Tutorial 2: Correlated Component Regression for a Continuous Dependent Variable

Goal

This tutorial shows how to use COREXpress to perform *regularized* linear regression (CCR-lm) with demo data set #1, consisting of data simulated according to the usual linear regression assumptions with true coefficients shown in Table 1A below. The tutorial consists of 2 parts – Part A, which illustrates the use of COREXpress with P=56 predictors with a large sample of size N=5,000, and Part B, which illustrates the use of COREXpress on *high-dimensional data* with the same number of predictors, P=56, but with a sample of size N=50, so that P>N. Some kind of regularization is needed in order to get reliable predictions based on *high-dimensional* data. The tutorial begins on page 5. Pages 1-4 provide some background information and a summary of results from some alternative regression approaches.

Overview

Tutorial Part A shows that the Correlated Component Regression (CCR) method, as implemented in COREXpress, outperforms stepwise regression (regardless of whether the forward or backward option is used), and also outperforms the penalized regression method lasso, an alternative regularized regression approach, based on a relatively large sample size with N=5,000 (the 'training' data). For these data, all but 14 of the P=56 predictors are either extraneous or completely irrelevant (i.e., only 14 predictors have true non-zero population coefficients). Additional cases from an independent data set of equal size (called the 'validation' or 'test' data) are not used at all during model development, but are available on the file to compare the performance of the models obtained by the different methods.

Tutorial Part B shows that the Correlated Component Regression (CCR) method also outperforms both stepwise regression and lasso in a *high-dimensional data* setting, created by performing the regression with a reduced N formed by taking a 1% sample from the 5,000 available training cases, the sample size thus being reduced to N=50. 'High-dimensional data' refers to data where the number of predictors P is large relative to the sample size N, and may exceed N by a small amount (P>N), or by a large amount (P>>N). Although the number of predictors (P=56) is relatively large, the data used in Part A are not considered 'high dimensional' since P << N.

The Data

Table 1A compares regression results obtained from CCR and stepwise regression using the entire training sample of size N=5,000. Column 1 lists the subset of predictors for which non-zero coefficient estimates were obtained in at least one of the regressions. The 14 true predictors

are listed on top, the first column listing the true population coefficients. The *population* R^2 for these simulated data is approximately .913, and as a benchmark, the R^2 values obtained by applying the true coefficients to the training and independent validation data is .911 and .914 respectively, the slight differences reflecting sampling variation in the two generated samples. Since only 14 of the 56 predictors are *valid* predictors (i.e., true coefficients are non-zero for these), the true regression is said to be *sparse* (many coefficients equal zero).

Table 1A: Comparison of Results for CCR and Stepwise Regression Models estimated on Training Data (N_{Tr} = 5,000) and Evaluate Using Validation (Test) Data (N_{Val} = 5,000)

		CCR	Stepwise Regression	
	TRUE	**K=8**	**forward**	**backward**
R-sq (Tr) =	0.911	0.911	0.912	0.912
R-sq (CV) =	N/A	0.911		
R-sq (Val) =	0.914	0.913	0.913	0.913
Predictors				
BRCA1	-2.13	-2.2	-2.2	-2.2
CD44	1.85	1.69	1.68	1.68
CD97	1.44	1.45	1.39	1.4
CDKN1A	2.33	2.34	2.34	2.33
EP300	-1.78	-1.64	-1.7	-1.69
GSK3B	4.56	4.59	4.55	4.56
IQGAP1	3.35	3.27	3.33	3.32
MAP2K1	2.75	2.48	2.64	2.73
MYC	-1.81	-1.77	-1.79	-1.77
RB1	-3.82	-3.68	-3.73	-3.75
RP5	5.75	5.8	5.77	5.78
SIAH2	1.15	1.12	1.14	1.14
SP1	-9.55	-9.44	-9.39	-9.39
TNF	2.24	2.25	2.26	2.27
Other1	0	0	0	-0.11
extra4	0	0	0	-0.13
extra5	0	0	0	0.06
extra13	0	0	0	0.05
extra14	0	0	0.06	0.08
extra16	0	0	0	-0.04
extra28	0	0	0	0.06

Of the 42 *extraneous* predictors (i.e., true coefficients equal zero for these), 14 (labeled 'other1-other14') are correlated with the 14 valid predictors, and each of the remaining 28 extraneous predictors (labeled 'extra1-extra28'), is uncorrelated with each of the other 55 predictors. Theoretically, prediction can never be improved by including any of the irrelevant predictors 'extra1-extra28' in the model, but if some of the valid predictors were excluded, it is possible that prediction can be improved by including one or more extraneous predictors 'other1-other14' that are correlated with the valid predictors excluded.

Column 2 in Table 1A contains results from the K=8-component CCR model. As K is reduced in value, the amount of regularization goes up. We selected the model with K=8 components by applying a tuning process based on 10-fold cross-validation, results of which are summarized in Table 1B. Table 1A shows that this model, CCR8, correctly yields non-zero coefficients for the 14 valid predictors and correctly excludes all of the extraneous predictors. The remaining columns in Table 1A show that stepwise (backward and forward) regression yields similar results in terms of the Validation R^2 based on this large sample of N=5,000. However, the stepwise solutions include at least 1 irrelevant predictor in the model.

In contrast to CCR and stepwise regression which are invariant to any linear transformation applied to the predictors, *penalized* regression methods such as lasso require that the predictors be standardized. Lasso also yields results similar to CCR based on this large sample size in terms of validation R^2 but is somewhat worse than both CCR and stepwise regression in terms of predictor recovery, resulting in 23 non-zero coefficients, including 7 irrelevant plus 2 extraneous variables. (The Lasso solution was obtained using GLMNET, tuned using the M-fold cross-validation procedure included in that package).

To determine the number of components K for the CCR model, Table 1B summarizes cross-validation output obtained from COREXpress, for K in the range 2-12. This output includes the cross-validation R^2 statistic (CV-R^2) supplemented by cross-validated predictor counts. Note that CV-R^2 steadily increases as K goes from 2 to 8, and then beginning with K=8 only increases slightly as K increases further for K = 9, 10, 11 and 12. For each K, the bottom row reports the number of predictors that maximize CV-R^2 when that number of predictors is included in the associated K-component model. Note that the correct number P*=14 is reported for K=7-9.

For each predictor, the body of Table 1B reports the number of the 10 CV-Subsamples for which the CCR step-down procedure included that predictor in the model. For example, for CCR8 and CCR9, when the CCR step-down procedure was applied in each of the 10 CV-subsamples (each subsample excluding one of the 10 folds), the 14 true predictors (and only these predictors) were correctly included in the model each and every time, for a total of 10. Table 1 is based on 1 round of 10-folds.

Although the models with K=10, 11, or 12 components yield a higher CV-R^2 than models with K=8 and K=9, CV-R^2 is only slightly higher for the former models, and the predictor counts are

less consistent than the latter models, reporting counts less than 10. Thus, by selecting the model with the smallest K among those having (approximately) the highest CV-R^2, we obtain greater consistency in terms of the predictor counts. This type of model selection criterion is similar to that recommended for lasso -- the selected model being the most parsimonious model among those for which the CV error rates are within 1 standard deviation of the lowest CV error rate. (The standard error for the CV-R^2 can be computed and is displayed in the COREXpress output, when more than 1 round of M-folds is requested.)

Table 1B: Frequency of predictor occurrence in 10 CV-Subsamples for specified K Components

# Components	12	11	10	9	8	7	6	5	4	3	2
CV-R^2	0.9111	0.9111	0.911	0.911	0.9109	0.909	0.8980	0.8911	0.8659	0.81	0.56
BRCA1	10	10	10	10	10	10	10	10	10	10	10
CD44	10	10	10	10	10	10	10	10			
CD97	10	10	10	10	10	10	10	10	10	10	
CDKN1A	10	10	10	10	10	10	10	10	10	10	10
EP300	10	10	10	10	10	9	2				
GSK3B	10	10	10	10	10	10	10	10	10	10	
IQGAP1	10	10	10	10	10	10	10	10	10	10	
MAP2K1	10	10	10	10	10	10	10	10	10	10	
MYC	10	10	10	10	10	10	10	10	10	10	10
RB1	10	10	10	10	10	10	10	10	10		
RP5	10	10	10	10	10	10	10	10	10	10	10
SIAH2	10	10	10	10	10	10	10	10	10	10	10
SP1	10	10	10	10	10	10	10	10	10	10	10
TNF	10	10	10	10	10	10	10	10	10	10	
Other1	10	10	7								
Other10							10	10		10	
Other12		1									
Other13	7	4				1	8	10		10	
Other14		2	2								
extra4	3	9									
extra5		4									
extra14	10	10	1								
# Predictors (P*)	17	18	15	14	14	14	15	15	12	13	6
Total count	170	180	150	140	140	140	150	150	120	130	60

For Part B, in order to provide comparisons that are comparable to those in Part A but based on 'high-dimensional data', we maintain the P=56 predictors but reduce the sample size by randomly dividing the training sample into 100 equal sized subsamples, each of size N=50. Thus for Part B we have P>N. Table 2 provides results from CCR (again, the 8-component model was obtained based on the CV criteria), and stepwise regression, for the first such subsample

('simulation' = 1). Results from the backward elimination option are not reported because this option cannot be performed with P>N due to singularity of the covariance matrix. The results from CCR8 are discussed in more detail in Tutorial Part B.

Table 2: Comparison of CCR and Stepwise Regression Results based on Simulation #1 (N=50)

	TRUE	CCR8	Stepwise Regression	
Rsq.Tr =	0.97	0.89	0.95	
Rsq.Val =	0.91	0.71	0.68	Reported
	Coefficients			p-vals
BRCA1	-2.13	-2.23	-1.51	0.004
CD44	1.85	0	0	
CD97	1.44	2.77	2.92	0.00005
CDKN1A	2.33	3.33	2.15	1.60E-06
EP300	-1.78	-1.6	0	
GSK3B	4.56	0	0	
IQGAP1	3.35	3.57	6.16	5.30E-07
MAP2K1	2.75	0	0	
MYC	-1.81	0	0	
RB1	-3.82	0	0	
RP5	5.75	6.25	6.63	4.40E-12
SIAH2	1.15	0	0.98	0.00023
SP1	-9.55	-8.66	-9.75	1.20E-14
TNF	2.24	2.78	2.43	2.00E-06
Other2	0	0	-1.68	0.009
Other3	0	0	0.56	0.001
Other4	0	0	-2.69	0.024
extra9	0	0	1.12	0.005

These data are available in an SPSS .sav file in demo data set #1 ('LMTr&Val.sav'). The file contains of training data ('Validation' = 1) consisting of all 100 simulated data sets of size N=50 ('Simulation' = 1-100), and an equal sized validation (test) data set ('Validation' = 1).

Note that Table 2 shows a relatively large training sample R^2 based on the true model, R^2 (Tr) =.97, which indicates that the observed dependent variable for this subsample tends to be more strongly related to the true model predictions than the average subsample. The associated validation R^2 of .91 is obtained using all N=9,950 cases outside the training data.

Comparing the top 2 rows of Table 2 shows that CCR outperforms stepwise regression on this subsample. The validation R^2 is higher for CCR (.71 vs. .68), and the drop-off in R^2 from the training to the validation sample is substantially smaller for CCR than for stepwise regression

(.89 - .71 = .18 vs. .95 - .68 = .27), indicating greater reliability. In addition, CCR includes 8 of the valid and none of the extraneous predictors, compared to 8 valid plus 4 extraneous predictors included by stepwise regression. The p-values (right-most column) reported in the stepwise regression output are substantially less than .05 for all predictors, mistakenly suggesting statistical significance. It is well known that these p-values have a downward bias due to the effects of selection.

Lasso again performs worse than either CCR or stepwise in terms of the validation R^2. The corresponding results from lasso are as follows:

a) Rsq (Tr) = .88, Rsq (Val) = .61,

b) 10 valid plus 5 irrelevant plus 2 extraneous predictors were included in the model.

Overall, across all 100 subsamples, CCR outperformed stepwise regression. On average, the CCR model includes 2 more valid predictors than stepwise regression (9.0 vs. 7.1) and approximately the same number of extraneous predictors (2.5 vs. 2.2). In addition, the average correlation between the CCR predicted score and the score obtained based on the true model is .942 compared to the smaller correlation of .907 using the predicted score obtained from stepwise regression. More details of these comparisons can be found in Magidson (2011, forthcoming).

Magidson, J. (2011). "Rethinking Regression, Prediction and Variable Selection in the Presence of High Dimensional Data: Correlated Component Regression." To be presented at Modern Modeling Methods Conference. U. of Connecticut. May 20, 2011.

Part A: CCR-Linear with Sample Size N=5000

Overview

In tutorial 1A, we will use COREXpress to estimate the 8-component model based on the training sample of size N=5,000. Fig. 1 shows CV-R^2 as a function of P for K=8. We see that for K=8 components, CV-R^2 achieves its maximum of .9109 with 14 predictors. Table 1 shows that the 14 predictors selected are in fact the 14 valid predictors, that the estimated coefficients are very close to the true coefficients, and that the R^2 based on the true coefficients is .9113. In addition, these 14 predictors were obtained in *each* of the 10 sub-analyses conducted as part of the 10-fold cross-validation.

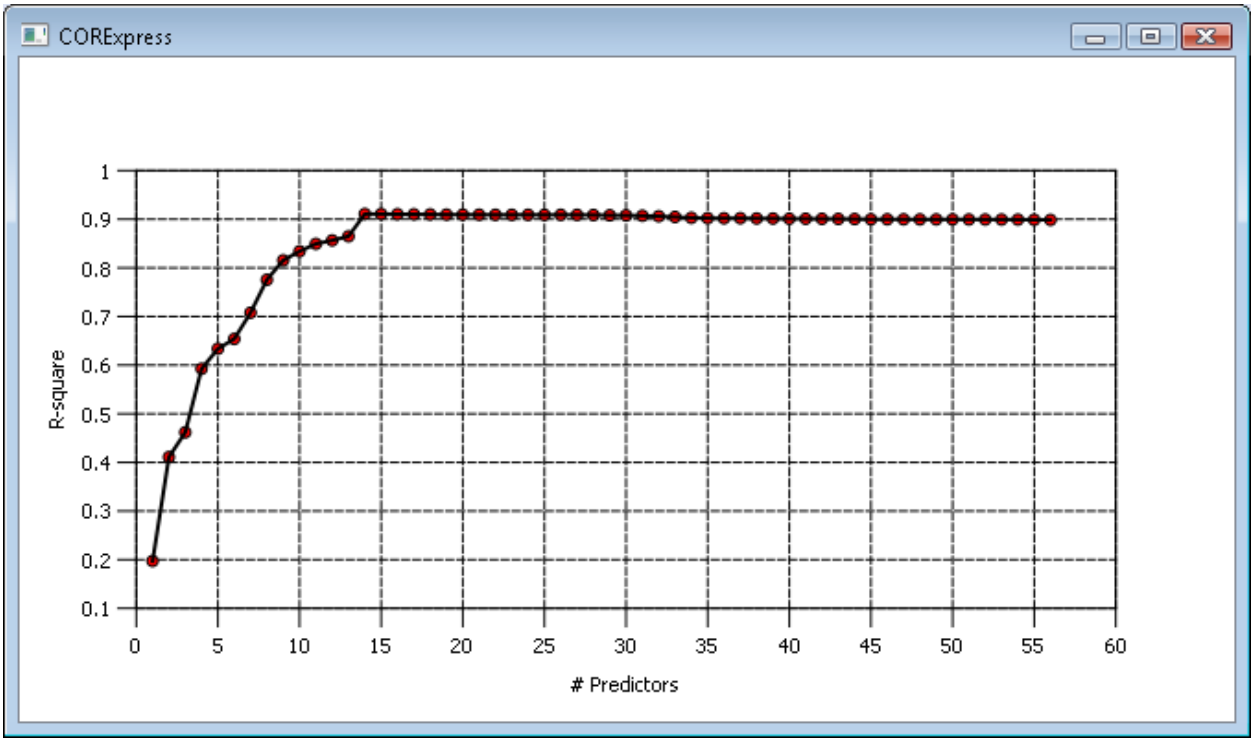


Fig. 1. CV-R^2 Plot of CV-R^2 for K=8. The maximum value for CV-R^2 occurs with P=14 predictors included in the model.

The follow steps show how to use COREXpress to obtain the CCR8 model coefficients for the 14 valid predictors as reported in Table 1A above and the graph in Fig. 1 which shows that the maximum value for CV-R^2 occurs with P=14 predictors included in the 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 'LMTr&Val.sav' and click Open to load the dataset

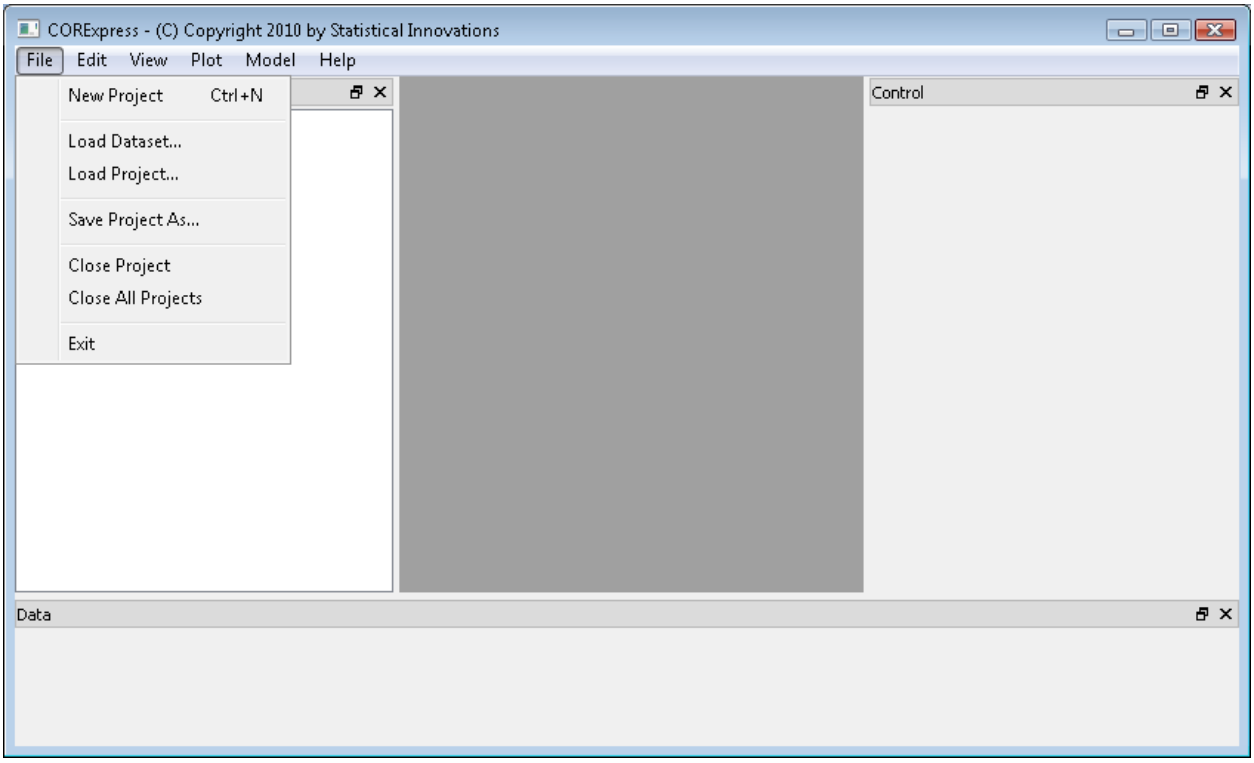


Fig. 2: File Menu

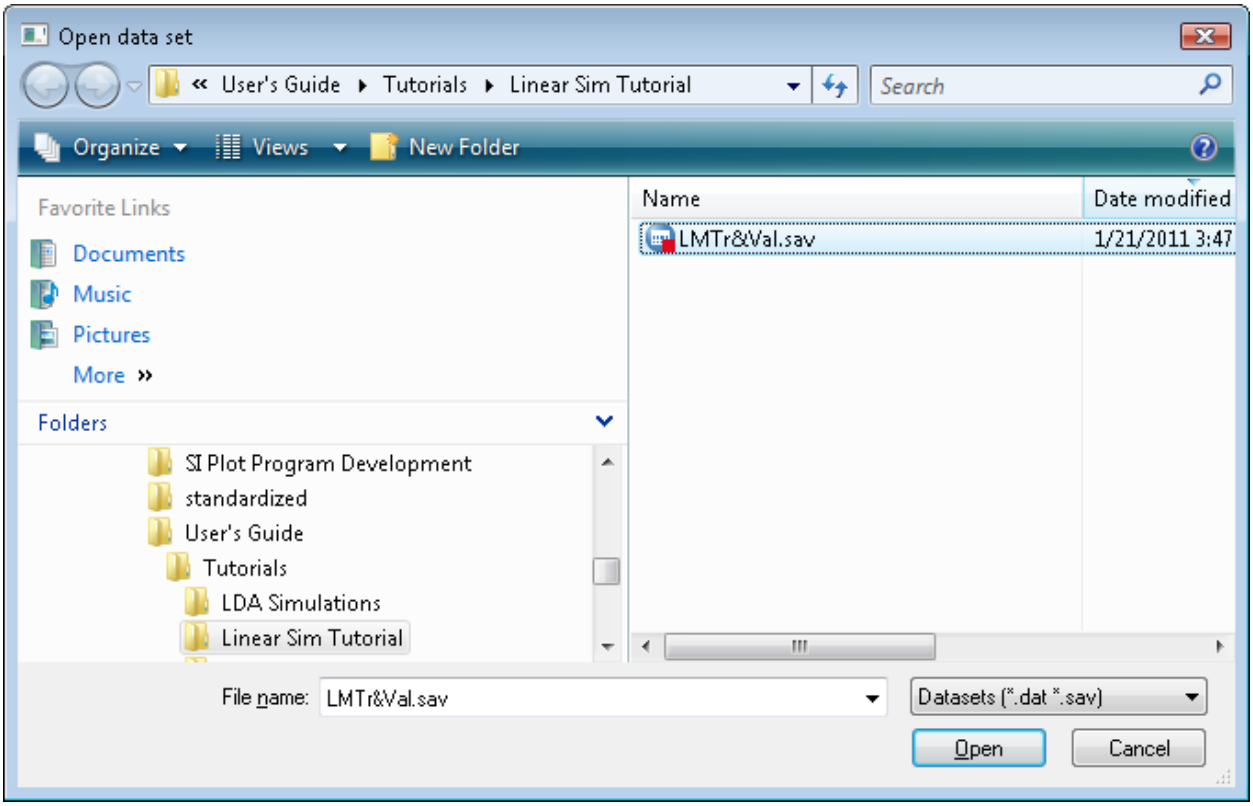


Fig. 3: Loading a Dataset

You will now see the "LMTr&Val" 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.

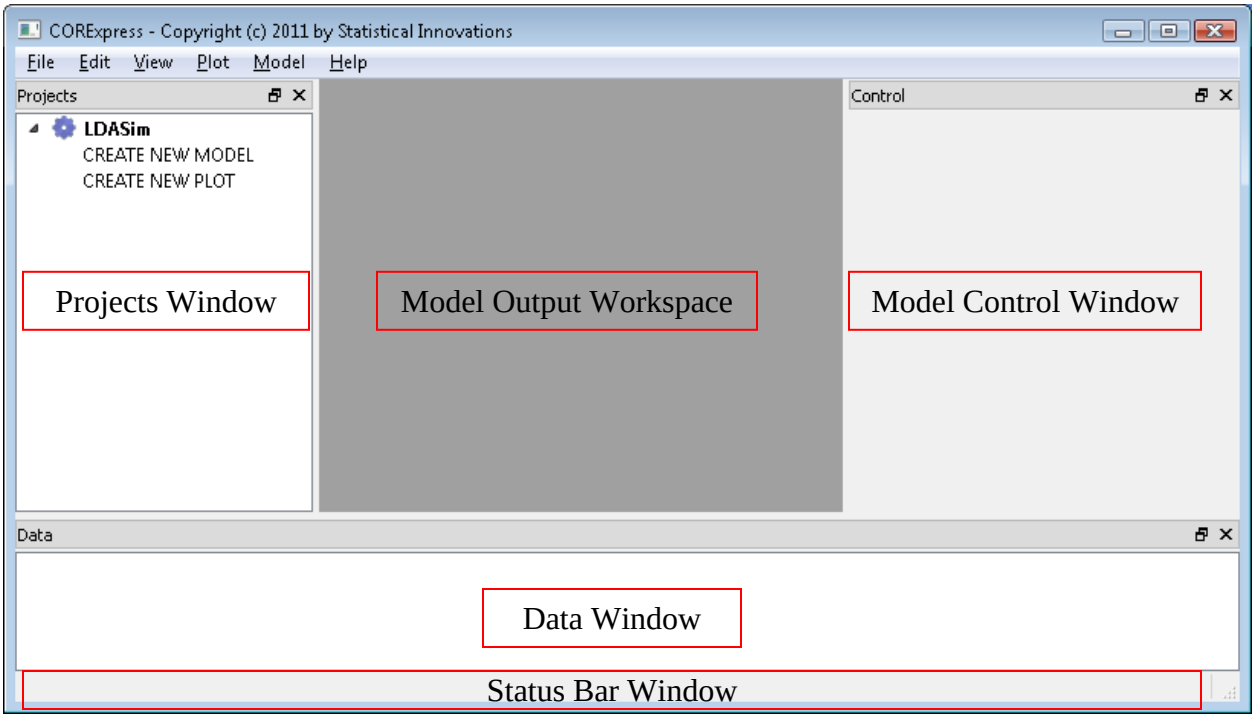
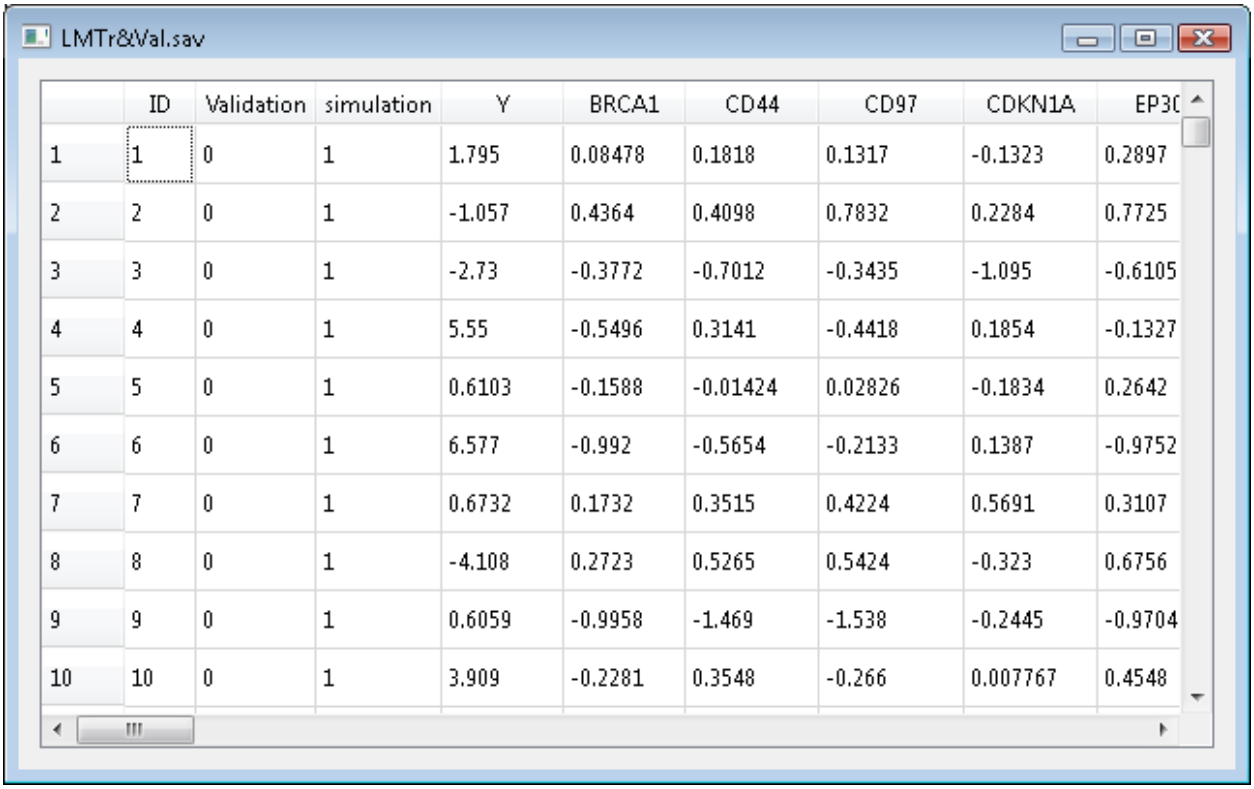


Fig. 4: COREXpress Windows

You can view the complete dataset in a new window by double clicking on "LMTr&Val" in the Projects window. After estimating a model, the predicted scores will automatically be added to the file, and any cases not used to estimate the model (holdout cases) will also be scored.



	ID	Validation	simulation	Y	BRCA1	CD44	CD97	CDKN1A	EP3(
1	1	0	1	1.795	0.08478	0.1818	0.1317	-0.1323	0.2897
2	2	0	1	-1.057	0.4364	0.4098	0.7832	0.2284	0.7725
3	3	0	1	-2.73	-0.3772	-0.7012	-0.3435	-1.095	-0.6105
4	4	0	1	5.55	-0.5496	0.3141	-0.4418	0.1854	-0.1327
5	5	0	1	0.6103	-0.1588	-0.01424	0.02826	-0.1834	0.2642
6	6	0	1	6.577	-0.992	-0.5654	-0.2133	0.1387	-0.9752
7	7	0	1	0.6732	0.1732	0.3515	0.4224	0.5691	0.3107
8	8	0	1	-4.108	0.2723	0.5265	0.5424	-0.323	0.6756
9	9	0	1	0.6059	-0.9958	-1.469	-1.538	-0.2445	-0.9704
10	10	0	1	3.909	-0.2281	0.3548	-0.266	0.007767	0.4548

Fig. 5: COREXpress Dataset View

Estimating a CCR Model

Selecting the Type of Model:

- Double click on "CREATE NEW MODEL" in the Workspace window under "LMTr&Val"

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 "Y" as the dependent variable.

Selecting the Predictors:

- In the Control window below "Predictors", click and hold on "BRCA1" and move the cursor down to "extra28" to highlight all 56 predictors. Click on the box next to "extra28" to select all 56 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 "BRCA1" and move the cursor down to "extra28" to highlight all 56 predictors in the left box.
- Click on the ">>" box in the middle to select all 56 predictors and move them to the right box as candidate predictors.

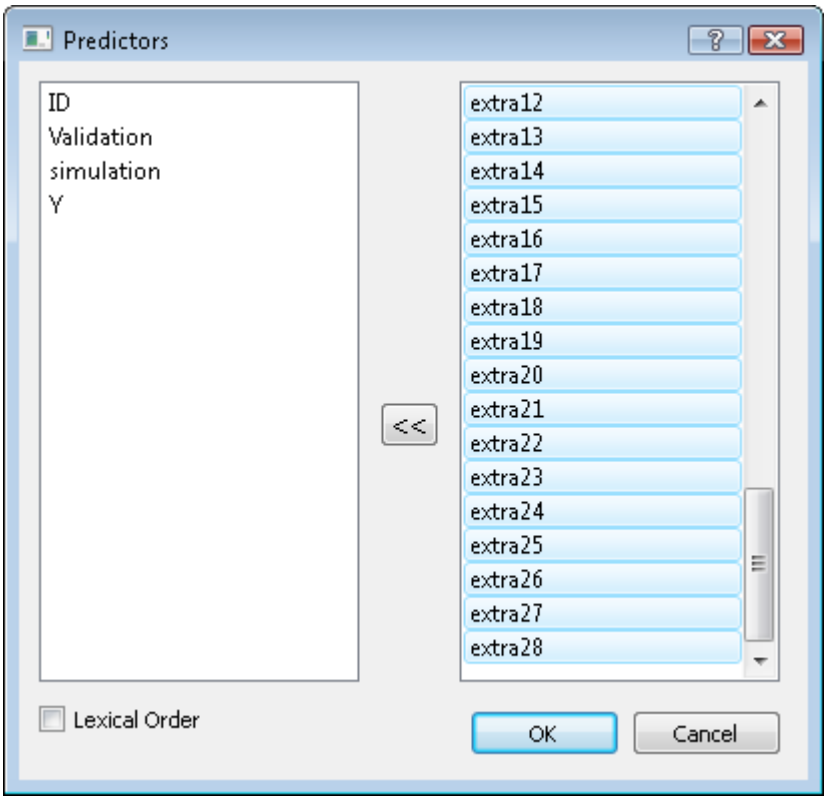


Fig. 6: Predictor Window

Specifying the Number of Predictors to Step Down:

- 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.
- In the "# Predictors:" box, keep the default number, "1"

Selecting the Number of Components:

- Under Options, click in the box to the right of "# Components", delete "4", and type "8"

Selecting the Model Type:

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

Your Control window should now look like this:

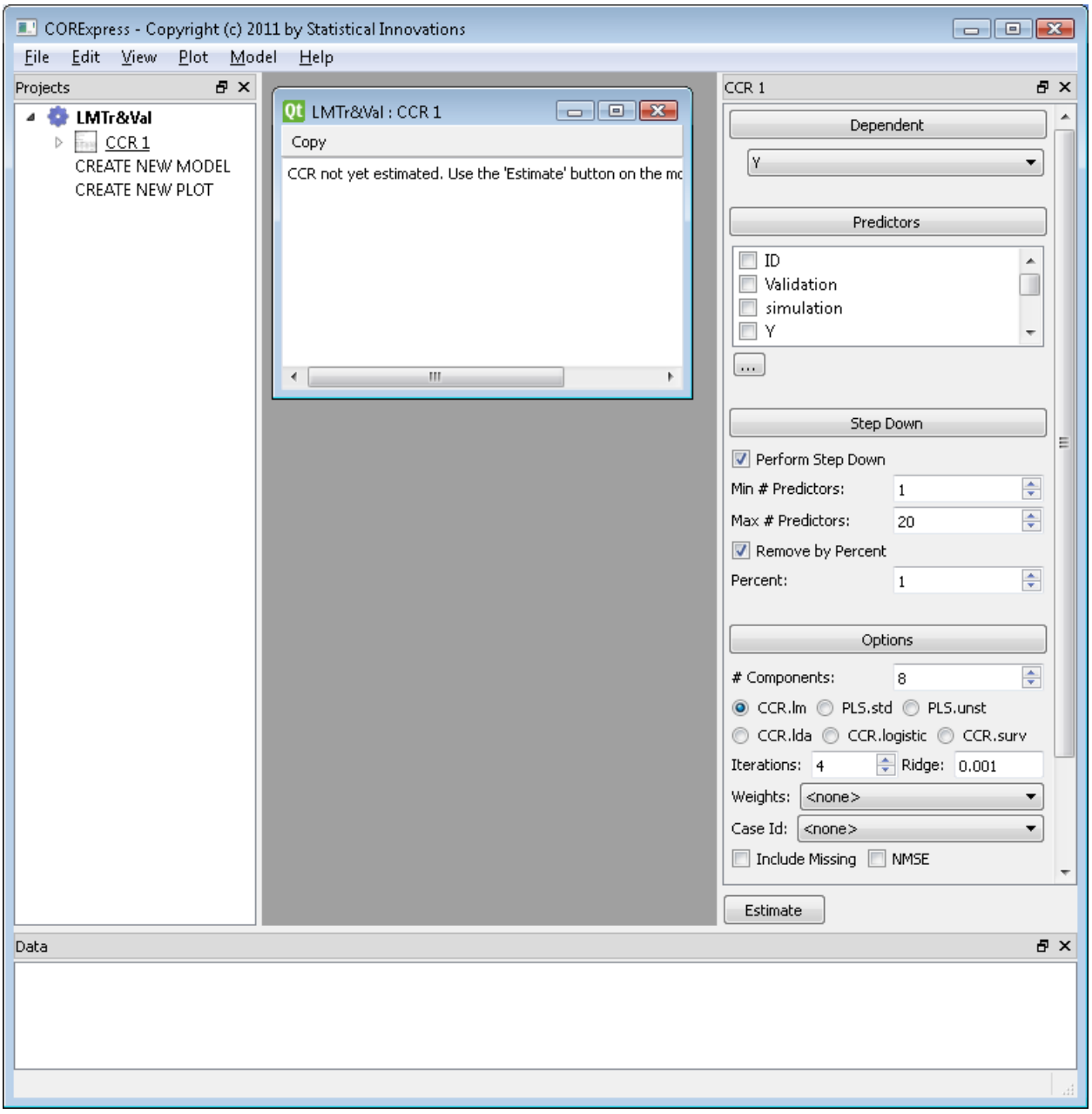


Fig. 7: Control Window

Specifying the Training Sample:

- Click on "Validation" and options will appear for selecting training and validation subgroups.
- Under the Training Subgroup, click on the "<select>" drop down menu and click on "Validation".
- Keep the default "=" in the drop down menu
- Keep the default "0" in the Training Subgroup numeric box.

Now, all records meeting the 'Validation=0' criterion will be selected as the Training sample, providing an analysis file of size N=5,000.

Specifying the Validation Sample:
By default, all cases meeting the criterion 'Validation~=0' will be automatically selected as the validation sample.

Specifying Cross Validation:

- 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, keep the default "1"
- In the "# Folds:" box, keep the default "10"
- Keep the "<none>" in the Fold Variable drop down drop down menu

This divides the analysis sample into 10 subsamples (folds) that can be used to obtain the optimal tuning parameters for the number of components K and the number of predictors P. The statistic to be used primarily will be CV-R^2, the cross-validation R^2. This statistic is computed using model scores obtained from the analysis sample, excluding cases a particular fold, and applied to cases in the excluded fold. The performance of the model is measured on the combined set of excluded cases, and thus is based solely on cases not used at all in the development of the model.

Your Control window should now look like this:

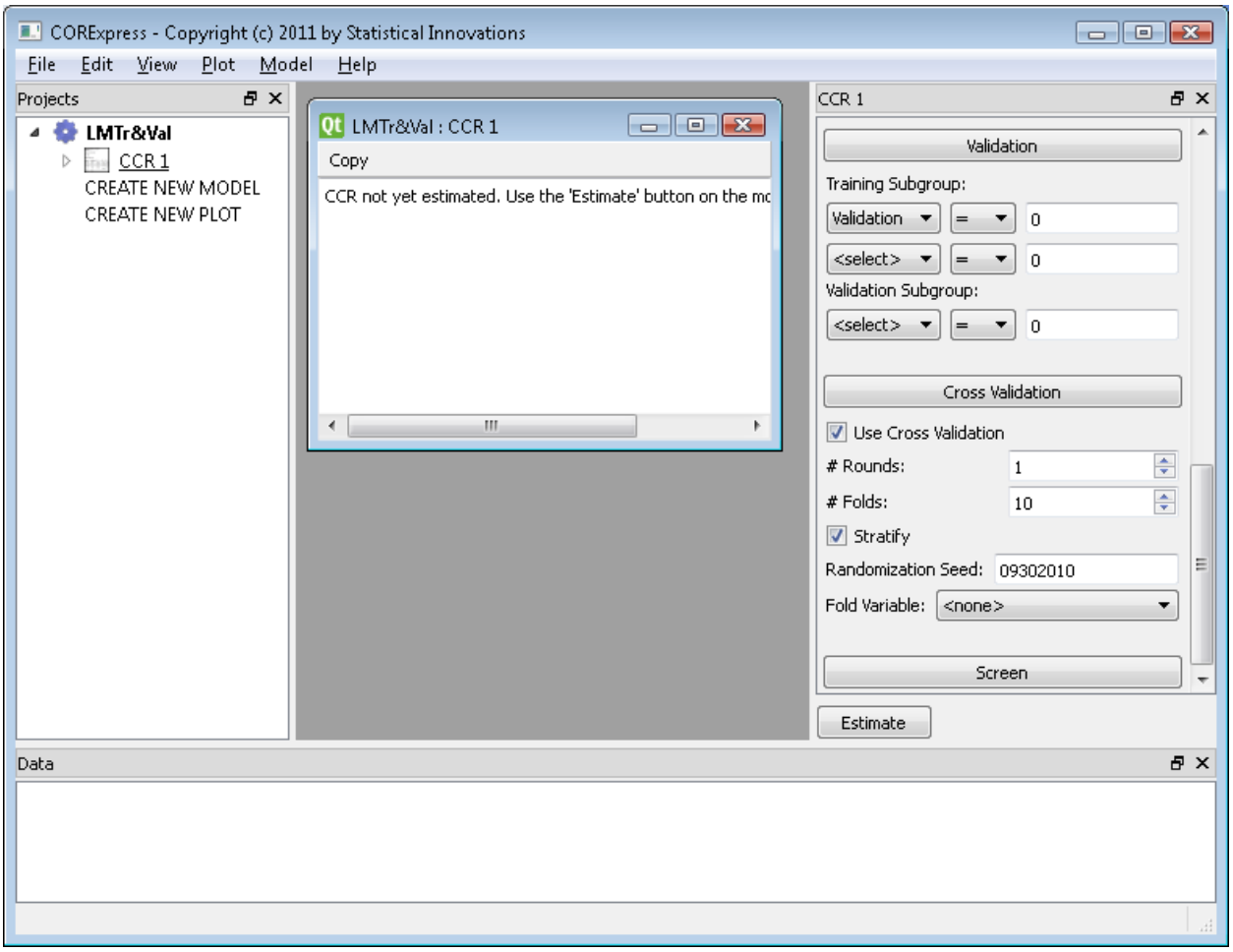


Fig. 8: Control Window

Estimate the Specified Model:

- Click on the "Estimate" box 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.

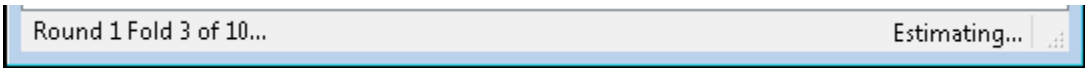


Fig. 9: COREXpress Status Bar

When the model is finished estimating, a new window will pop up: "COREXpress" (CV-R^2 Plot)

View Model Output

Viewing CV-R^2 Plot:

- Click on the "LMTr&Val : CCR 1" window (CV-R^2 Plot)

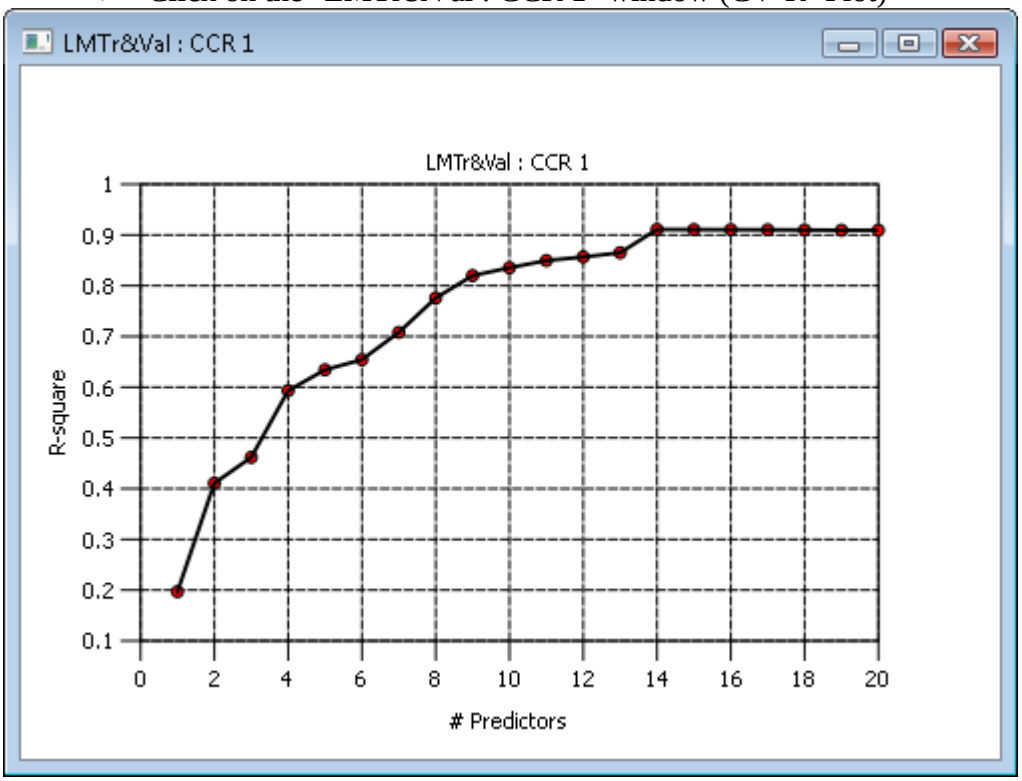


Fig. 10: CV-R^2 Plot

The CV-R^2 plotted in the graph corresponds to the cross-validation R^2 based on the 8-component model for number of predictors P ranging from 56 down to 1. Since we kept the default "20" for the "Max # Predictors", only predictors ranging from 20 down to 1 are shown. The number of predictors is determined as the smallest P that yields the maximum value for CV-R^2. The program then uses all cases in the analysis file to estimate the K-component CCR model with that number of predictors (and the specified value for K).

Viewing CV-R^2 Output:

- Click on the "LMTr&Val : CCR 1" window in COREXpress
- The CV-R^2 as well as the corresponding R^2 for the training and validation data are provided at the top of the output window.
- The unstandardized and standardized coefficients are provided in the main body of the output window.
- When the step-down procedure is selected with a specified range for the number of predictors, the cross-validation R^2 (CV- R^2) is reported next for each value of P in the selected range.
- Scroll down the " LMTr&Val: CCR 1" window past the coefficients

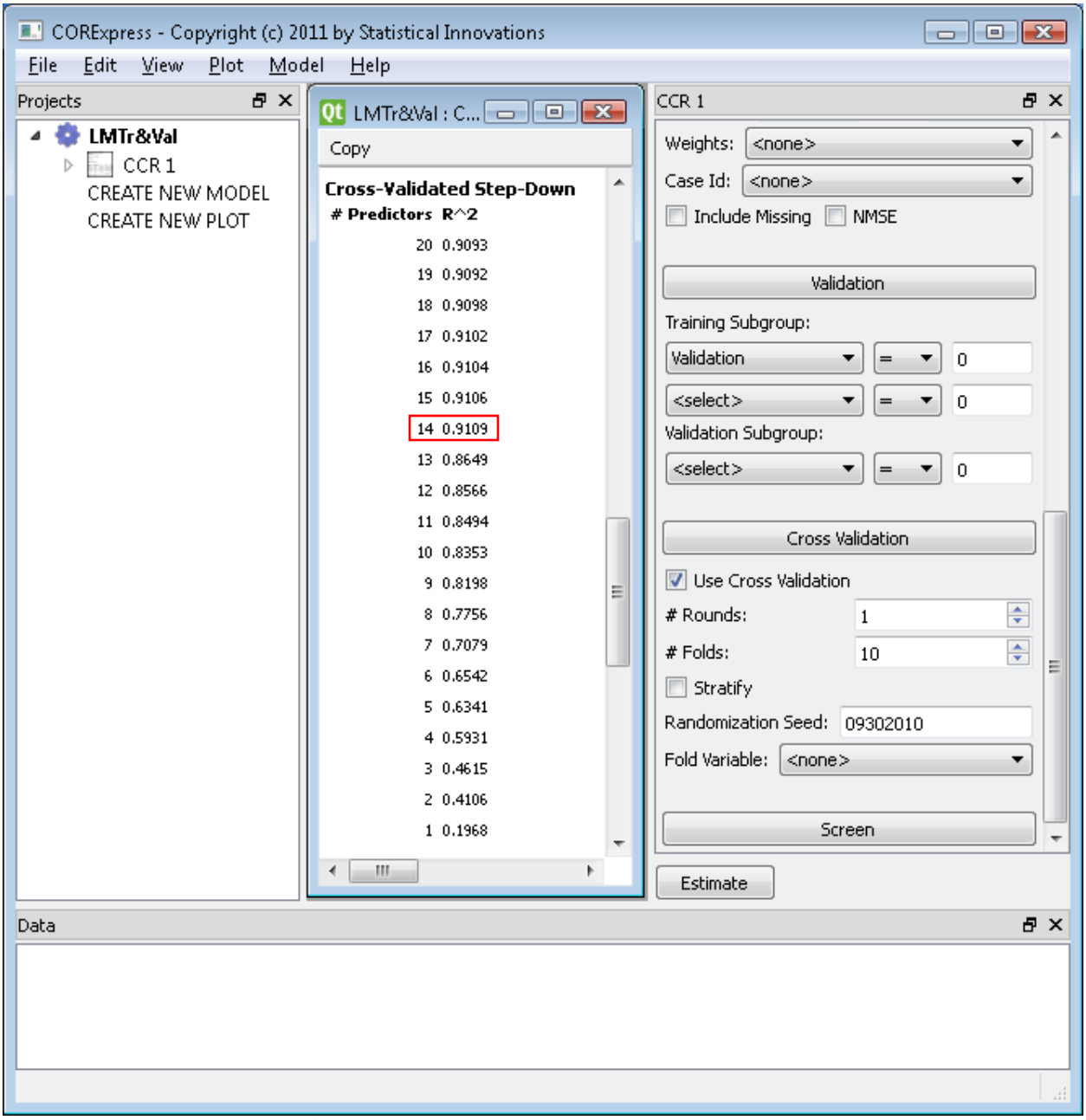


Fig. 11: Cross Validation R^2 Output in the Model Output Window

Model Output Window along with the CV-AUC and CV-ACC for each number of predictors. By default, the model estimated and shown in the model output window is the 'optimal' one -- the one with P* predictors, where P* is the value for P with the highest CV-R^2. In the case of ties, the optimal number of predictors P* is taken to be the smallest value for P among those with the same highest value for CV- R^2.

Viewing the 'Optimal' Model Output:

- Click on the "LMTr&Val : CCR 1" window in COREXpress
- Scroll to the top of the window

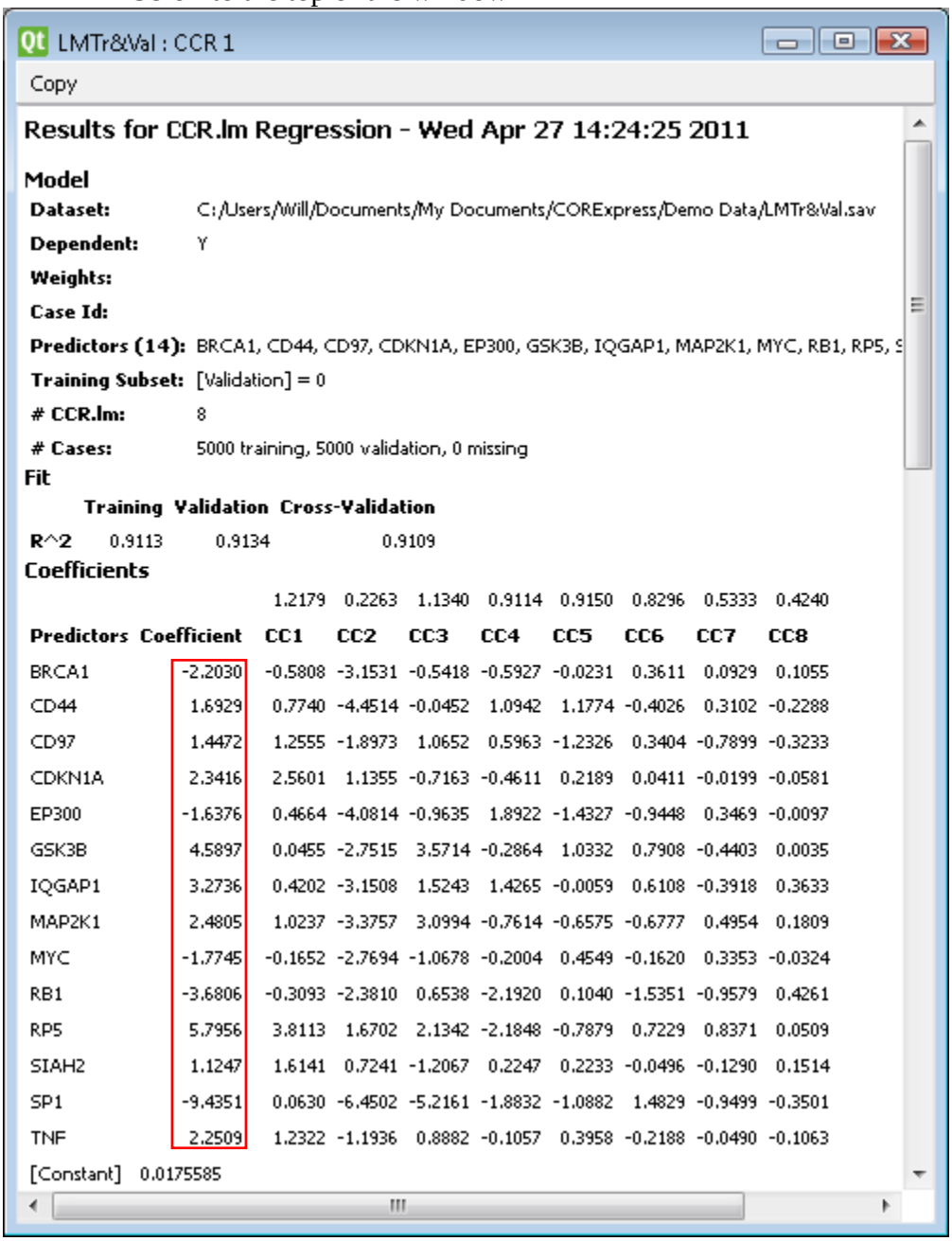


Fig. 12: Unstandardized Coefficients for K=8 in the Model Output Window
Note that for the Training, the R^2=0.9113 and for the Validation the R^2=0.9134.

When both the predictor Step-down Selection option and the Cross-validation option are both selected, the table of cross-validated predictor counts is provided at the bottom of the output window, followed by the processing time. (We will see that the processing time of about 1 minute, is substantially reduced in Part B when we reduce the sample size from 5,000 to 50.)

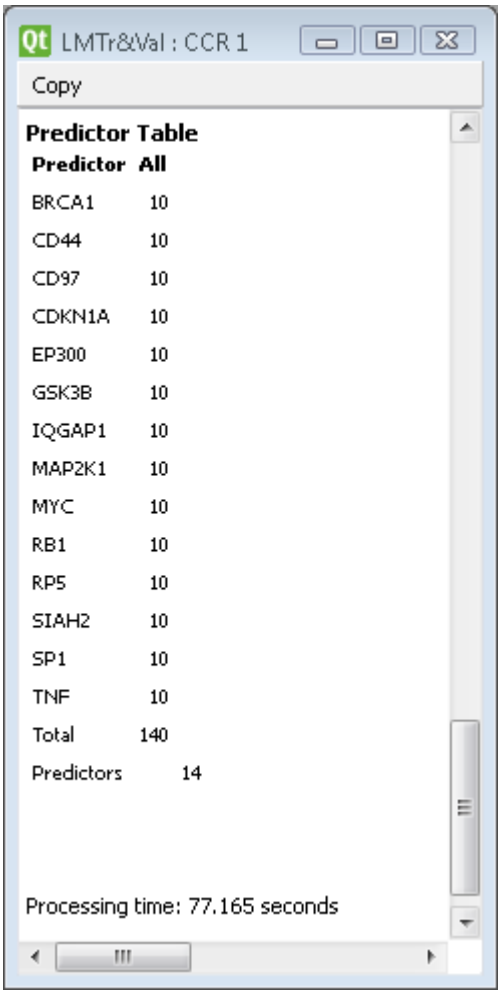


Fig. 13. Predictor Counts Table

Note that these match the counts provided in Table 1B above for the 8-component model.

Determining the Optimal Number of Components

By default, the optimal number of components K* is taken to be the value for K that achieves the highest CV- R^2 based on 1 round of M-fold cross-validation. In the case of 2 or more values for K that achieve approximately the same CV- R^2 value, a table of predictor counts can be examined to assist in the selection of K*, where each CV-Subsample contributes to the count based on the top P*(K) predictors for that subsample. If necessary, a more extended table of predictor counts can be provided based on R>1 rounds of M-folds to make a more informed decision, and output the standard error for the CV-R^2 statistic.

It is possible to get an extended prediction table containing information on more than 1 round of folds. For example, if R = 2 rounds of 10-folds are requested, the first round will be the same as the original and the second round is based on a different random selection of CV-Subsamples. In this case, the CV-R^2 is computed as the average of the CV-R^2 from both rounds.

Recall that K = 10 achieved the highest CV-R^2, but based on a single round of 10-folds, the results were somewhat inconsistent (recall Table 1B). Next we will request a second round of 10-folds to provide additional information.

To obtain results including a 2nd round of 10-folds for the K=10 component model

Specifying the # components and the 2nd round:

- Double click on CCR 1 from the Projects window
- In the Control Window, under Options, click in the box to the right of "# Components", delete "8", and type "10"
- In the "# Rounds" box in Cross Validation section of the Control Window, delete "1" and type "2"

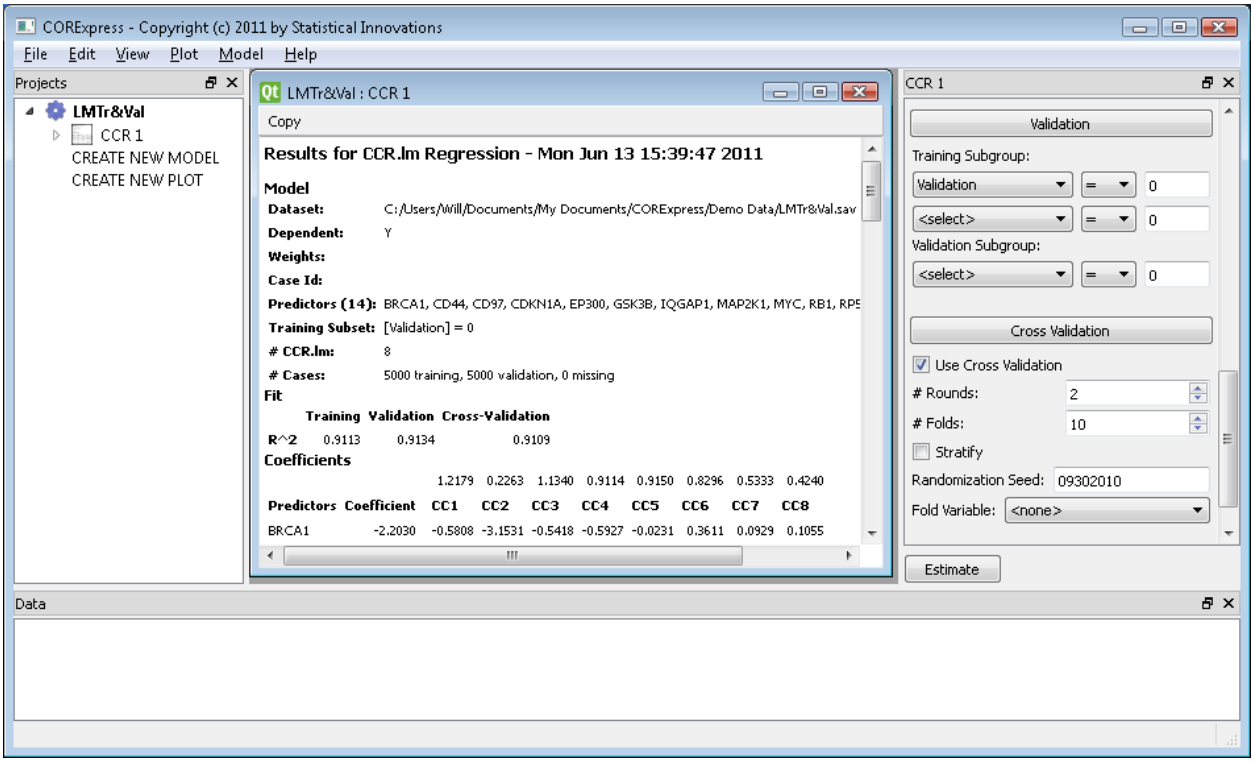


Fig. 14: Control Window

- Click "Estimate"

Since we request more than 1 round of M-folds, the standard error for the CV-R^2 now appears to the right of the CV-R^2 (see Fig. 14).

- Scroll to the bottom of the Control Window to see the updated predictor table of counts

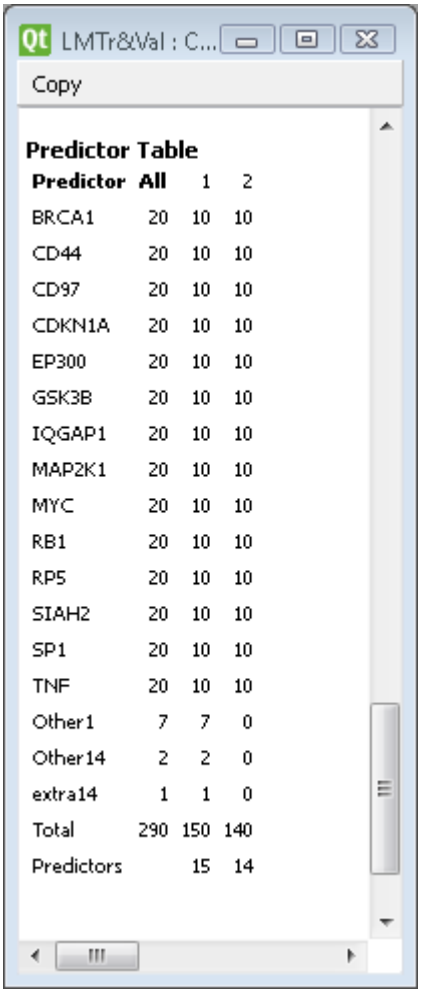
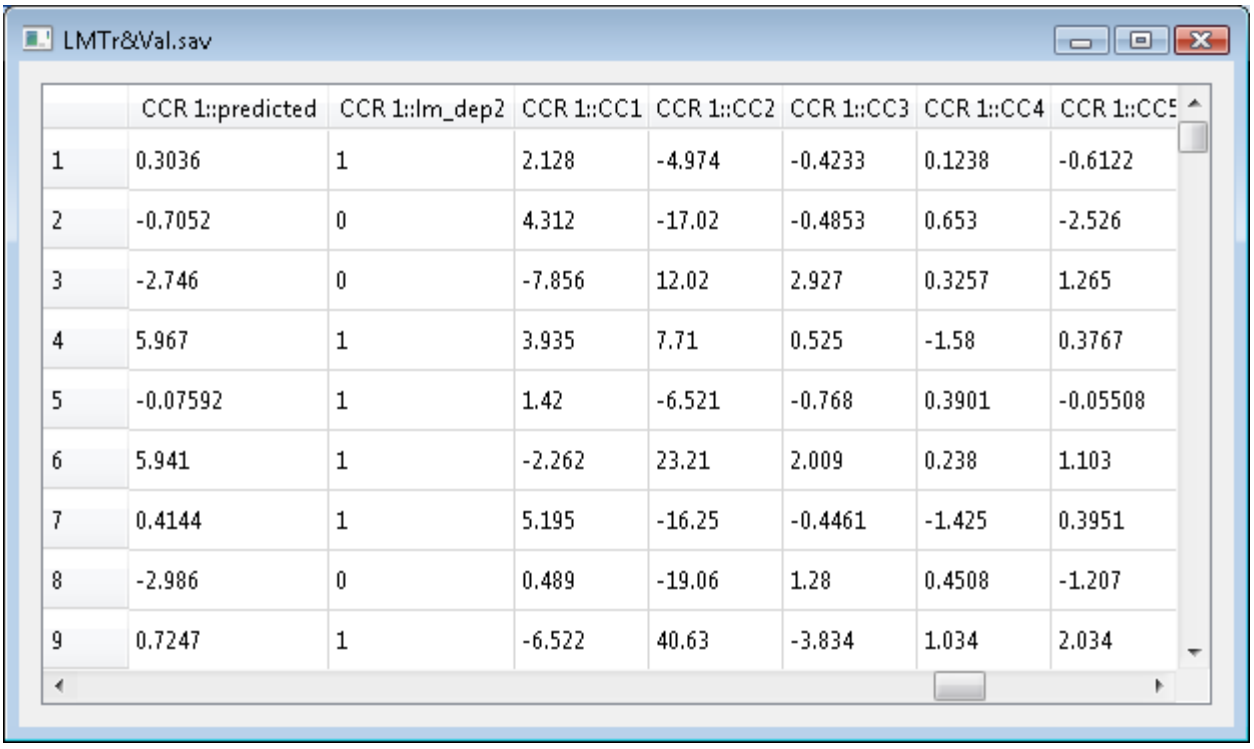


Fig. 15: Predictor Counts Table for K=10-component model with 2 rounds of 10-folds

Notice that round 2, based on a different random seed than round 1, resulted in all 14 valid predictors and no extraneous predictors being selected into the model.

Viewing the K Components and Predicted Scores on the Dataset:

- Double click on "LMTr&Val" in the Projects window
- Click on the window with the dataset and scroll all the way to the right.



	CCR 1::predicted	CCR 1::lm_dep2	CCR 1::CC1	CCR 1::CC2	CCR 1::CC3	CCR 1::CC4	CCR 1::CC5
1	0.3036	1	2.128	-4.974	-0.4233	0.1238	-0.6122
2	-0.7052	0	4.312	-17.02	-0.4853	0.653	-2.526
3	-2.746	0	-7.856	12.02	2.927	0.3257	1.265
4	5.967	1	3.935	7.71	0.525	-1.58	0.3767
5	-0.07592	1	1.42	-6.521	-0.768	0.3901	-0.05508
6	5.941	1	-2.262	23.21	2.009	0.238	1.103
7	0.4144	1	5.195	-16.25	-0.4461	-1.425	0.3951
8	-2.986	0	0.489	-19.06	1.28	0.4508	-1.207
9	0.7247	1	-6.522	40.63	-3.834	1.034	2.034

Fig. 16: K Components and Predicted Scores in the COREXpress Dataset View

The right-most variables contain the scores for each of the K components as well as the predicted score for the K-component model. It also contains the folds generated when cross-validation is performed, unless a specific fold variable is specified, along with all of the other variables on the file. After estimating another model, you can retrieve an updated data file window containing the updated model information by closing out of the data file window and double clicking on "CCR Demo" again. The new data file window will now contain the scores for the most recently updated model.

To copy the predicted scores and other variables from the data set window, click on the desired variables and type the shortcut "CTRL+C"> A window will pop up asking you if you also wish to copy the variable name.

Saving the Current Project

Save the Current Project:

- Click on File→Save Project As…
- A dialog box will pop up with the option to save the current project in the same directory as the dataset file.
- Type "LMTr&Val"
- Click "Save" to save the project.

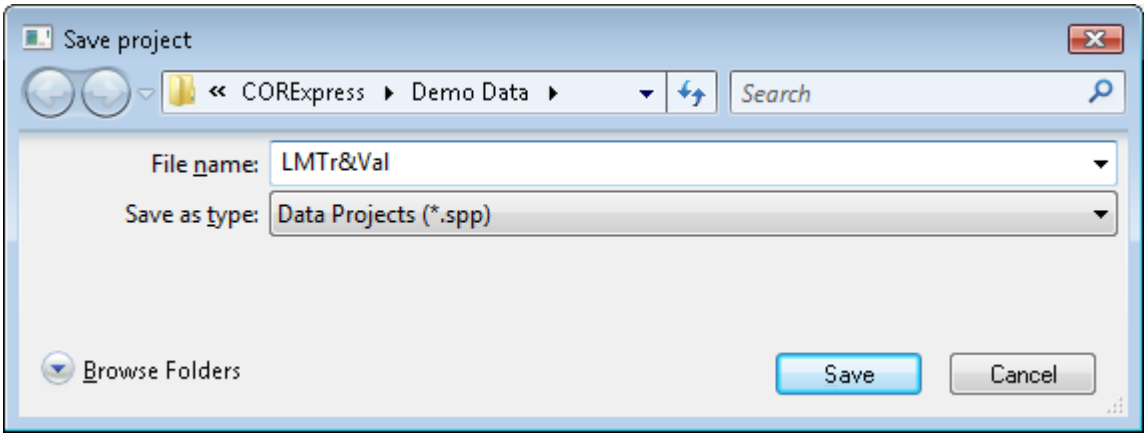


Fig. 17: Saving a Current Project

Part B: Performing CCR-Linear on High Dimensional Data

Now we will select a small subset for our analysis sample. If you did not just complete Tutorial #1A, we will begin by opening the saved COREXpress project from Tutorial #1A.

Opening the Previously Saved Project:

- File → Load Project…
- Select 'LMTr&Val.spp" and click Open to load the project

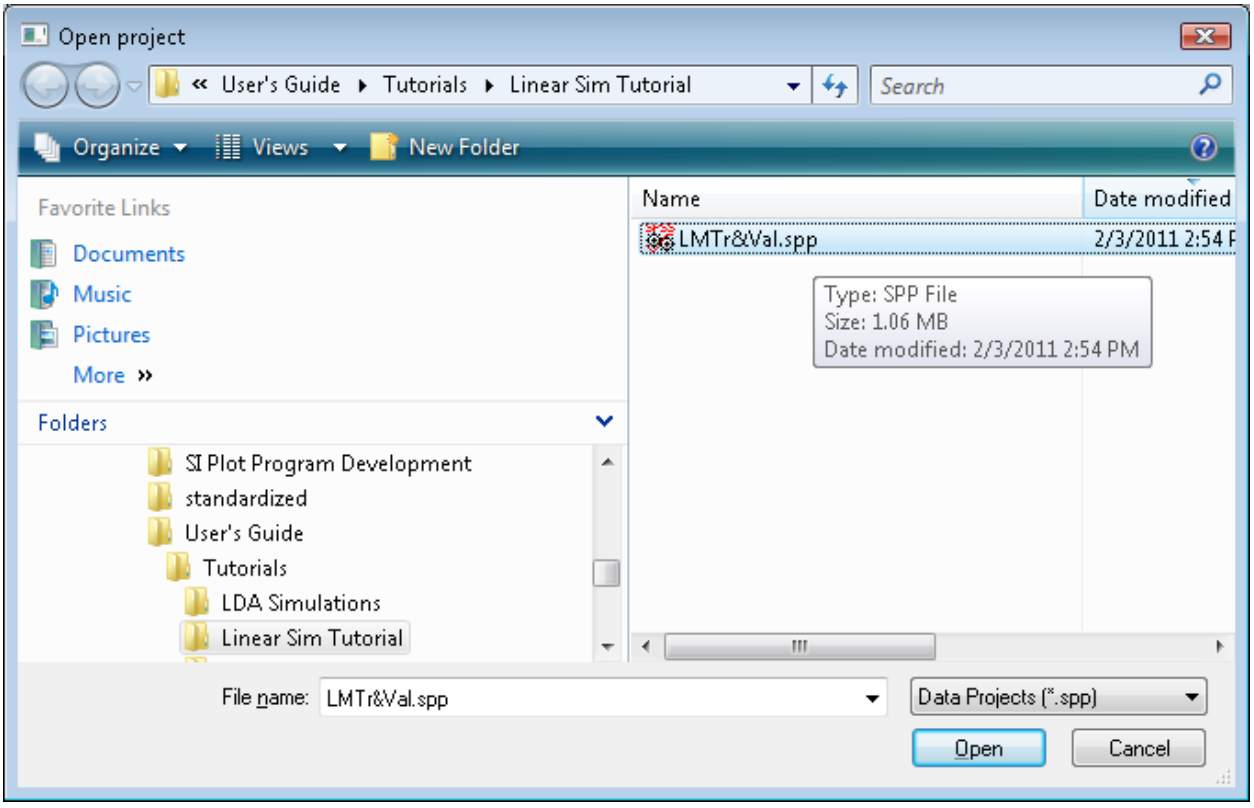


Fig. 18: Loading a previously saved project

Viewing Model Specifications & Output from Previously Saved Project

Opening the Model Specifications for the Saved Project:

- Double click on "CCR 1" in the Projects window

The control window will now show the saved model specifications and the model output window will show the previously saved model output corresponding to the model specifications.

Viewing the K Components and Predicted Scores from the Previously Saved Project:

- Double click on "LMTr&Val" from the Projects window

- Scroll to the right to see that COREXpress automatically saves K Components and Predicted Scores from the previously generated runs.

Performing CCR-Linear on High Dimensional Data

Selecting the Number of Components:

- Under Options, click in the box to the right of "# Components", delete "10", and type "8"

Specifying the Training Dataset:

- Click on "Validation" and options will appear for selecting training and validation datasets. Currently, under the Training Subset, there should be one specification: "Validation = 0"
- Click on (<select>) in the 2nd row of selection options and choose "simulation"
- Keep the default "=" in the drop down drop down menu
- In the Training Subset numeric box type "1"

Now, all records with Validation=0 and simulation=1 will be selected as the Training dataset, providing an analysis size of N=50.

Specifying Cross Validation:

- Click on the "Cross Validation" box and cross validation options will appear.
- In the "# Rounds:" box, delete "2" and type "10"

Estimate the Specified Model:

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

Your program should now look like this:

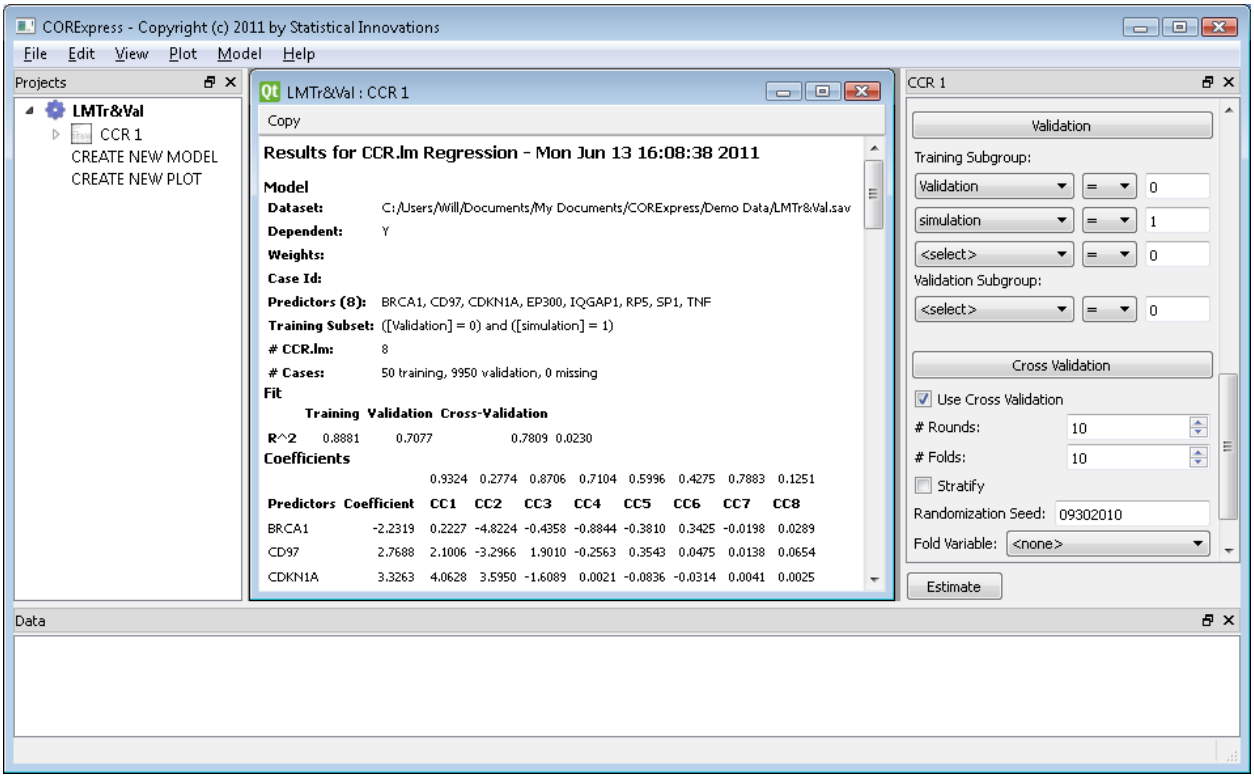
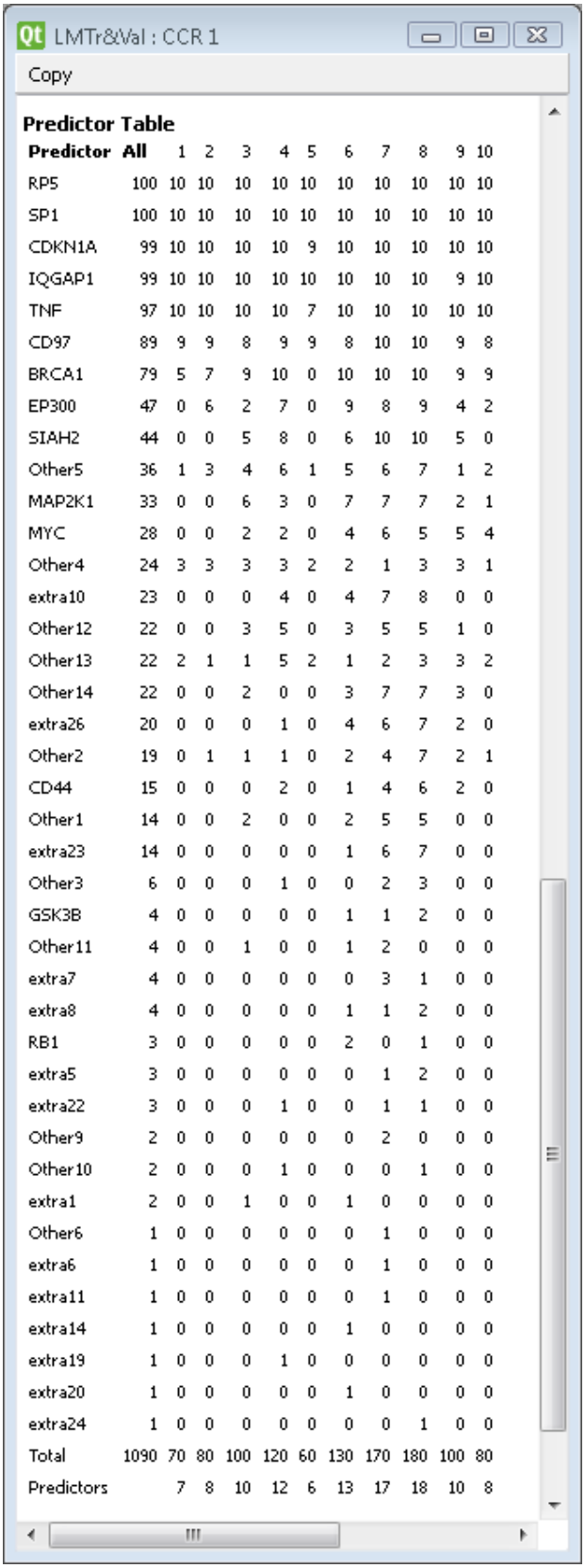


Fig. 19: Control Window

- Scroll to the bottom of the Control Window to see the updated predictor table of counts

Note the reduced estimation time on N=50 despite estimating 10 rounds of 10-fold.



Predictor Table

Predictor	All	1	2	3	4	5	6	7	8	9	10
RP5	100	10	10	10	10	10	10	10	10	10	10
SP1	100	10	10	10	10	10	10	10	10	10	10
CDKN1A	99	10	10	10	10	9	10	10	10	10	10
IQGAP1	99	10	10	10	10	10	10	10	10	9	10
TNF	97	10	10	10	10	7	10	10	10	10	10
CD97	89	9	9	8	9	9	8	10	10	9	8
BRCA1	79	5	7	9	10	0	10	10	10	9	9
EP300	47	0	6	2	7	0	9	8	9	4	2
SIAH2	44	0	0	5	8	0	6	10	10	5	0
Other5	36	1	3	4	6	1	5	6	7	1	2
MAP2K1	33	0	0	6	3	0	7	7	7	2	1
MYC	28	0	0	2	2	0	4	6	5	5	4
Other4	24	3	3	3	3	2	2	1	3	3	1
extra10	23	0	0	0	4	0	4	7	8	0	0
Other12	22	0	0	3	5	0	3	5	5	1	0
Other13	22	2	1	1	5	2	1	2	3	3	2
Other14	22	0	0	2	0	0	3	7	7	3	0
extra26	20	0	0	0	1	0	4	6	7	2	0
Other2	19	0	1	1	1	0	2	4	7	2	1
CD44	15	0	0	0	2	0	1	4	6	2	0
Other1	14	0	0	2	0	0	2	5	5	0	0
extra23	14	0	0	0	0	0	1	6	7	0	0
Other3	6	0	0	0	1	0	0	2	3	0	0
GSK3B	4	0	0	0	0	0	1	1	2	0	0
Other11	4	0	0	1	0	0	1	2	0	0	0
extra7	4	0	0	0	0	0	0	3	1	0	0
extra8	4	0	0	0	0	0	1	1	2	0	0
RB1	3	0	0	0	0	0	2	0	1	0	0
extra5	3	0	0	0	0	0	0	1	2	0	0
extra22	3	0	0	0	1	0	0	1	1	0	0
Other9	2	0	0	0	0	0	0	2	0	0	0
Other10	2	0	0	0	1	0	0	0	1	0	0
extra1	2	0	0	1	0	0	1	0	0	0	0
Other6	1	0	0	0	0	0	0	1	0	0	0
extra6	1	0	0	0	0	0	0	1	0	0	0
extra11	1	0	0	0	0	0	0	1	0	0	0
extra14	1	0	0	0	0	0	1	0	0	0	0
extra19	1	0	0	0	1	0	0	0	0	0	0
extra20	1	0	0	0	0	0	1	0	0	0	0
extra24	1	0	0	0	0	0	0	0	1	0	0
Total	1090	70	80	100	120	60	130	170	180	100	80
Predictors		7	8	10	12	6	13	17	18	10	8

Fig. 20. Predictor Count Table

Available Plots:

A wide range of traditional plot options for linear regression are also provided. To see these, double click "CREATE NEW PLOT" from the Projects Window

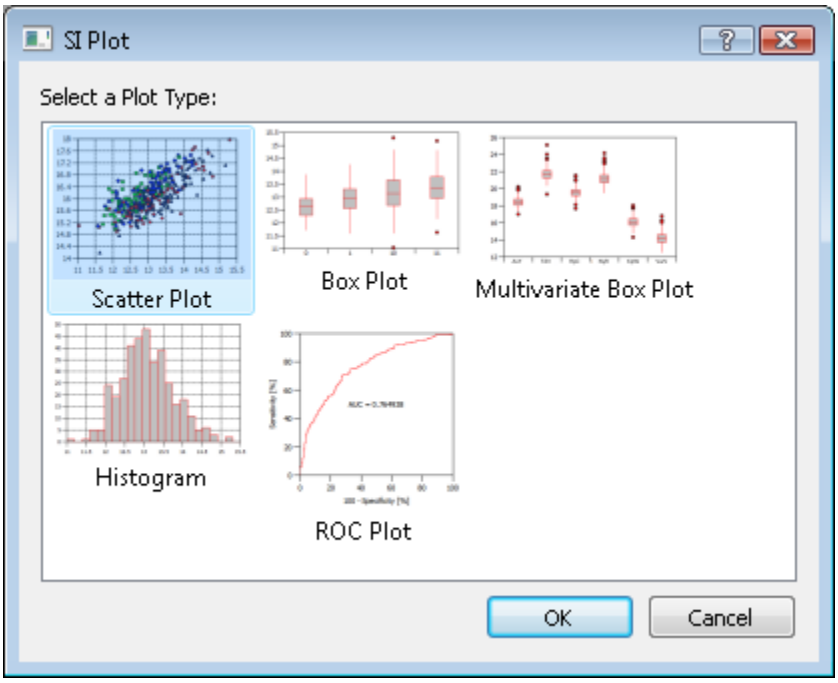


Fig. 21. Create New Plot Options

In particular:

Select the **Scatter Plot** option to construct additional scatterplots for the training data only, for the validation data only, for any selected subset of the data, or plots for each of the above.

Select the **Box Plot** to compare the distribution side by side for the 2 dependent variable groups.

Select the **Histogram** option to examine the distribution for any variable on the file, within any selected subset of cases.

To open the Project Settings menu:

- Right click on LMTr&Val at the top of the Projects Window.
- Select "Project Settings"

The following window appears:

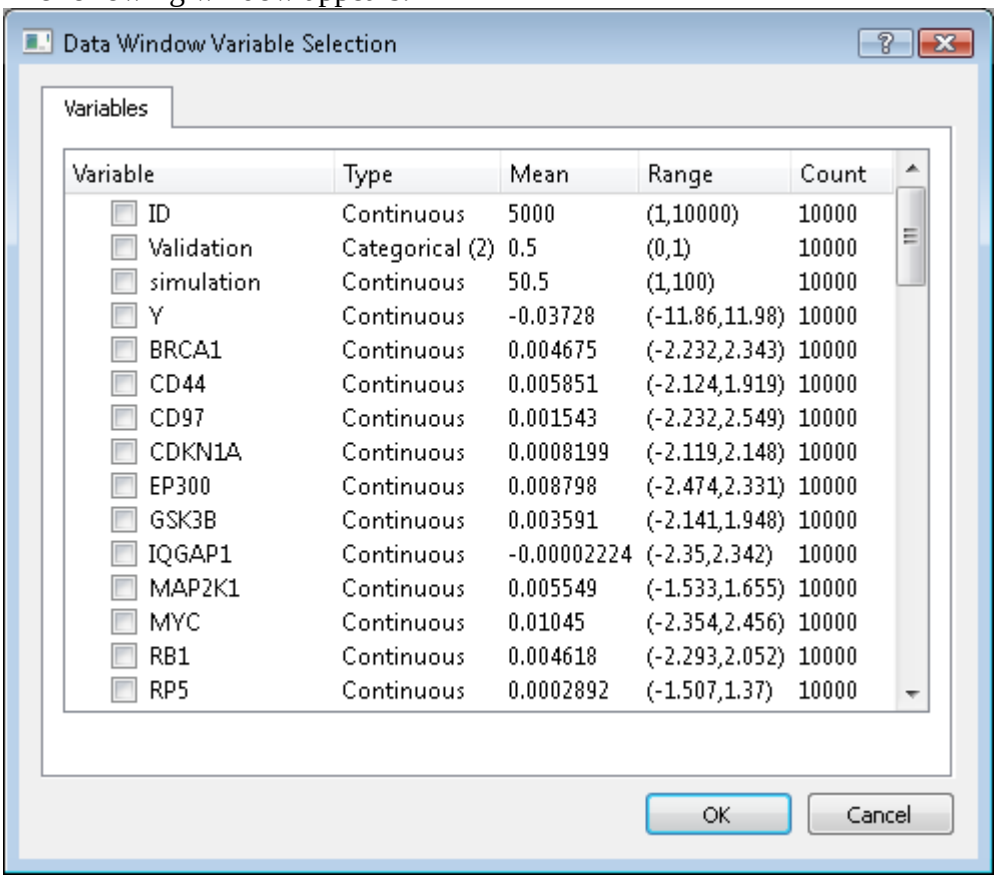


Fig. 22. Project Settings

Select the variable(s) that you wish to appear in the Data Window by checking the box to the left of the variable name.