Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Tutorial 3: Correlated Component Regression for a Dichotomous Dependent Variable

Overview

In regression analyses involving a dichotomous dependent variable (group 1 vs. group 2), a logistic regression model is most often the model of choice. In the special case where the predictor variables are continuous, and the within-group variances and covariances are identical within each group, the assumptions of linear discriminant analysis (LDA) are met, and the coefficients in the logistic regression model can be more efficiently estimated by LDA methods.

Often in practice, many predictor variables are available, and the number of predictors P may approach or even exceed the sample size N, a situation known as *high-dimensional data*. In such cases some kind of regularization is needed in order to get reliable predictions. For example, with high dimensional data that meets LDA assumptions, Bickel and Levina (2004) provided theoretical results showing that LDA performs poorly and is outperformed by a substantial margin by Naïve Bayes (NB), an approach which imposes an extreme form of regularization.

In COREXpress, the amount of regularization is determined by the number of components K that are included in the Correlated Component Regression (CCR) model. CCR1, the model with only K=1 component, provides the most extreme form of regularization, and is equivalent to NB. CCR2, the 2-component CCR model, almost always outperforms CCR1 on real data. The saturated CCR model, with K=N-1 components, imposes no regularization at all, and is equivalent to traditional regression models. Taking K as a tuning parameter, COREXpress implements M-fold cross-validation (CV) to help users select the optimal value for K. In practice, we have found that the best performance is generally achieved with K<10 regardless of the number of predictors. Estimation with a small value of K is fast.

In addition, predictions can be improved by excluding *extraneous predictor*s (those with true coefficients equal to zero) from the model. For a given K, COREXpress relies on results from M-fold CV to automatically determine the number of predictors P* to be included in the model, and then estimates the model with P* predictors, excluding the least important predictors. In this tutorial we show how COREXpress employs the CCR-lda model to analyze simulated data (demo data set #2) in a high dimensional setting involving P=84 predictors and sample size of N=100 or N=50. Note that in the latter situation, P > N. Results are better than those from stepwise LDA.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

The Data: Analysis Based on 100 Simulated Datasets

Data were simulated according to the assumptions of Linear Discriminant Analysis. The number of available predictors is $P = G_1 + G_2 + G_3$ where $G_1 = 28$ valid predictors (those with nonzero population coefficients given in Table 1), which include 15 relatively weak predictors (valid predictors with importance scores < .85), $G_2 = 28$ irrelevant predictors (named 'extra1' – 'extra28') uncorrelated with both the dependent variable and with the 28 valid predictors but correlated with each other, and $G_3 = 28$ additional irrelevant predictors ('INDPT1' – 'INDPT28'), each uncorrelated with all other variables. Correlations and variances mimic real data. We generated 100 simulated samples, each consisting of N=50 cases, with group sizes $N_1 = N_2 = 25$.

Table 1: True Linear Discriminant Analysis (LDA) Model Coefficients

Predictors	Unstandardized	Standardized*	Importance	Importance Rank
SP1	-9.55	-5.72	5.72	1
GSK3B	4.56	2.48	2.48	2
RB1	-3.82	-2.30	2.30	3
IQGAP1	3.35	2.13	2.13	4
BRCA1	-2.13	-1.36	1.36	5
TNF	2.24	1.32	1.32	6
CDKN1A	2.33	1.29	1.29	7
MAP2K1	2.75	1.20	1.20	8
MYC	-1.81	-1.19	1.19	9
EP300	-1.78	-1.15	1.15	10
CD44	1.85	1.03	1.03	11
CD97	1.44	0.92	0.92	12
SIAH2	1.15	0.87	0.87	13
MAPK1	1.64	0.79	0.79	14
RP5	1.94	0.76	0.76	15
S100A6	1.22	0.74	0.74	16
ABL1	1.44	0.73	0.73	17
NFKB1	1.22	0.70	0.70	18
MTF1	-1.01	-0.62	0.62	19
CDK2	1.20	0.61	0.61	20
IL18	-0.79	-0.56	0.56	21
PTPRC	-0.98	-0.53	0.53	22
SMAD3	-0.57	-0.35	0.35	23
C1QA	-0.29	-0.30	0.30	24
TP53	0.45	0.26	0.26	25
CDKN2A	-0.31	-0.23	0.23	26
CCNE1	-0.21	-0.19	0.19	27
ST14	-0.18	-0.14	0.14	28

*Standardized coefficient = Unstandardized coefficient multiplied by standard deviation of predictor

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Opening the Data File

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

To open the file, from the menus choose:

- File → Load Dataset
- Select 'LDASim.sav' and click Open to load the data

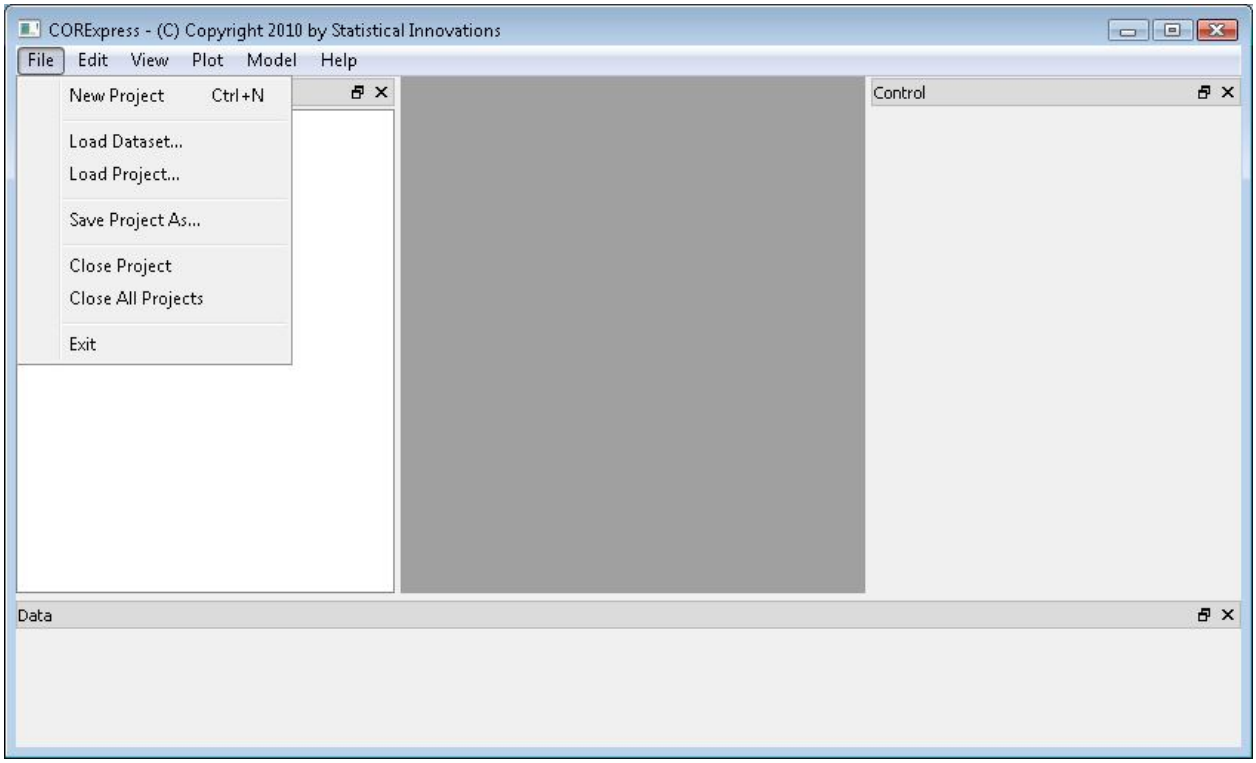


Fig. 1. File Menu

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

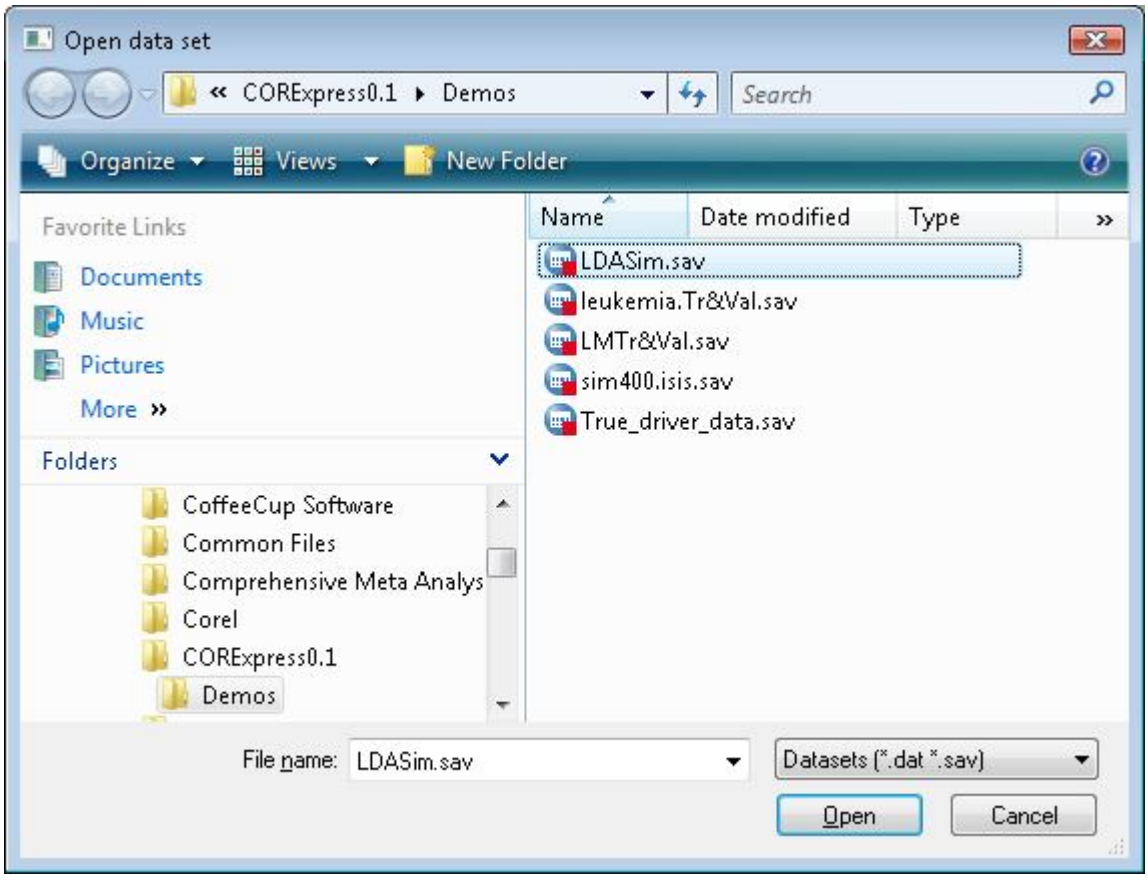


Fig 2. Loading a Dataset

Figure 3 shows the "LDASim" dataset loaded in the "Datasets" Outline Window on the left. The middle section contains the workspace (currently a dark gray box), where the "Model Output" appears following estimation of a CCR model. On the right is the "Model Control Setup" window, where model setup is done. The "Data" window on the bottom shows data from cases included in the dataset.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

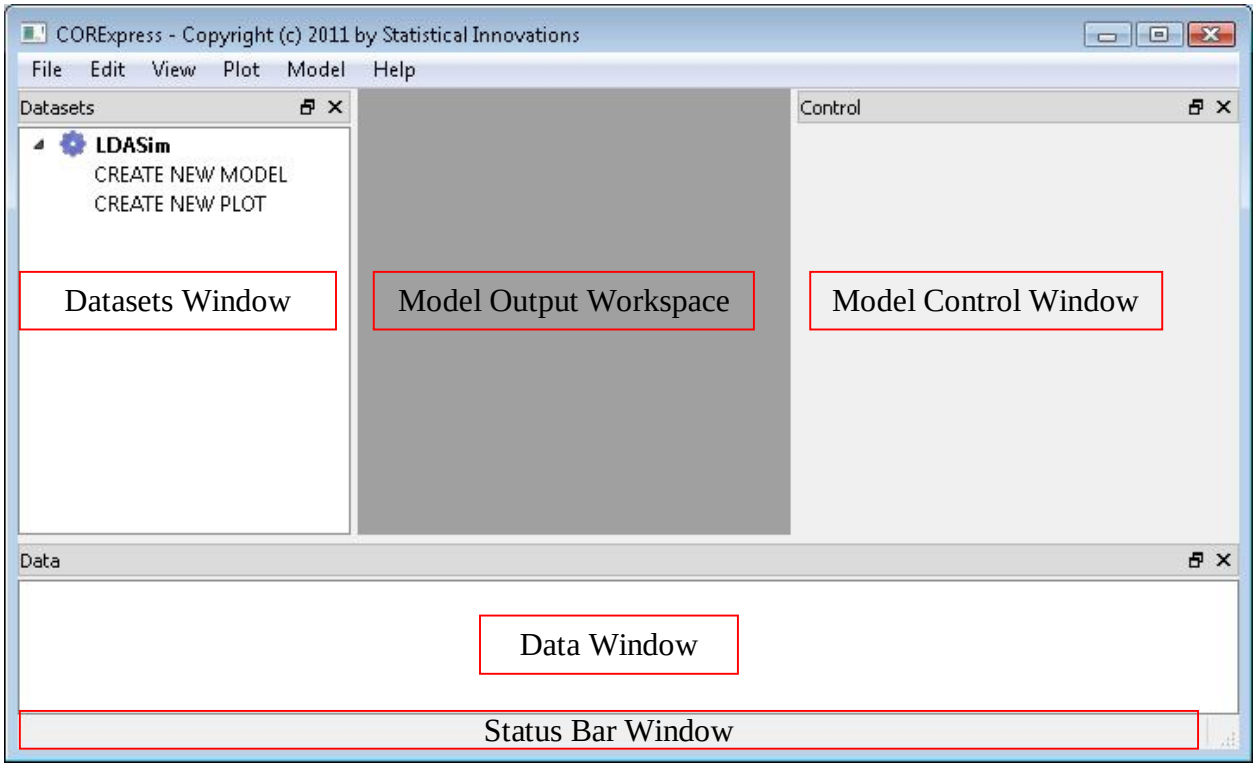
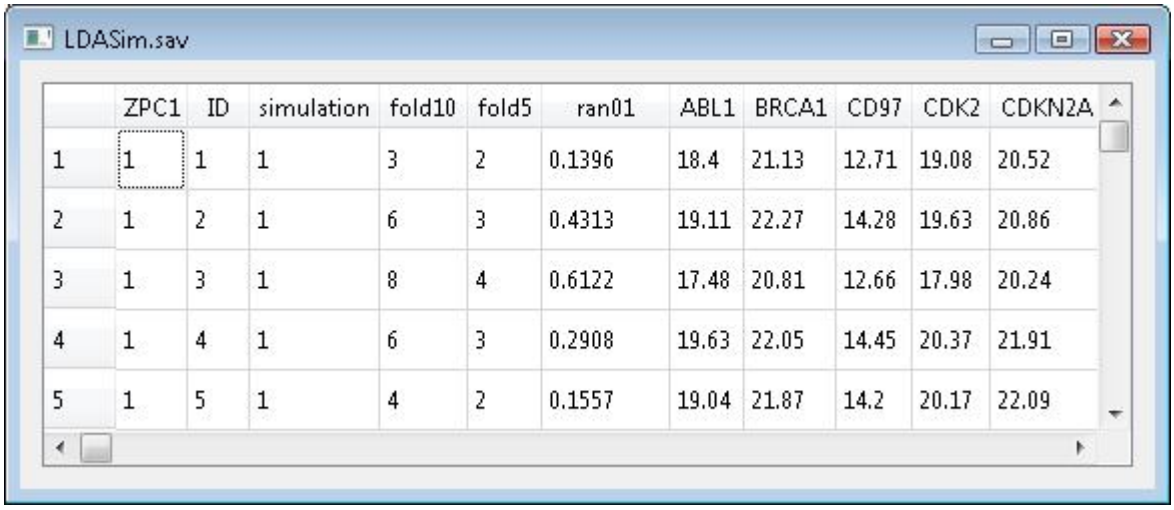


Fig. 3: COExpress Windows

You can view the complete dataset in a new window by double clicking on "LDASim" in the Datasets window.



	ZPC1	ID	simulation	fold10	fold5	ran01	ABL1	BRCA1	CD97	CDK2	CDKN2A
1	1	1	1	3	2	0.1396	18.4	21.13	12.71	19.08	20.52
2	1	2	1	6	3	0.4313	19.11	22.27	14.28	19.63	20.86
3	1	3	1	8	4	0.6122	17.48	20.81	12.66	17.98	20.24
4	1	4	1	6	3	0.2908	19.63	22.05	14.45	20.37	21.91
5	1	5	1	4	2	0.1557	19.04	21.87	14.2	20.17	22.09

Fig. 4: COExpress Dataset View

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Estimating a CCR Model

Selecting the Type of Model:

- Double click on "CREATE NEW MODEL" in the Workspace window under "LDASim"

Model setup options appear in the Control window.

Selecting the Dependent Variable:

- In the Control window below "Dependent", click on the drop down menu and select "ZPC1" as the dependent variable.

Selecting the Predictors:

- In the Control window below "Predictors", click and hold on "ABL1" and move the cursor down to "INDPT28" to highlight all 84 predictors. Click on the box next to "INDPT28" to select all 84 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 "ABL1" and move the cursor down to "INDPT28" to highlight all 84 predictors in the left box.
- Click on the ">>" box in the middle to select all 84 predictors and move them to the right box as candidate predictors.

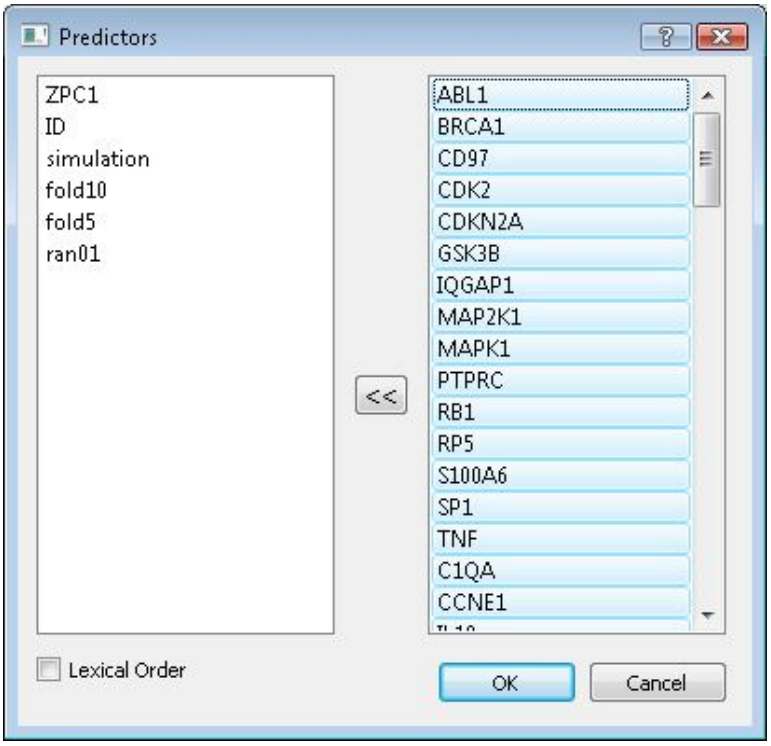


Fig. 5. Predictor Window

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Specifying the Number of Predictors to Include in the Model:

- 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"
- In the "Max # Predictors:" box, keep the default number, "20"

The estimation begins with all P=84 predictors in the model, and the CCR step-down procedure is applied, eliminating the weakest predictors until 20 remain. Since we will also activate the Cross-validation (CV) feature, it then accumulates the CV statistics and evaluates all models in the specified range 1-20. The model output displayed will be for the model with P* predictors, where P* is the one in the range 1-20 that achieves the highest cross-validation accuracy (CV-ACC). In the case of 2 or more values for P tied for best, the cross-validated Area Under the Curve (CV-AUC) will be used to break the tie, and if ties still remain, the smallest P* will be selected from among those tied.

CV information is provided for all models within the range, should you wish to estimate additional models containing a different number of predictors or change the specific predictors included in the preliminary model.

By default, at each step of the step-down procedure, the 1% of the predictors that are weakest (lowest importance coefficient) are excluded. If the check-mark to the left of "Remove by Percent" is removed, the weakest predictors are removed 1 at a time at each step.

If the CV feature were *not* activated, the step-down algorithm will eliminate the weakest predictors until the selected number (here '1") remains in the model. In this case, the Max # Predictors specified is ignored by the program.

Selecting the Number of Components:

- Under Options, click in the box to the right of "# Components", delete "4", and type "3".

Specifying the Model Type:

- Select CCR.lda to specify the LDA form of CCR

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Your Control window should now look like this:

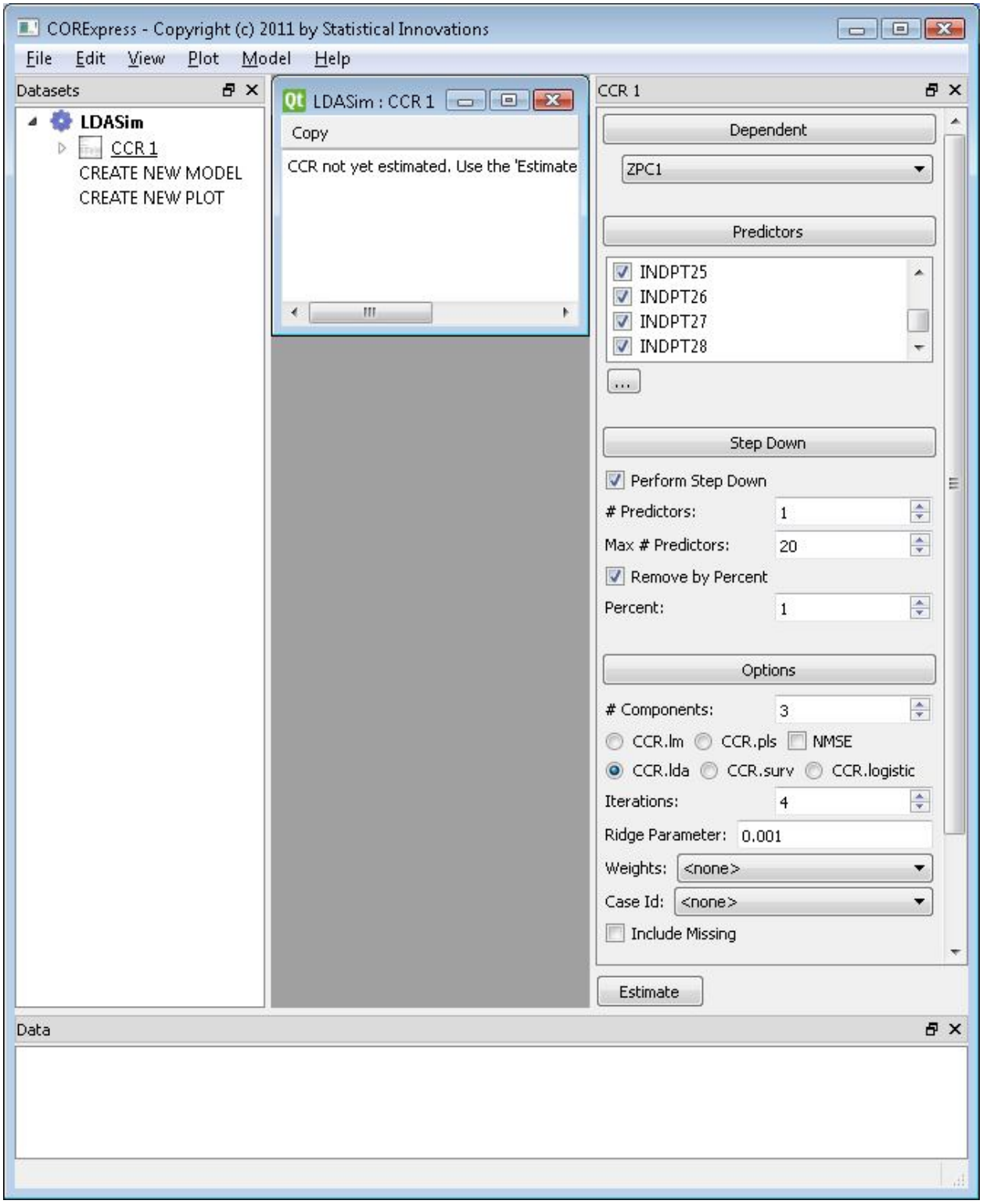


Fig. 6: Control Window

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Selecting the Training Sample:

- Click on "Validation" and options will appear for selecting training and validation sample cases.
- Under 'Training Subset', click on the "<select>" drop down menu and click on "simulation".
- Click on the "=" drop down drop down menu and click on "<".
- Click in the Training Subset numeric box and delete the number 0. Type "3".

Now, all records with simulation<3 will be selected as the Training sample. This selects the 100 cases in simulations 1 and 2 for the analysis sample, providing group sample sizes of $N_1 = N_2 = 50$.

Specifying the Validation Sample:

Unless otherwise specified, all cases other than those selected to be used in the training sample are automatically selected as the validation sample. This corresponds to the N=4,900 cases for which simulation>3.

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 "# Folds:" box, delete "10" and type "5"
- Click on the "<none>" Fold Variable drop down drop down menu and click on "fold5".

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. The folds are defined by the variable 'fold5' on the data file. If a fold variable is not specified, COREXpress assigns cases randomly to each fold, and if the 'Stratify' option is selected, each fold will have the same dependent variable distribution (or as close as possible). The variable 'fold5' is one implementation of stratified random assignment, exactly 10 cases being assigned to each fold, 5 from each group. Later, we will let the program assign cases to the folds.

M-fold cross-validation is a common technique used in datamining. The statistics CV-ACC and CV-AUC, are 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 statistics. Thus, the performance of the model is measured using cases not used at all in the development of the model.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Your Control window should now look like this:

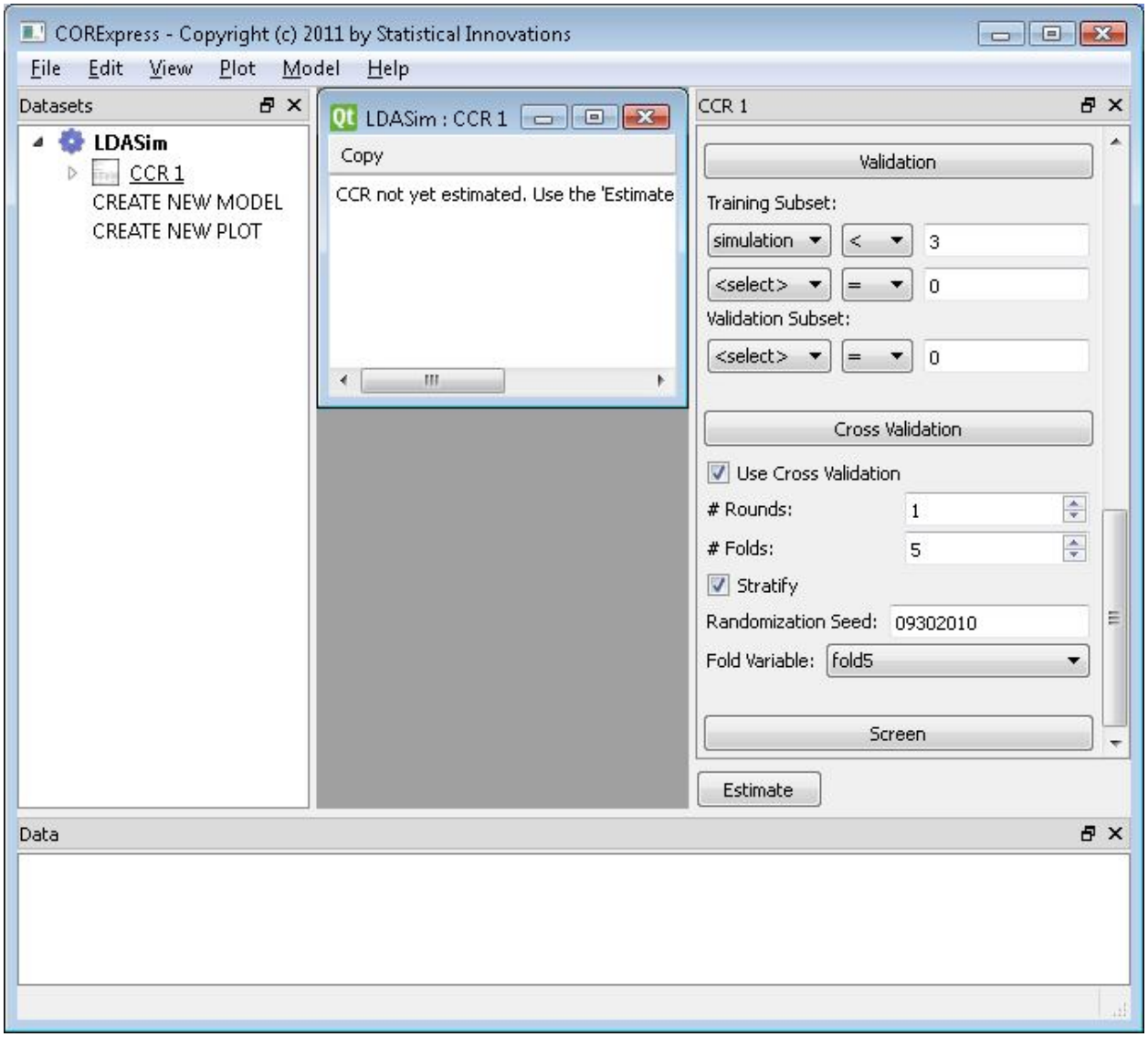


Fig. 7: Control Window

Estimate the Specified Model:

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

A new window containing the CV-ACC / CV-AUC Plot pops up, which summarizes graphically, CV results for predictors within the selected range.

View Model Output

Viewing CV-ACC / CV-AUC Plot:

- Click on the "COREXpress" window (CV-ACC / CV-AUC Plot)

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

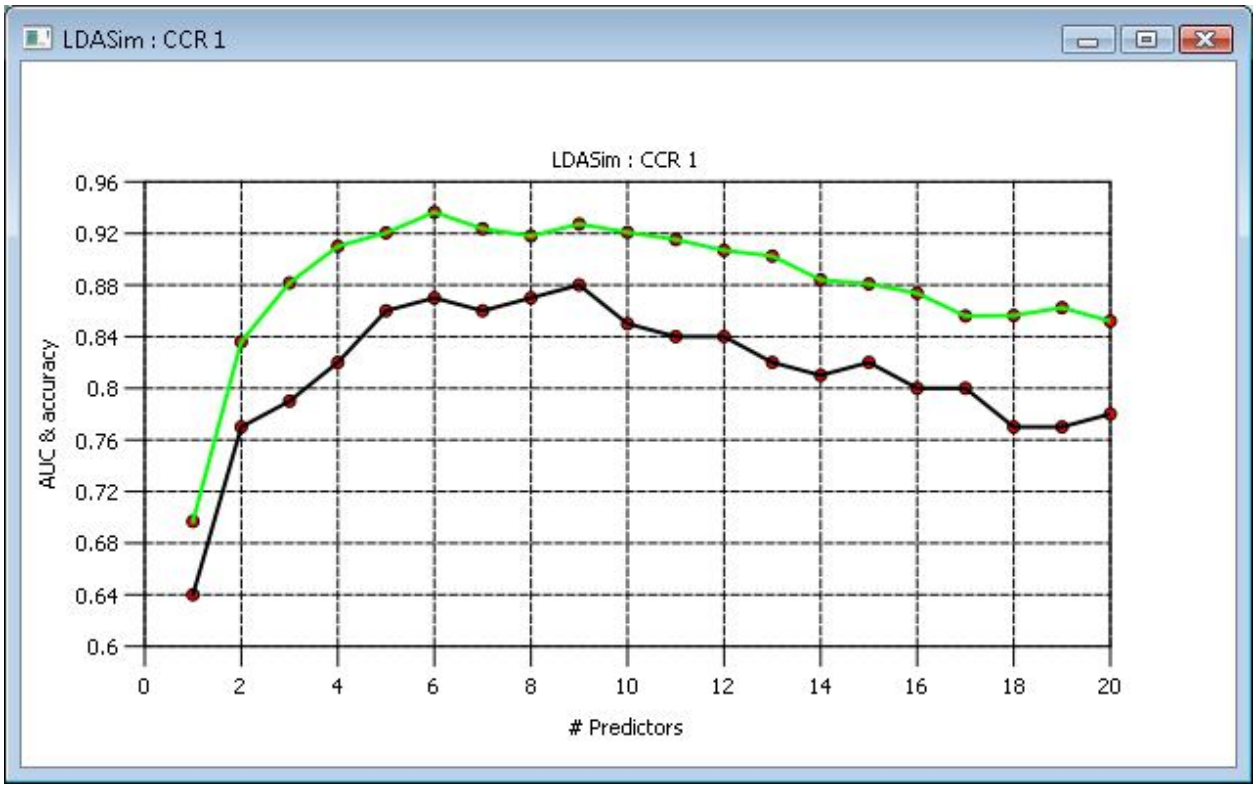


Fig. 8: CV-AUC and CV-ACC Plot

The CV-AUC and CV-ACC plotted in the graph corresponds to the cross-validation AUC and model accuracy based on the 3-component model for numbers of predictors P ranging from 20 down to 1. Given K = 3, the highest CV-ACC of .88 occurs with P* = 9 predictors. (As an exercise, you can repeat the estimation for other values of K, and confirm that the resulting models yield lower values for CV-ACC, which means that K*=3.) The performance is also better than that of stepwise discriminant analysis for this sample.

Note that when the algorithm steps down to P = 3 predictors, the model becomes saturated -- meaning K=P. Since it is not possible to estimate a model where K>P, for P< 3 COREXpress automatically reduces the number of components, maintaining a saturated model. Thus, for P> 2, K=3 components are maintained and for P< 3, K is reduced accordingly.

Viewing CV-ACC / CV-AUC Output:

- Click on the "LDASim : CCR 1" window (the Model Output Window) in COREXpress
- Scroll to the bottom of the "LDASim : CCR 1" window

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

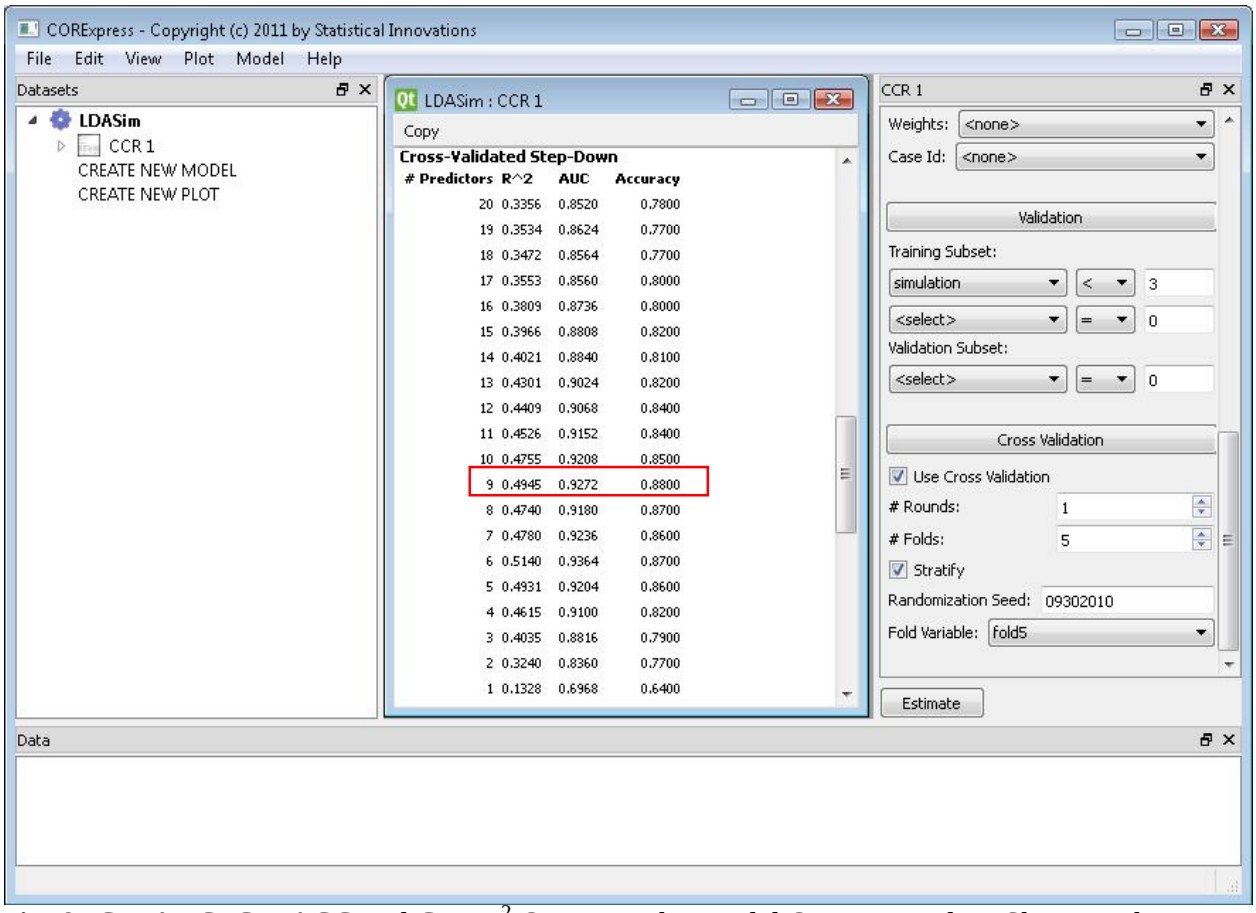


Fig. 9: CV-AUC, CV-ACC and CV- R^2 Output in the Model Output Window Showing the Highest CV-ACC Occurs with P*=9 Predictors

The cross-validation AUC (CV-AUC) is located at the bottom of the CCR 1 Model Output Window along with the CV-ACC for each number of predictors. By default, the model estimated and shown in the model output window is the one based on the tuned parameter value for P-- the one with P* predictors. Here, P* is the value for P with the highest CV-ACC among the eligible 3-component models. As mentioned above, if there were ties, P* is taken to be the one with the highest CV-AUC among those with the highest CV-ACC. For the example here, P*=9 which has the highest CV-ACC (no ties). It also turns out to be the value of P with the highest CV-AUC.

Viewing the 'Optimal' Model Output:

- Click on the "LDASim : CCR 1" window (the Model Output Window) in COREXpress
- Scroll to the top of the window to view the summary statistics, predictors included in the model and coefficients enclosed in the red box.

Loadings defining each component CC1, CC2, and CC3 as a linear combination of the predictors are listed beneath columns labeled 'CC1', 'CC2', and 'CC3', and each predictor

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

coefficient can be expressed as a weighted average of their loadings, the component weights being listed above the components. For example, the coefficient for predictor ABL1 is:

2.7184 = 1.3297*1.6961 + .4266*(-2.5205) + .8704*1.7675

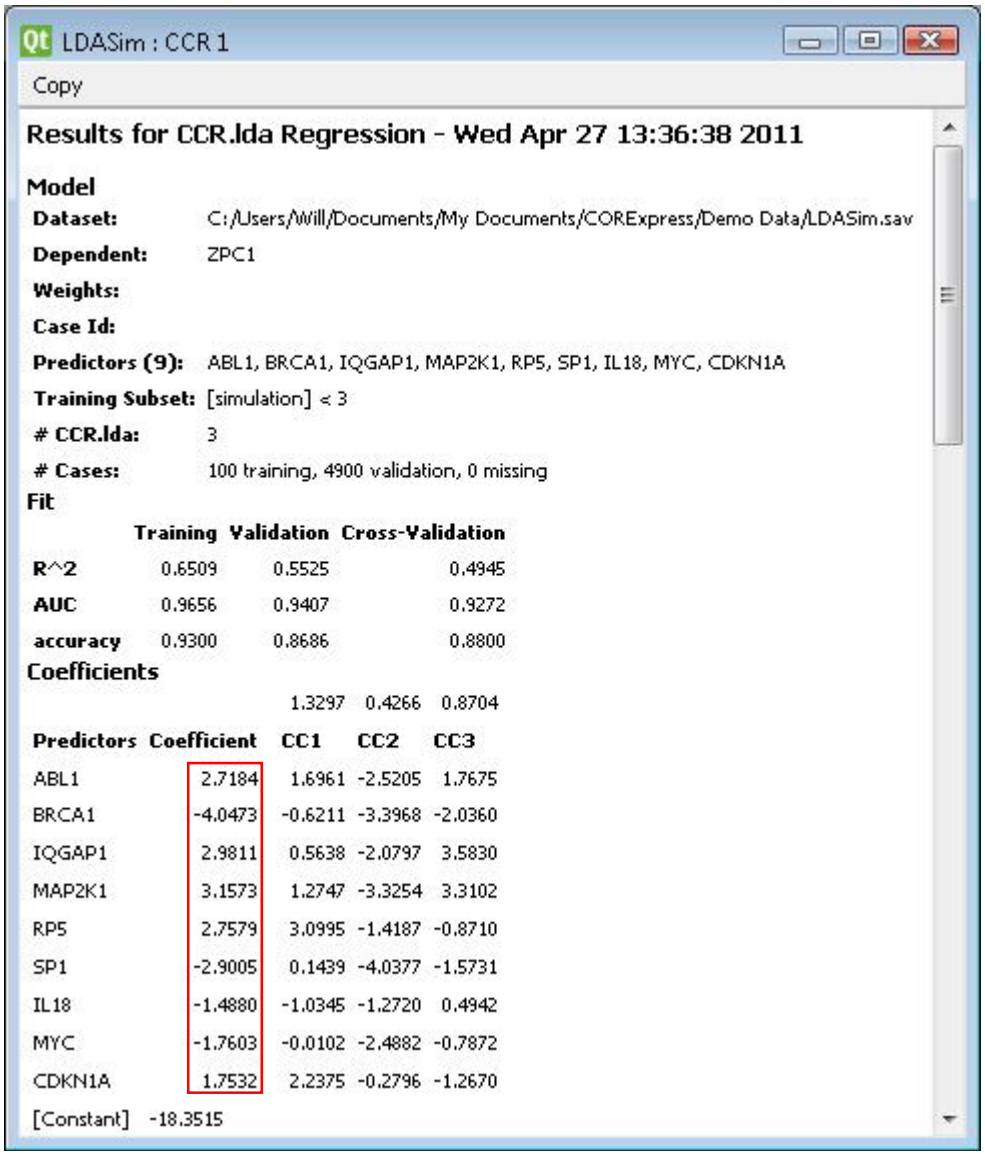


Fig. 10: Unstandardized Coefficients for K=3 in the Model Output Window

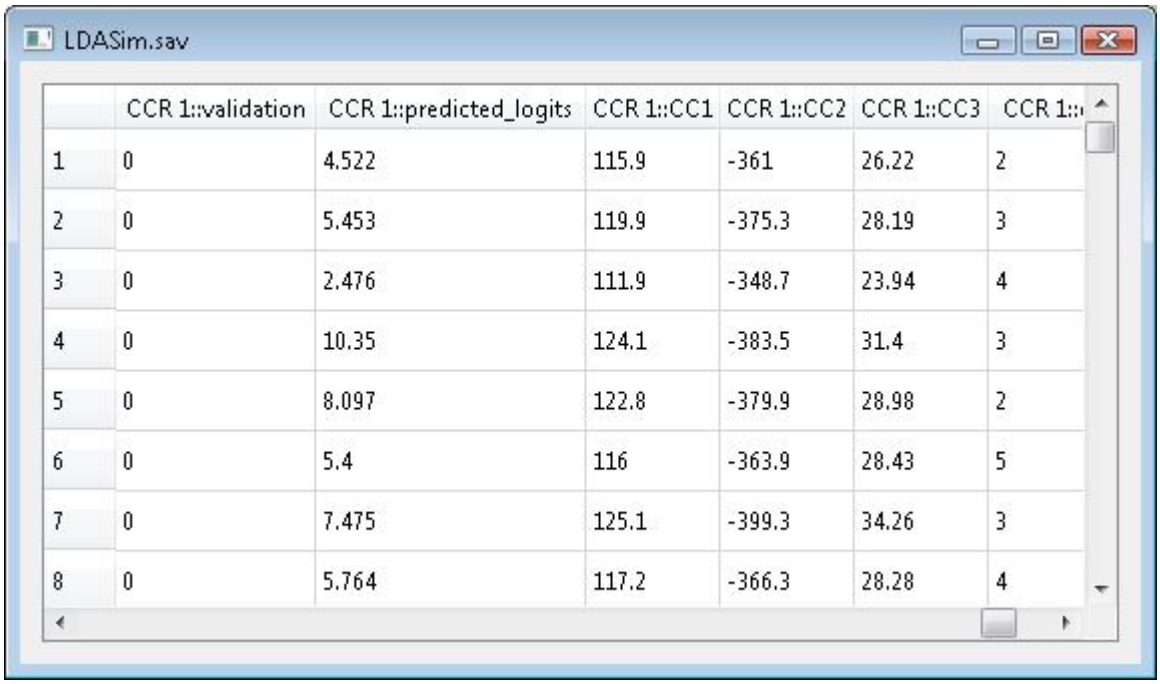
Note that for the Training, the CV-ACC=0.88 and for the Validation the ACC=0.8686.
Note that the validation accuracy and AUC are .8686, and .9407, close to the corresponding cross-validation quantities, CV-ACC = .88, and CV-AUC=0.9272. The results are quite good based on this sample -- the drop-off in performance from the Training to the Validation sample is fairly small, and the 9 predictors included in the model are all among the valid predictors.

Viewing the K Components and Predicted Scores on the Dataset:

- Close the Datafile window

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

- Double click on “LDASim” to re-open the Datafile window, which is now updated with some new variables including the predicted logit scores and the 3 components.
- Click on the Datafile window and scroll all the way to the right to view these newly constructed variables



	CCR 1::validation	CCR 1::predicted_logits	CCR 1::CC1	CCR 1::CC2	CCR 1::CC3	CCR 1::…
1	0	4.522	115.9	-361	26.22	2
2	0	5.453	119.9	-375.3	28.19	3
3	0	2.476	111.9	-348.7	23.94	4
4	0	10.35	124.1	-383.5	31.4	3
5	0	8.097	122.8	-379.9	28.98	2
6	0	5.4	116	-363.9	28.43	5
7	0	7.475	125.1	-399.3	34.26	3
8	0	5.764	117.2	-366.3	28.28	4

Fig. 11: 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 logit score for the 3-component model. If the folds were randomly generated, the fold assignments would also appear on this file. After estimating another model, this file will be updated further. To view the updated data file, first close the current data file window and then double click on "LDASim" again. The new data file window will now contain the scores for the most recently updated model.

To copy the predicted scores and any other variables from the data set window to the clipboard, click on the desired variables (or CTRL-click to select non-adjacent variables) and type the shortcut "CTRL+C"> A pop-up window asks whether to include the variable name in the copy. You can then paste the new variables into other programs, with or without the variable names.

Viewing the Training and Validation Interactive Plots:

Two interactive plots are available -- one for the training, the other for the validation data.

- Click on the drop down arrow next to “CCR 1” in the Datasets window
- Double click on “~ ROC Training”

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

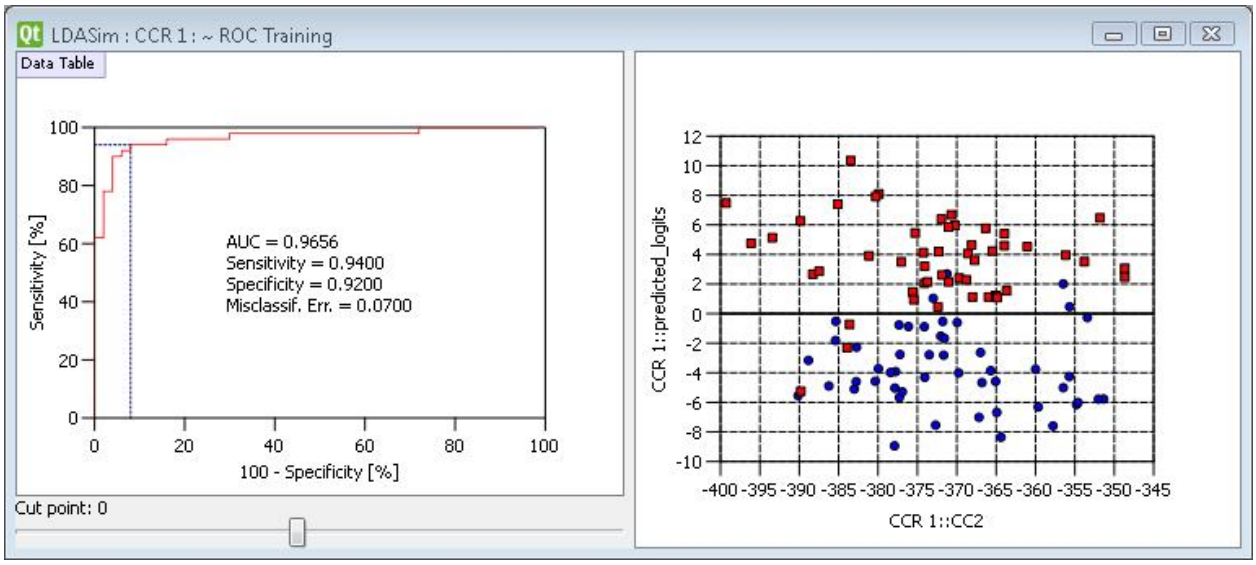


Fig. 12: Training Dataset ROC & Scatter Plot

Each point on the red ROC curve corresponds to a particular logit cut-point depicted by a horizontal reference line in the associated scatterplot. By default, the cut-point = 0, the predicted logit of zero corresponding to a predicted probability of .5. Cases above the cut-point (above the horizontal equal-probability reference line in the scatterplot) are predicted to be in group ZPC1=1, those below the cut-point being predicted to be in group ZPC1=0.

The specific point on the red ROC curve corresponding to this horizontal reference line is identified at the intersection of the dotted lines. The blue dotted lines define the sensitivity and 1-specificity for the given cut-point. For example,

- The sensitivity given on the vertical axis is .94, meaning that 94% of the red points in the scatter-plot are correctly classified as being above the reference line (i.e., above the cut-point).
- The specificity of .92 ('1-specificity' = .08) means that 92% of the blue points in the scatter-plot are correctly classified below the reference line (below the cut-point)

  The slider, located in the Control Window beneath the Cutpoint box can be used to see how the sensitivity and specificity changes with different cut-points. To increase the cut-point:
  - Position the curser on the slider, left-click and move it to the right to raise the reference line to the new cut-point of .7, so it now lies above 1 blue point that was incorrectly classified previously.

This blue point is now correctly classified, raising the specificity from 92% to 94%. However, 1 red point that was correctly classified previously is now incorrectly classified (it is now below the new reference line), and the sensitivity is reduced from 94% to 92%.

The blue dotted line in the ROC plot shifts to show the updated position on the ROC curve, and the sensitivity and specificity quantities are updated accordingly.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

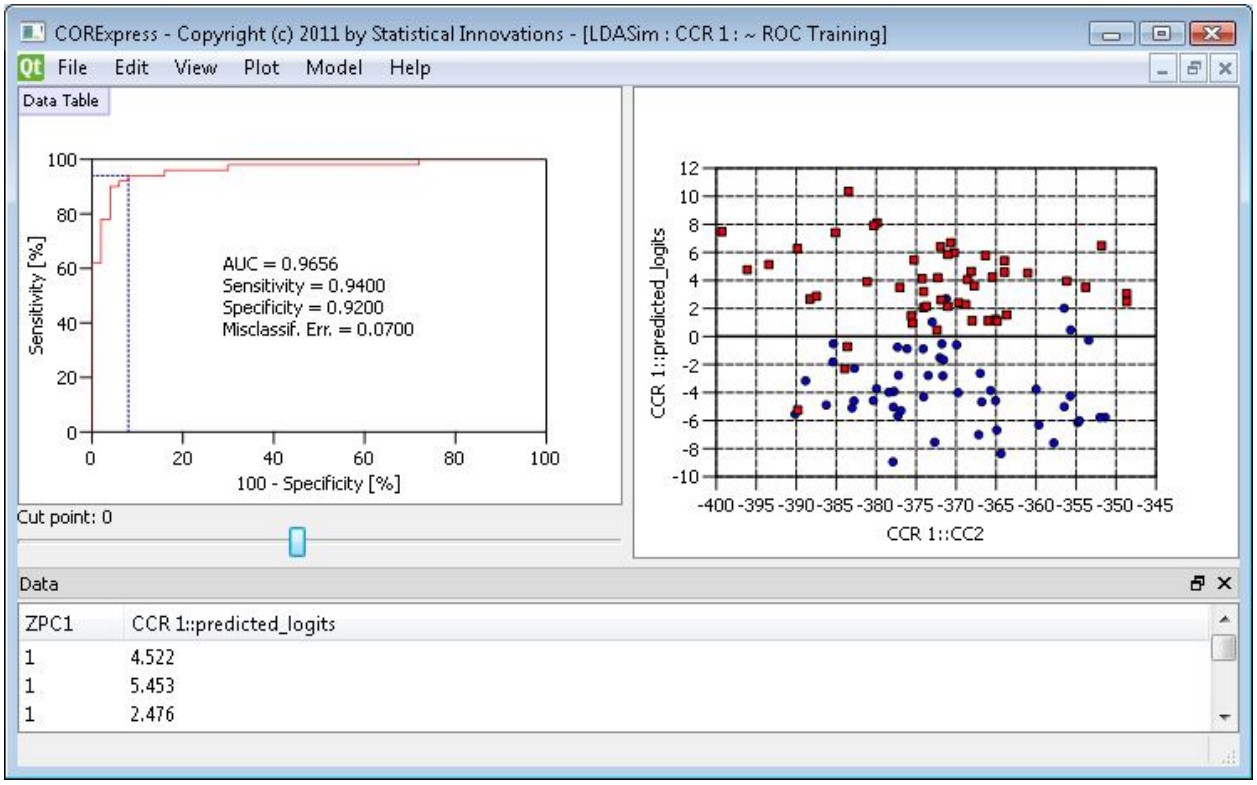


Fig. 13. Training Dataset ROC & Scatter Plot

A similar plot is available for the validation data. Since there are N=4,900 cases in the validation sample, this window will take longer to open, and due to the large number of red and blue points, it is not as easy to interpret.

- Double click on "~ ROC Validation" to open the Validation plot

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

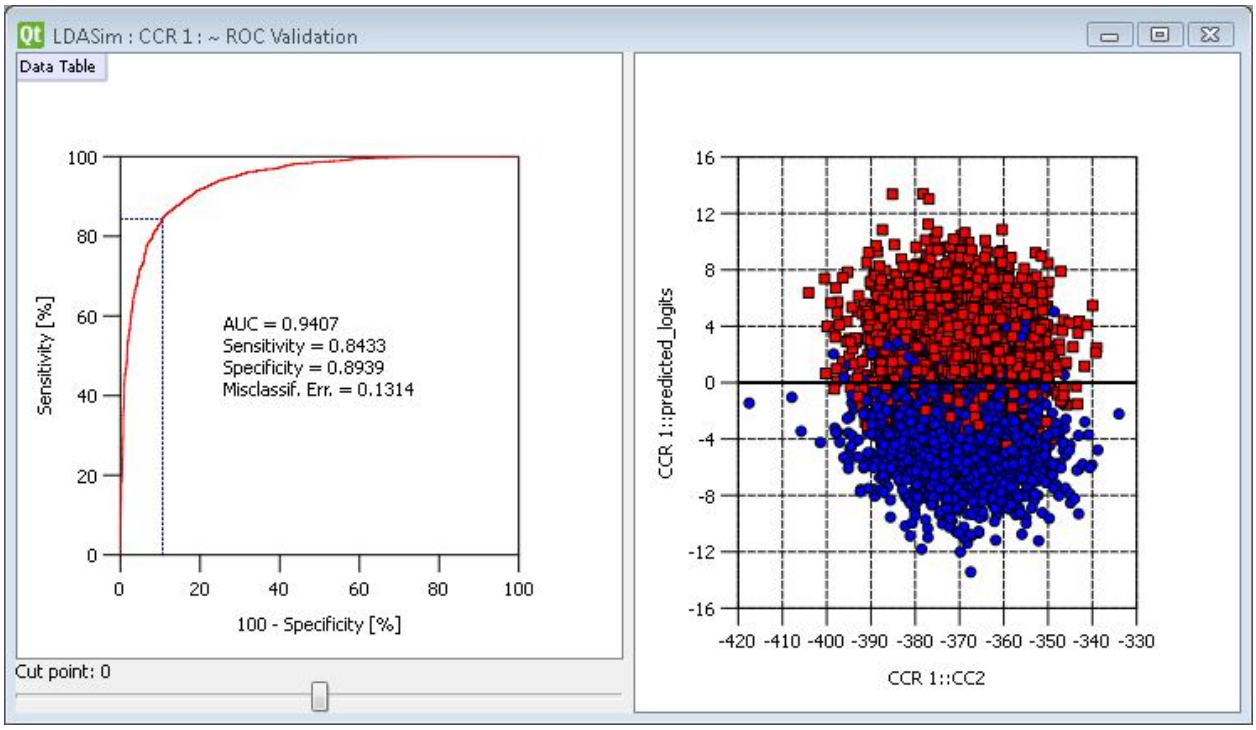


Fig. 14: Validation Dataset ROC & Scatterplot

To reduce the number of points to make it easier to visualize:

- Click on the plot to make it active

This changes the Control Window settings so it can be used to modify the active plot.

We will now change the Validation plot to show only the subset of cases in simulation=2:

Selecting a Subset of the Cases for the Validation Plot:

- Click on the "Subsets" box in the "~ ROC Validation" Control window
- Click on the "CCR 1 : Validation" box and select the variable named 'simulation'
- Click in the Corresponding Subset numeric box and type "2"
- Click the "Update" button.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Your screen should now look like this:

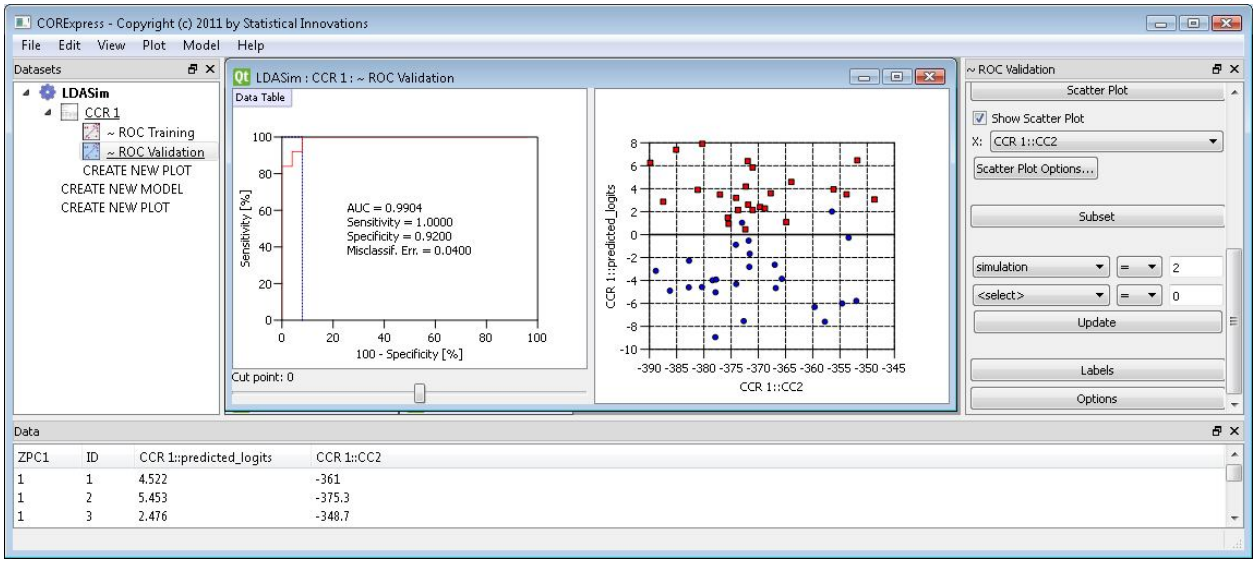


Fig. 15: Validation Dataset ROC & Scatterplot with simulation=2

Alternatively, the variable 'ran01' on the file can be used to display a random subset of cases. 'Ran01' consists of random numbers between 0 and 1. For example, by specifying 'ran01< .1', the plot will be updated to show only 10% of the validation sample cases.

Notice that for each case, the predicted logits and other variables are provided in the data window below the plots. Clicking on a point in the plot (or highlighting several points) identifies the associated cases in the data window. Similarly, selecting a case in the data window highlights the associated point in the plot.

Additional variables can be displayed in the data window by selecting the desired variables from the Project settings menu.

To open the Project Settings menu:

- Right click on 'LDASim' at the top of the Datasets Window.
- Select "Project Settings"

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

The following window appears:

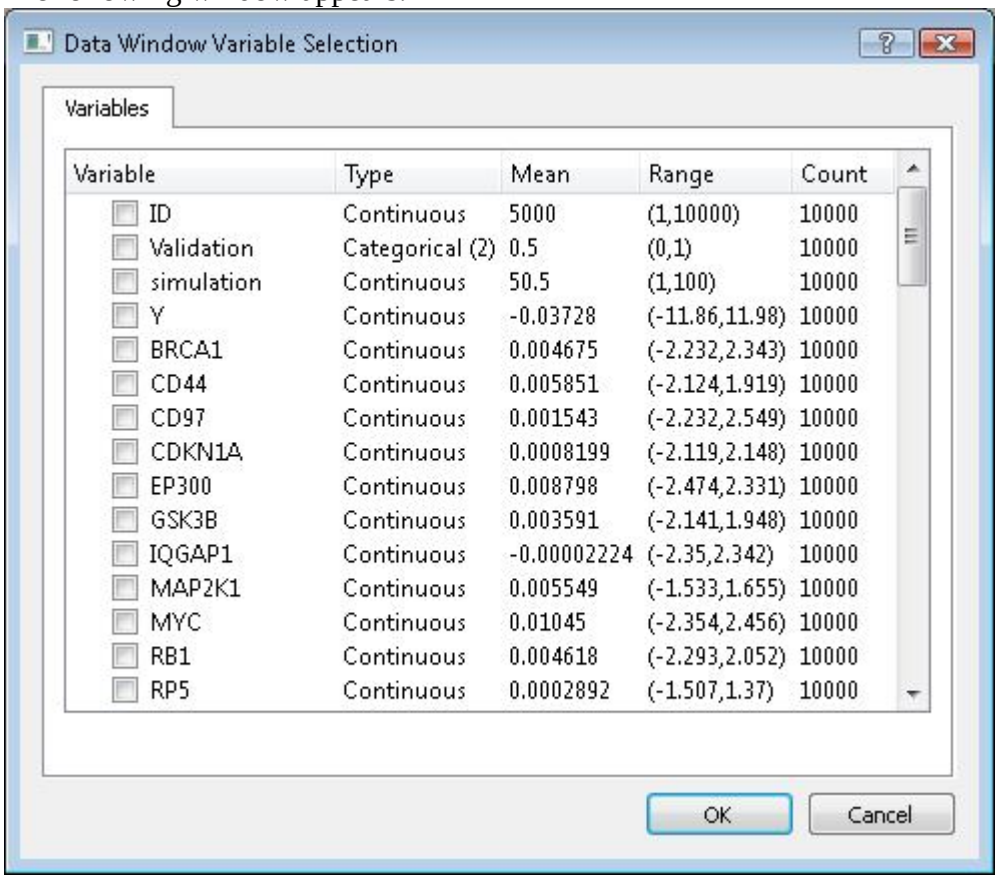


Fig. 16. Project Settings Menu Option: Data Window Variable Selection

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

A wide range of traditional plot options for linear regression are also provided. To see these, double click “CREATE NEW PLOT”

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

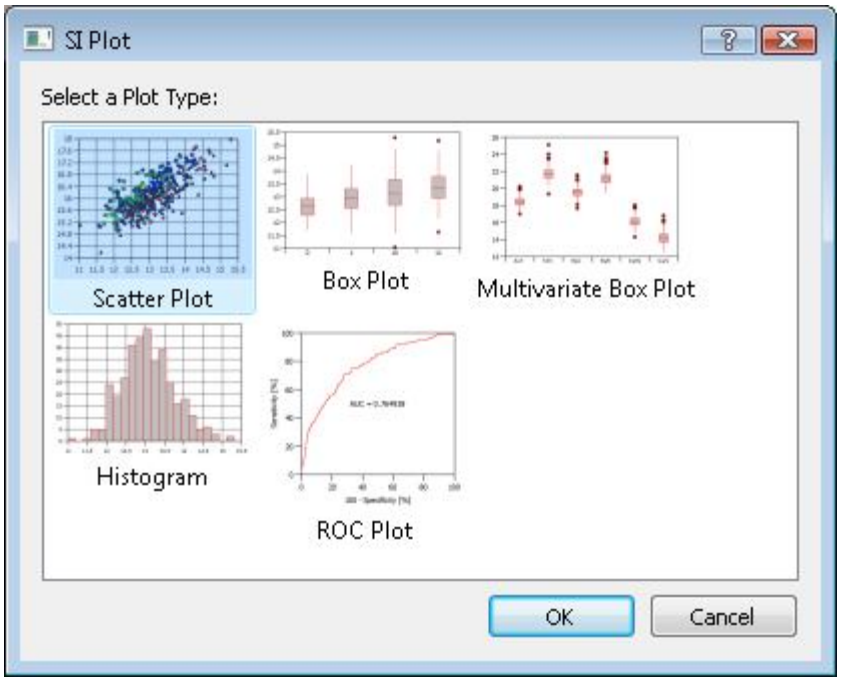


Fig. 17. Create New Plot Options

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.

Analysis of Simulated Sample #1 (N=50)

In our first example, we analyzed data based on a sample of size N=100, by limiting the training data to cases in simulated samples #1 and #2. Now, we will reduce the sample size further, using only simulation #1 as our analysis file. (As an exercise, these analyses can be repeated for each of the other 99 simulated samples on the file.)

Specifying the Training Sample:

- Double click on 'CCR' in the Datasets window to make the model active.
- In the Control Window, click on 'Validation' and options appear for selecting training and validation samples.
- Click on the "<" drop down drop down menu and click on "=".

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

- Click in the Training Subset numeric box and delete the number 3. Type “1”.

Now, all records with simulation=1 will be selected as the Training sample, providing group sample sizes of $N_1 = N_2 = 25$. (Note: For comparability with our earlier example, we use the same 5-fold CV assignments, as specified by the variable ‘fold5’. (The number of folds, 5, was selected because the group sample size of 25 is evenly divisible by 5.)

Selecting the Number of Components:

- Under Options, click in the box to the right of “# Components”, delete “3”, and type “6”, to specify K=6 components. K=6 turns out to yield the highest (CV-ACC, CV-AUC) combination for this sample.

Naming the Model:

- Right-click on the current model name CCR1
- Select ‘Rename’ to enter into EDIT mode
- Type ‘Sim1.K6’

Estimate the New Model:

- Click the “Estimate” button

View Model Output

Viewing CV-ACC / CV-AUC Plot:

- Click on the "COREXpress" window (CV-ACC / CV-AUC Plot)

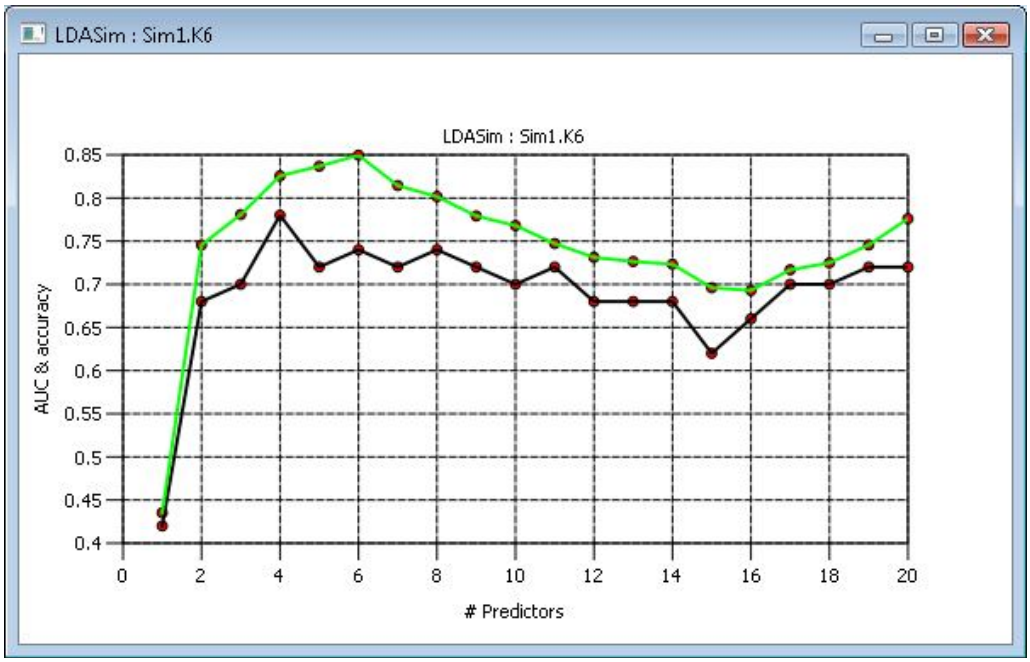


Fig 18. CV-AUC and CV-ACC Plot

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

The CV-ACC and CV-AUC plotted in the graph correspond to the cross-validation accuracy and area under the ROC curve based on the 6-component model for number of predictors P ranging from 20 down to 1.

As the number of predictors reaches P=6, the model becomes saturated -- meaning K=P. For P<6, COREXpress automatically reduces the number of components, maintaining a saturated model. Thus, for P>5, K=6 components are maintained and for P<6, K is reduced accordingly. Here, the tuned value is P*=4, so K is also reduced to 4 components (reported as '#CCR.lda: 4' in Fig. 20), yielding CV-ACC=.8.

Viewing CV-ACC / CV-AUC Output:

- Click on the "LDASim : CCR 1" window in COREXpress
- Scroll to the bottom of the "LDASim : CCR 1" window

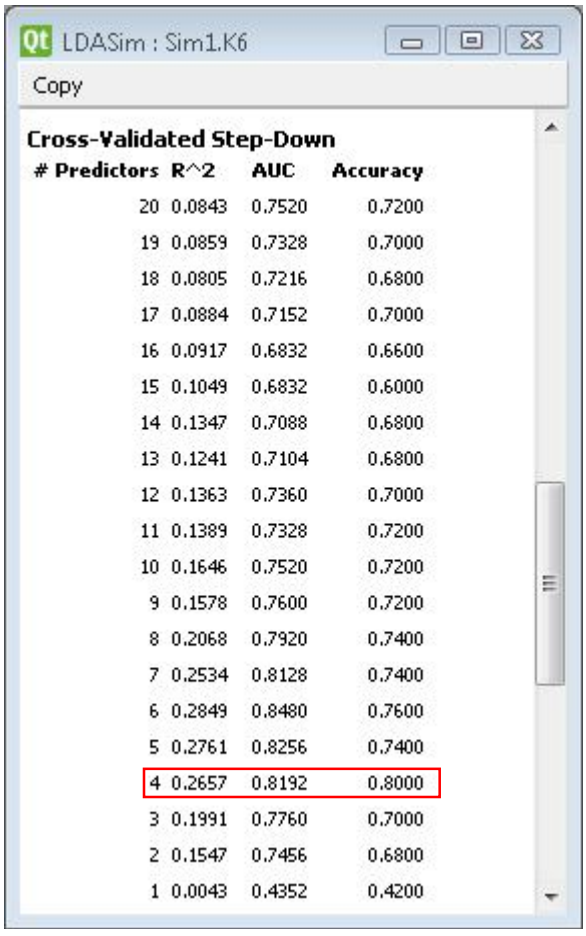


Fig. 19. Cross-Validation Step-down Output

The cross-validation accuracy CV-ACC is located at the bottom of the CCR 1 Model Output Window along the AUC (CV-AUC) and CV-R^2 for each number of predictors. (CV-R^2 is used as the primary statistic for CCR linear regression models which involve a continuous dependent variable.) As explained earlier, the results shown in the model output window is for the 'tuned'

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

model -- the one with P* predictors, where P* is the value for P with the highest CV-ACC. (In the case of ties, the tuned number of predictors P* is taken to be the one with the highest CV-AUC among those with the highest CV-ACC.) For the example here, the results are shown for the model with P*=4 which is the only one with CV-ACC as high as .8.

Since the model with 4 predictors is saturated (K=P=4), it is equivalent to a LDA model. In particular, this tuned CCR model turns out to be a LDA model with 4 valid predictors, all 4 being among the *valid* predictors.

This model validates very well, as the corresponding accuracy obtained by applying this model with the particular coefficients (and cut-point) estimated based on the N=50 training cases to new cases on the independent large validation (test) sample of N=4,950 is .802.

We will now show that the coefficients reported by COREXpress for this model are in fact identical to that obtained from Fisher's LDA (which is always the case since saturated CCR models are equivalent to the traditional regression models, LDA in this case). The table below computes the logit coefficients directly from the output obtained from linear discriminant analysis:

Table 2. Logit Coefficients for the Saturated CCR.lda Model
Match Coefficients Obtained from Fisher's LDA

Classification Function Coefficients for CCR Model			
	ZPC1		Logit
	0	1	Coefficients
BRCA1	74.10	69.50	-4.60
CD97	-24.78	-20.67	4.11
IQGAP1	-88.09	-81.44	6.66
SP1	82.11	76.26	-5.85
(Constant)	-698.64	-652.28	46.36
Fisher's linear discriminant functions			
		Val-ACC	0.802

Traditionally, with a small sample of N = 50 and P = 84 predictors, the stepwise LDA option would typically be used. For purposes of comparison, Table 3 below reports the results obtained from stepwise LDA. This results in a model with 6 predictors, 2 of which are irrelevant predictors, which are mistakenly included in the model. The resulting accuracy of this model, again estimated using the large test sample, is 77.8%, lower than the 80.2% accuracy obtained for the CCR model by COREXpress.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Table 3. Logit Coefficients Obtained from Stepwise LDA

Classification Function Coefficients for Stepwise LDA			
	ZPC1		Logit
	0	1	Coefficients
BRCA1	82.21	75.49	-6.72
IQGAP1	-89.34	-78.89	10.45
SP1	31.49	25.39	-6.10
CDKN1A	41.85	46.86	5.01
INDPT9	6.11	2.41	-3.70
INDPT23	7.73	5.68	-2.05
(Constant)	-873.22	-861.75	
			11.47
Fisher's linear discriminant functions			
		Val-ACC	0.778

To obtain the standard error for CV-ACC:

Since CV-ACC is the primary criteria used by CCR to determine the number of predictors in this model, we might be interested in knowing the standard error of this statistic. To compute the standard error, we need to re-estimate the model with more than 1 round of M-folds. We will allow COREXpress to randomly assign cases to 10 different sets of 5-folds. To request R = 10 rounds of 5-folds:

- In the Cross Validation section in the Control Window, change the "# Rounds" from 1 to 10
- Click on the drop down menu for "Fold Variable" and select "<none>"
- Click "Estimate"

Recall that previously based on a single round of 5-folds, the assignments being defined by the variable 'fold5', achieved CV-ACC = .80 for P*=4. In this analysis, with 10 rounds of 5-folds, we obtain CV-ACC = .7420 with standard error of .0503 for the model with P=4 predictors. The CV step-down output reported in Fig. 20 show that the resulting CV-ACC is again highest for this model with the 4 valid predictors.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

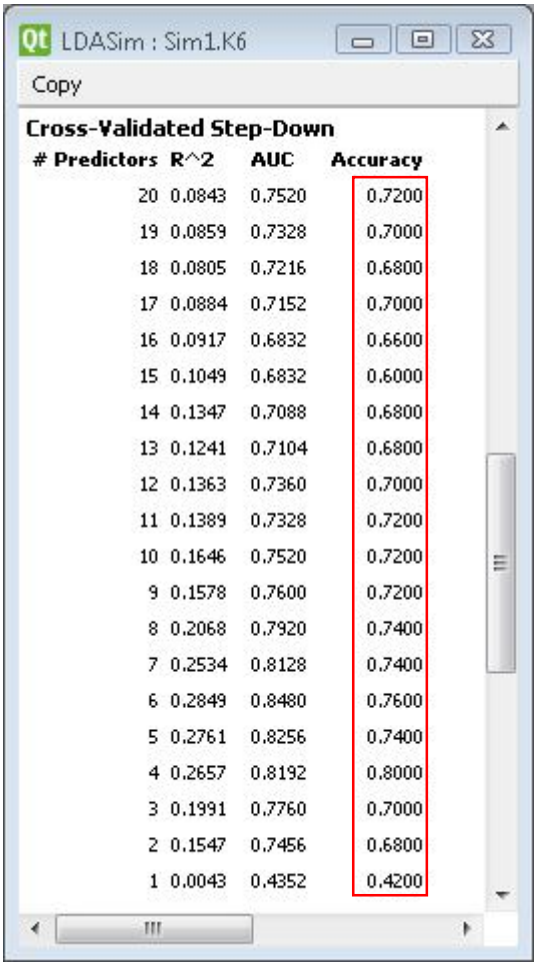


Fig. 20. Cross-Validation Step-down Output

The CV-ACC values reported in Fig. 20 are averages of the CV-ACC obtained from the 10 rounds, and the associated standard errors are computed as the standard deviation of these 10 values. This output is used to determine P*, the value of P yielding the highest average CV-ACC.

With multiple rounds of M-folds it is also possible to compute an additional informative statistic. From each round, COREXpress records the maximum CV-ACC, among the eligible values for P. For example, for round 1, the max CV-ACC may occur with P=7 predictors, while for round 5, it may occur with P=4 predictors. We then compute the average and standard deviation for the maximum CV-ACC, and report it in the Model Summary Results shown at the top of the Model Output Window. This is reported in Fig. 21, yielding CV-ACC = .7680, with standard error of .0391.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

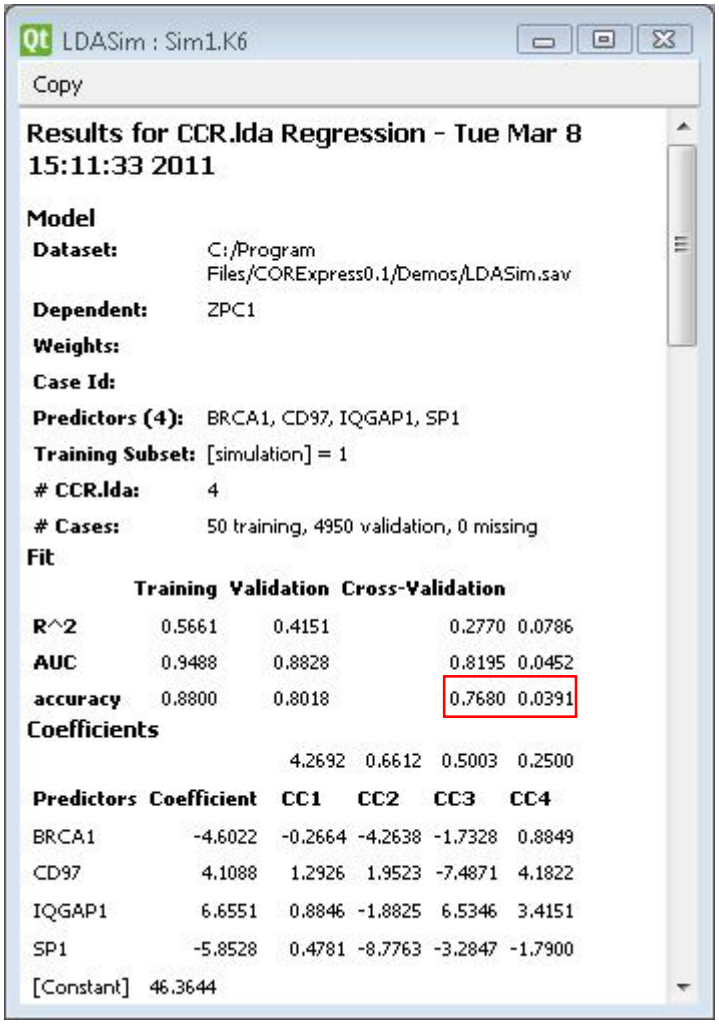


Fig. 21: Results from CCR4, Obtained based on 10 Rounds of 5-Folds
CV-ACC = 0.7680 with Associated Standard Error = 0.0391

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 "LDASim"
- Click "Save" to save the project.

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

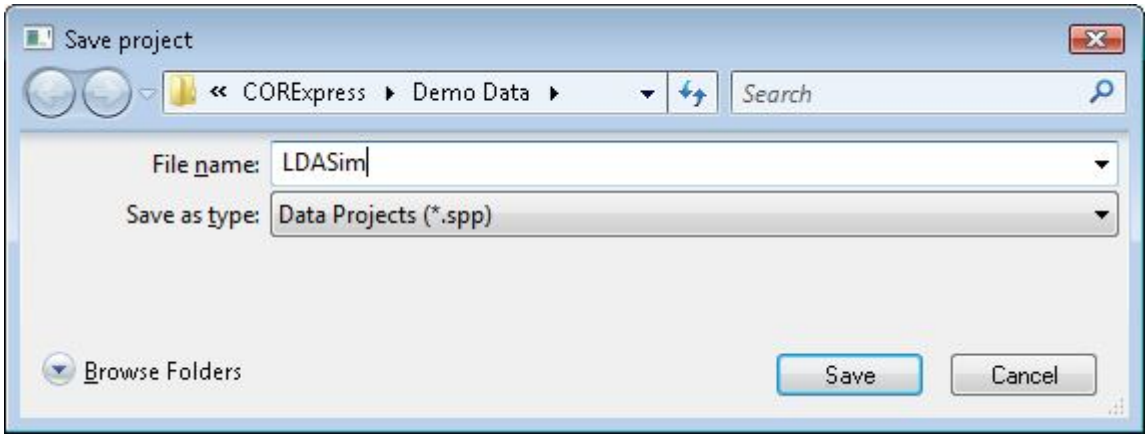


Fig. 22: Saving a Current Project

Viewing Model Specifications & Output from Previously Saved Project

Opening the Previously Saved Project:

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

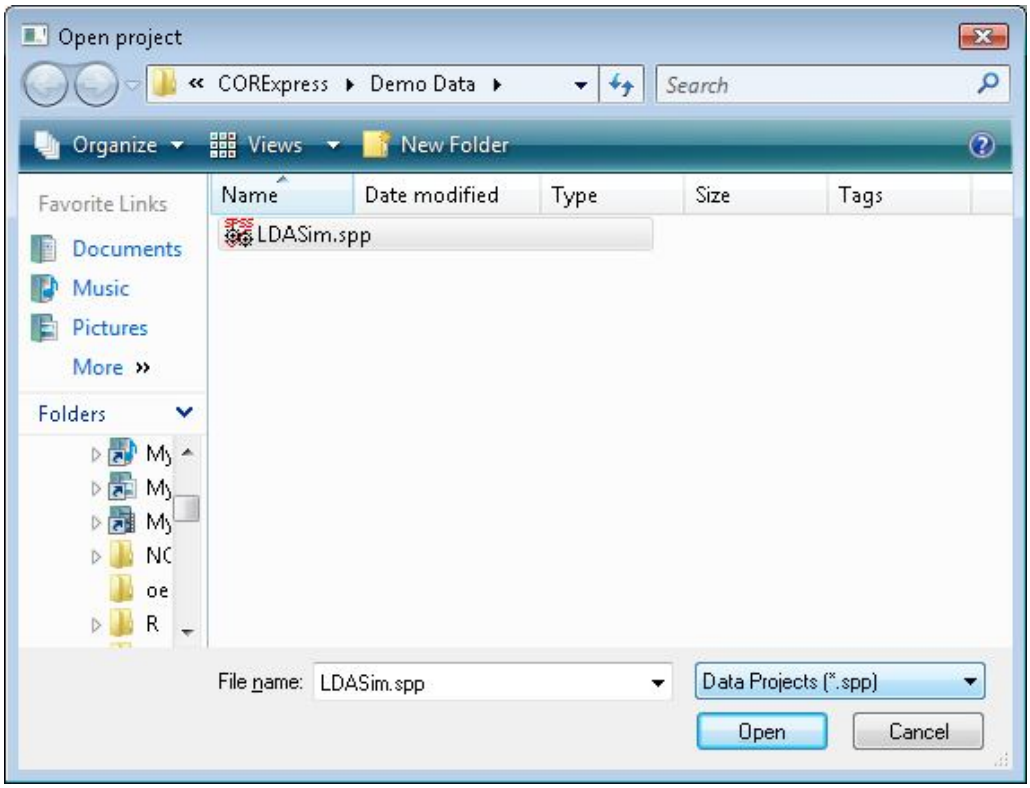


Fig. 23: Loading a previously saved project

Copyright © 2011 by Statistical Innovations Inc. All rights reserved. (version: 4/27/2011)

Viewing Model Specifications & Output from Previously Saved Project

Opening the Model Specifications for the Saved Project:

- Double click on "Sim1.K6" in the Datasets 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 "LDASim" from the Datasets window
- Scroll to the right to see that COREXpress automatically saves K Components and Predicted Scores from the previously generated runs.

Viewing the ROC and Scatter Plots from the Previously Saved Project:

- Click on the drop down arrow next to "Sim1.K6" in the Datasets window
- Double click on "~ ROC Training"
- Confirm that the ROC and Scatter Plots were saved from the previously generated runs.