Step Down" box to enable the step down feature.
- Keep the default values for "Min # Predictors" and "Max # Predictors".

Activation of the step-down option automatically requests the step-down predictor selection and the Predictor Count table.

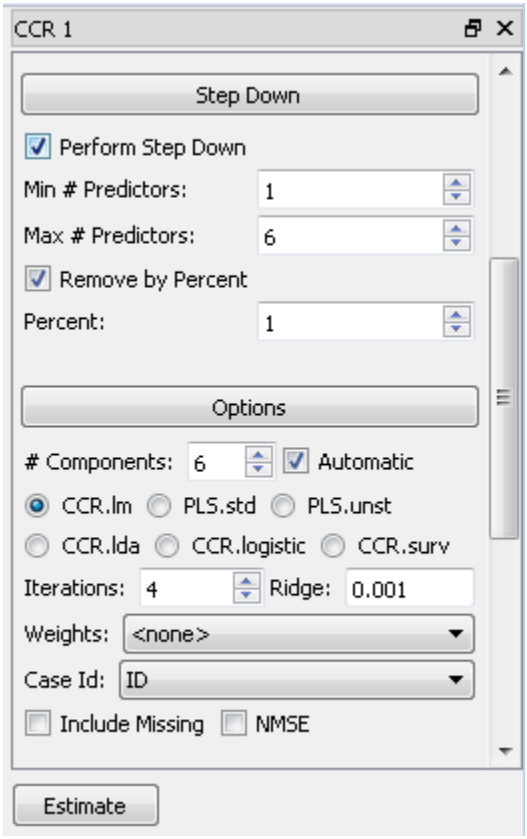


Fig. 11. Model Control Window with Step-down options specified

Estimate the Specified Model:

- Click on the "Estimate" button to estimate the specified model.

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

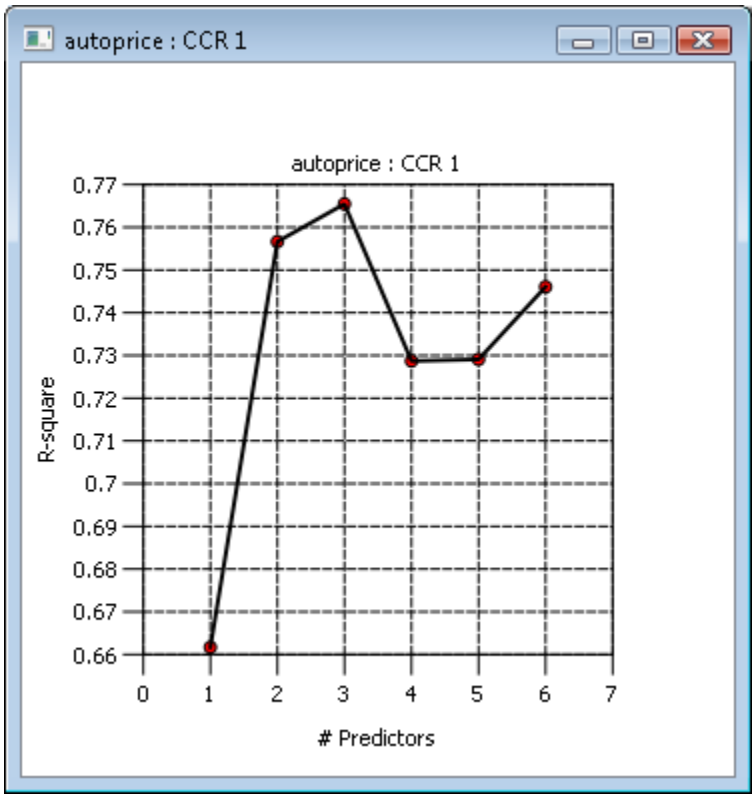


Fig. 12.

The Predictor Count table suggests that POWER and WEIGHT are the most important predictors, being included in 60 and 59 of the 186 cross-validated regressions respectively. Also, we see that among the 10 rounds, P*=3 predictors was obtained as the optimal number 7 of the 10 times.

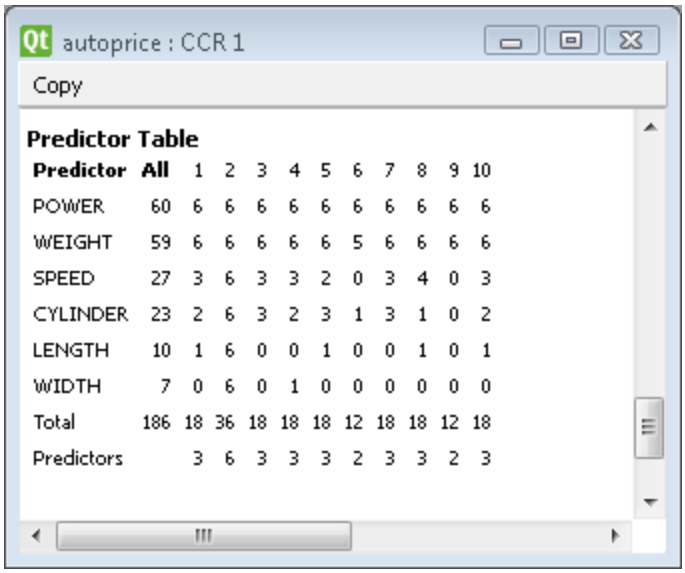


Fig. 13.

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

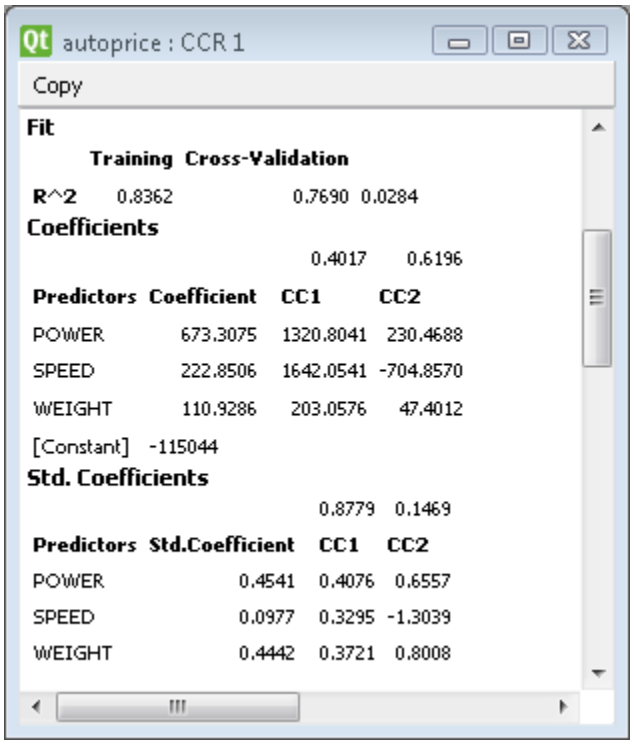


Fig. 14

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.