Session 2

LDA Tutorial

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

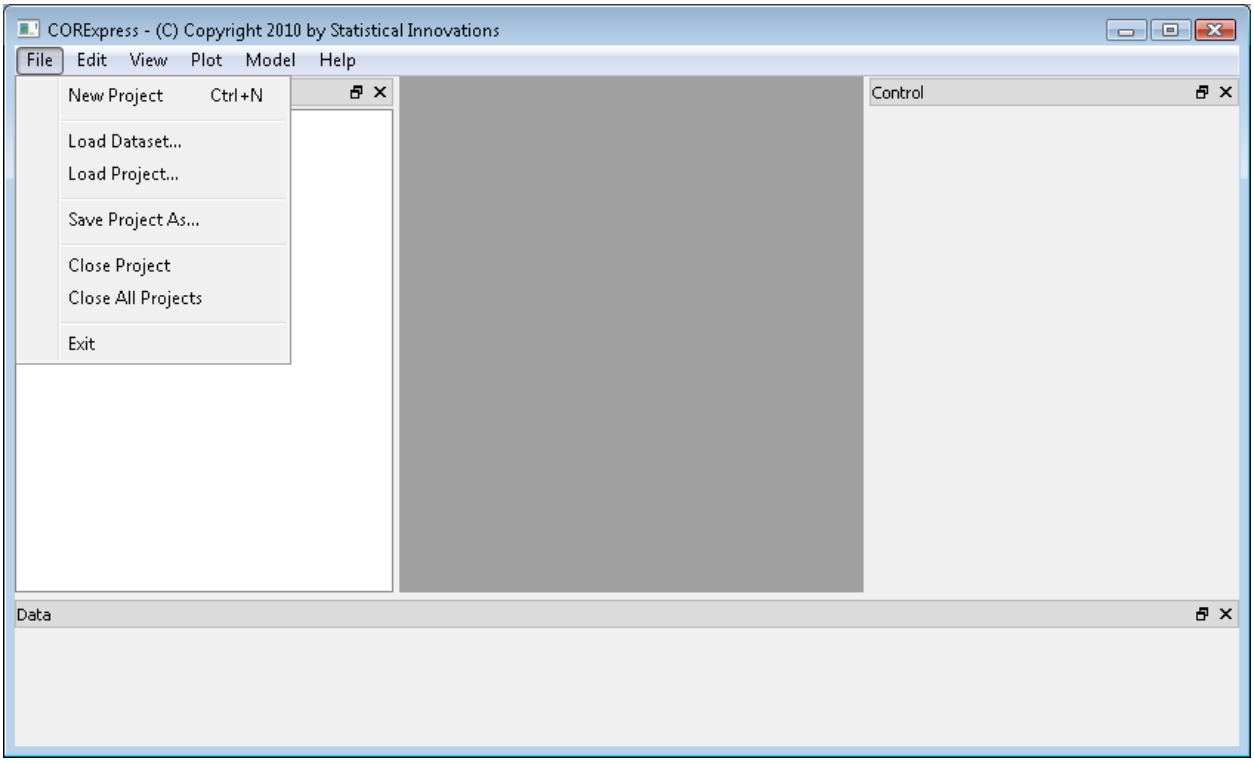


Fig. 1: File Menu

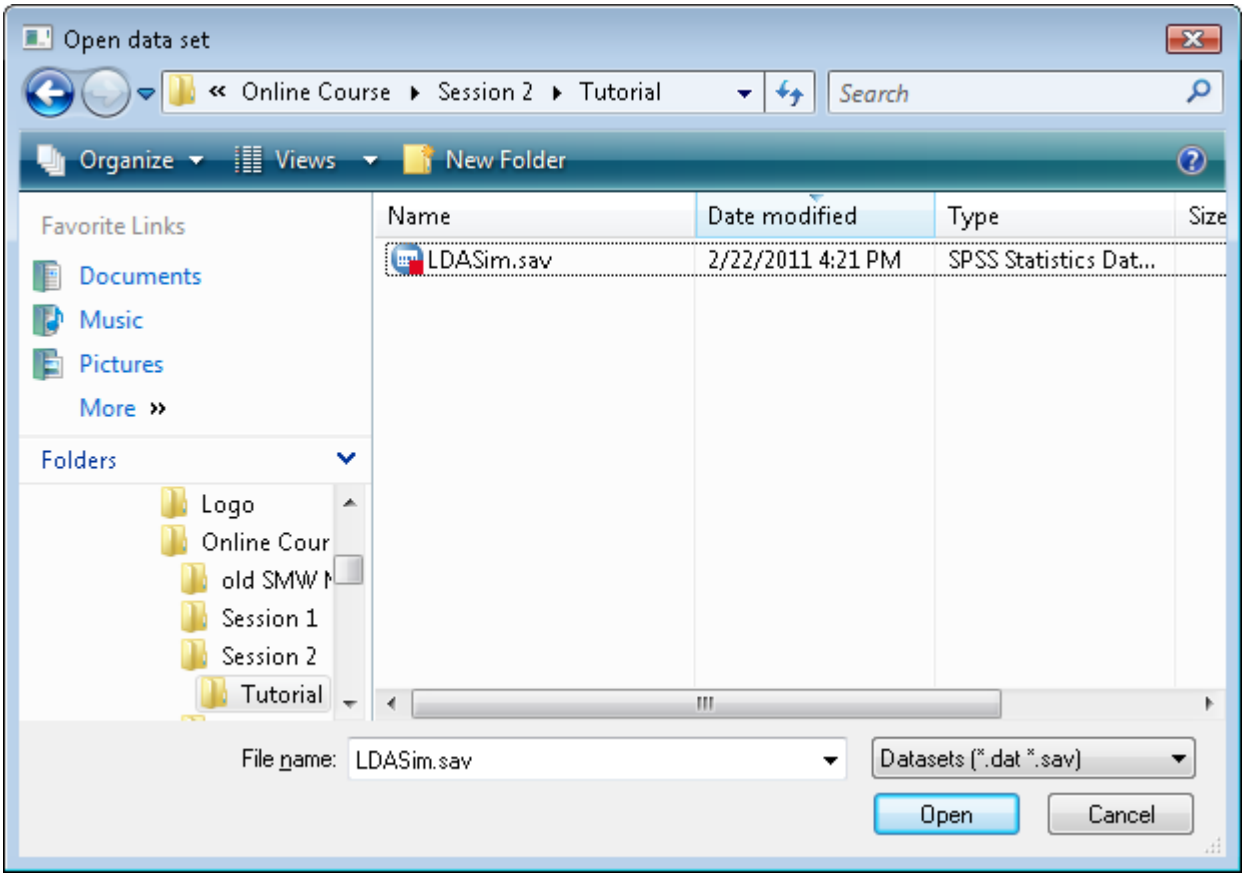


Fig. 2: Loading a Dataset

You will now see the "LDASim.sav"" dataset loaded in the "Datasets" Outline 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 Setup" window, where models can be specified and graphs can be updated. The "Data" Window on the bottom shows various data from the dataset.

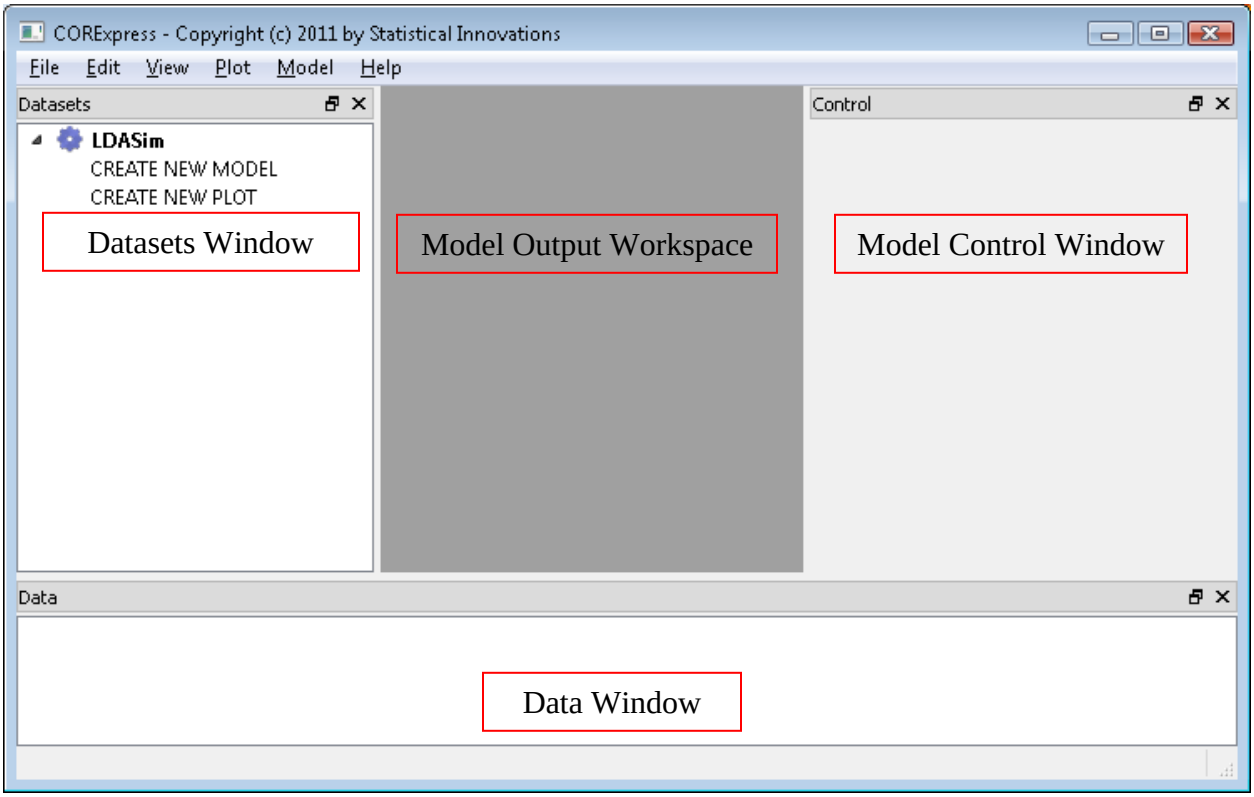
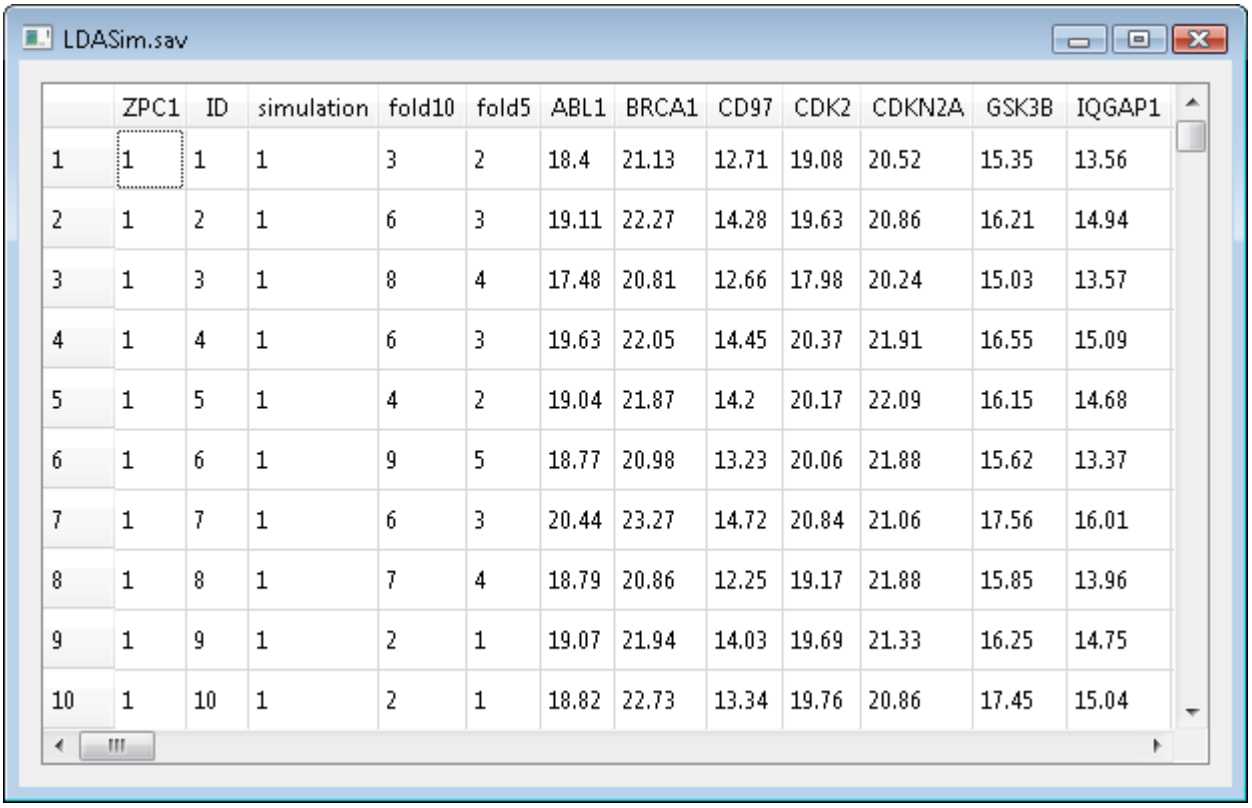


Fig. 3: COREXpress Windows

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



	ZPC1	ID	simulation	fold10	fold5	ABL1	BRCA1	CD97	CDK2	CDKN2A	GSK3B	IQGAP1
1	1	1	1	3	2	18.4	21.13	12.71	19.08	20.52	15.35	13.56
2	1	2	1	6	3	19.11	22.27	14.28	19.63	20.86	16.21	14.94
3	1	3	1	8	4	17.48	20.81	12.66	17.98	20.24	15.03	13.57
4	1	4	1	6	3	19.63	22.05	14.45	20.37	21.91	16.55	15.09
5	1	5	1	4	2	19.04	21.87	14.2	20.17	22.09	16.15	14.68
6	1	6	1	9	5	18.77	20.98	13.23	20.06	21.88	15.62	13.37
7	1	7	1	6	3	20.44	23.27	14.72	20.84	21.06	17.56	16.01
8	1	8	1	7	4	18.79	20.86	12.25	19.17	21.88	15.85	13.96
9	1	9	1	2	1	19.07	21.94	14.03	19.69	21.33	16.25	14.75
10	1	10	1	2	1	18.82	22.73	13.34	19.76	20.86	17.45	15.04

Fig. 4: COExpress Dataset View

Estimating a CCR Model

Selecting the Type of Model:

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

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 "ZPC1" as the dependent variable.

Selecting the Predictors:

- In the Control window below "Predictors", click and hold on "ABL1"
- Scroll down, and move the cursor down to "INDPT28" to highlight all 84 predictors. Click on the box next to "INDPT28" to select all 84 extraneous 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.

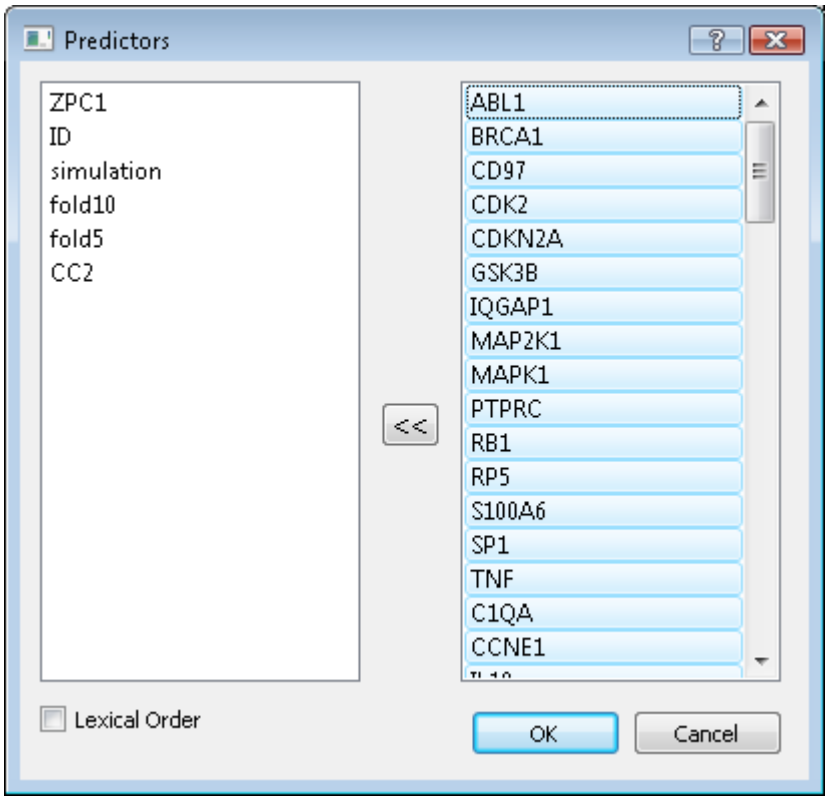


Fig. 5: Predictor Window

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 "# Variables:" box, delete "1" and type "10"

Selecting the Number of Components:

- Under Options, keep the default "# Components" (4 components)

Selecting the Model Type:

- Click on CCR.lda to select a CCR linear discriminant analysis model

Your Control window should now look like this:

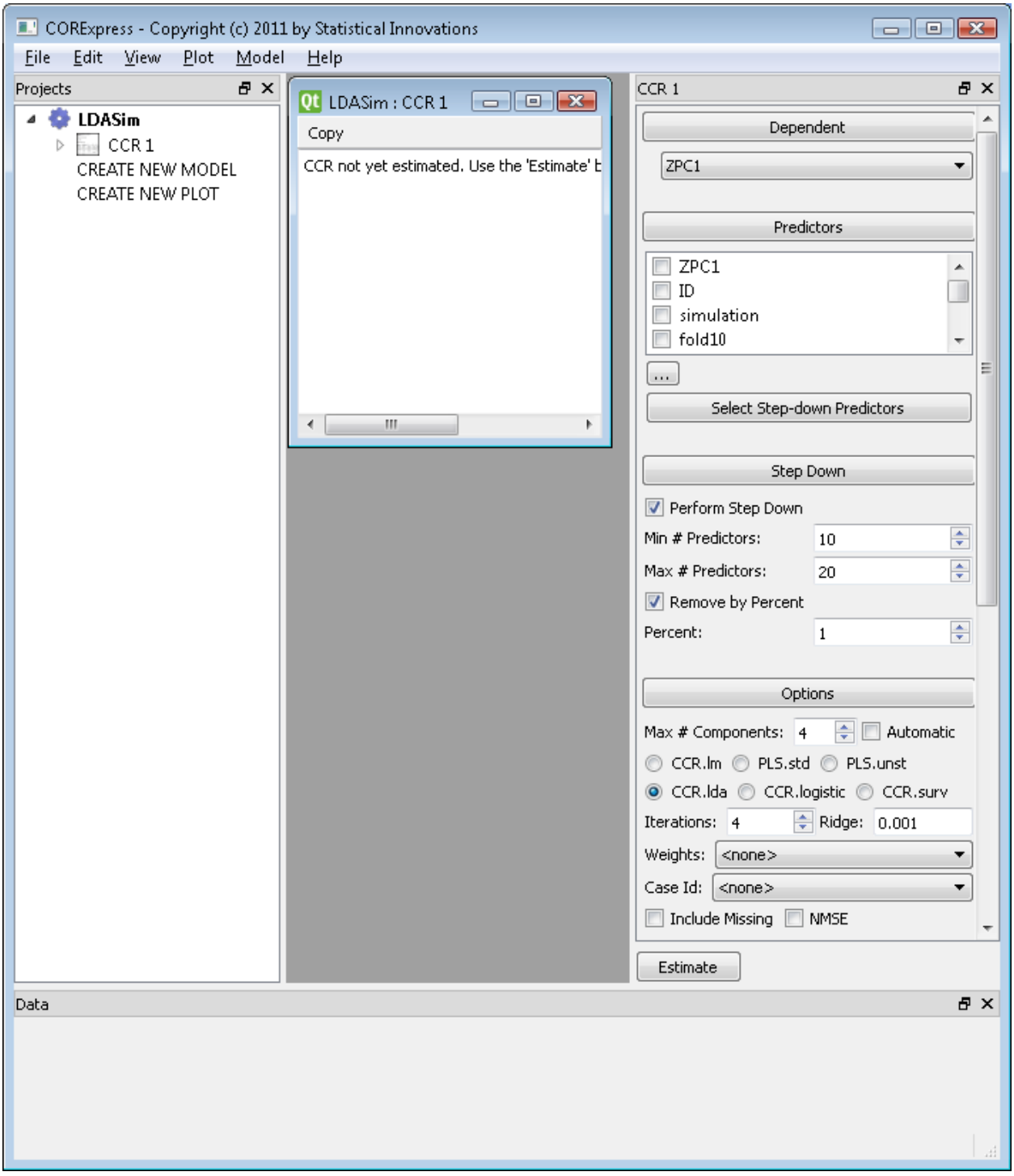


Fig. 6: Control Window

Specifying the Training Dataset:

- Click on "Validation" and options will appear for selecting training and validation datasets.
- Under the Training Subset, click on the "<select>" drop down menu and click on "simulation".
- Click on the corresponding "=" drop down menu and change the default "=" to "<"
- Delete the default "0" in the corresponding Training Subset numeric box and type 3.

Now, all records with simulation<3 will be selected as the Training dataset, providing an analysis size of N=100.

Specifying the Validation Dataset:
All simulation>3 will be automatically selected as the validation dataset.

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, delete the default "1" and type "10"
- In the "# Folds:" box, change the default "10" to "5"
- Keep the "<none>" in the Fold Variable drop down drop down menu

This divides the analysis sample into 5 subsamples (folds) that will 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 will be CV-ACC, the cross-validated ACC. This statistic is estimated based on model scores obtained from the analysis sample, excluding a particular fold, and applied to the fold excluded. Thus, the performance of the model is measured using cases not used at all in the development of the model.

Your Control window should now look like this:

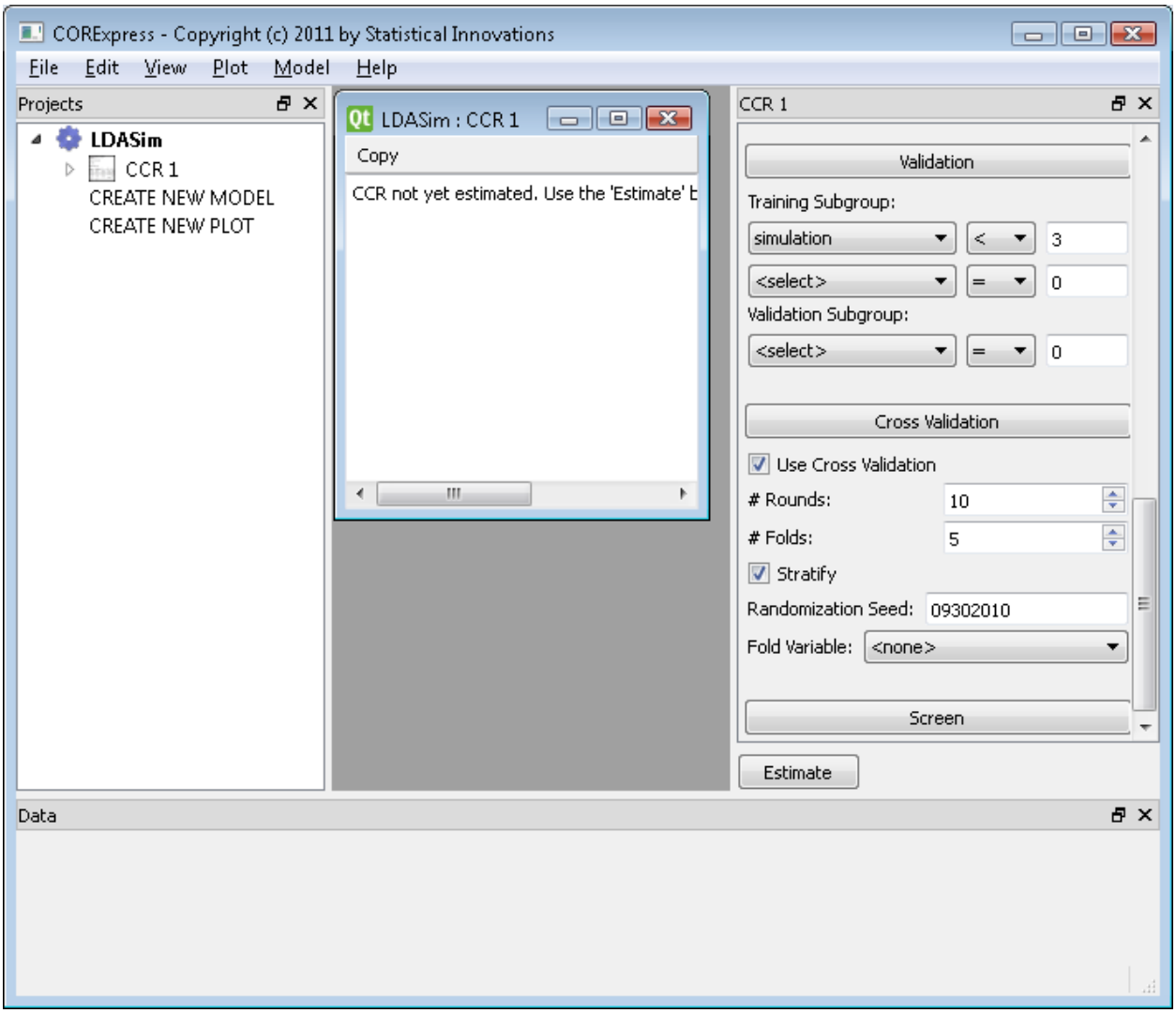


Fig. 7: Control Window

Estimate the Specified Model:

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

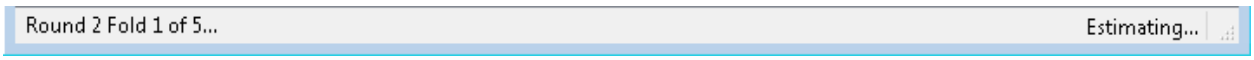


Fig. 8: COExpress Status Bar

When the model is finished estimating, a new window will pop up: "LDASim : CCR 1" (CV-AUC /ACC Plot)

View Model Output

Viewing CV-R^2 Plot:

- Click on the "LDASim : CCR 1" window (CV-AUC /ACC Plot)

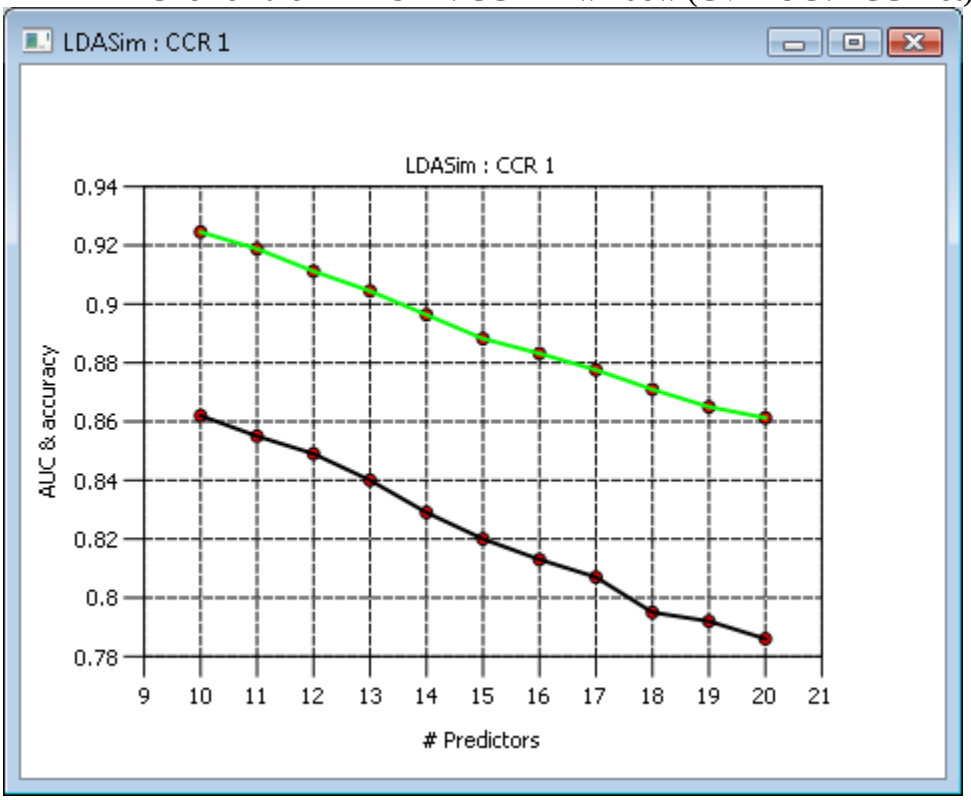


Fig. 9: CV-AUC /ACC Plot

The CV-AUC and ACC plotted in the graph corresponds to the cross-validation AUC and ACC based on the 4-component model for number of predictors P ranging from 20 down to 10.

Viewing CV-AUC / CV-ACC Output:

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

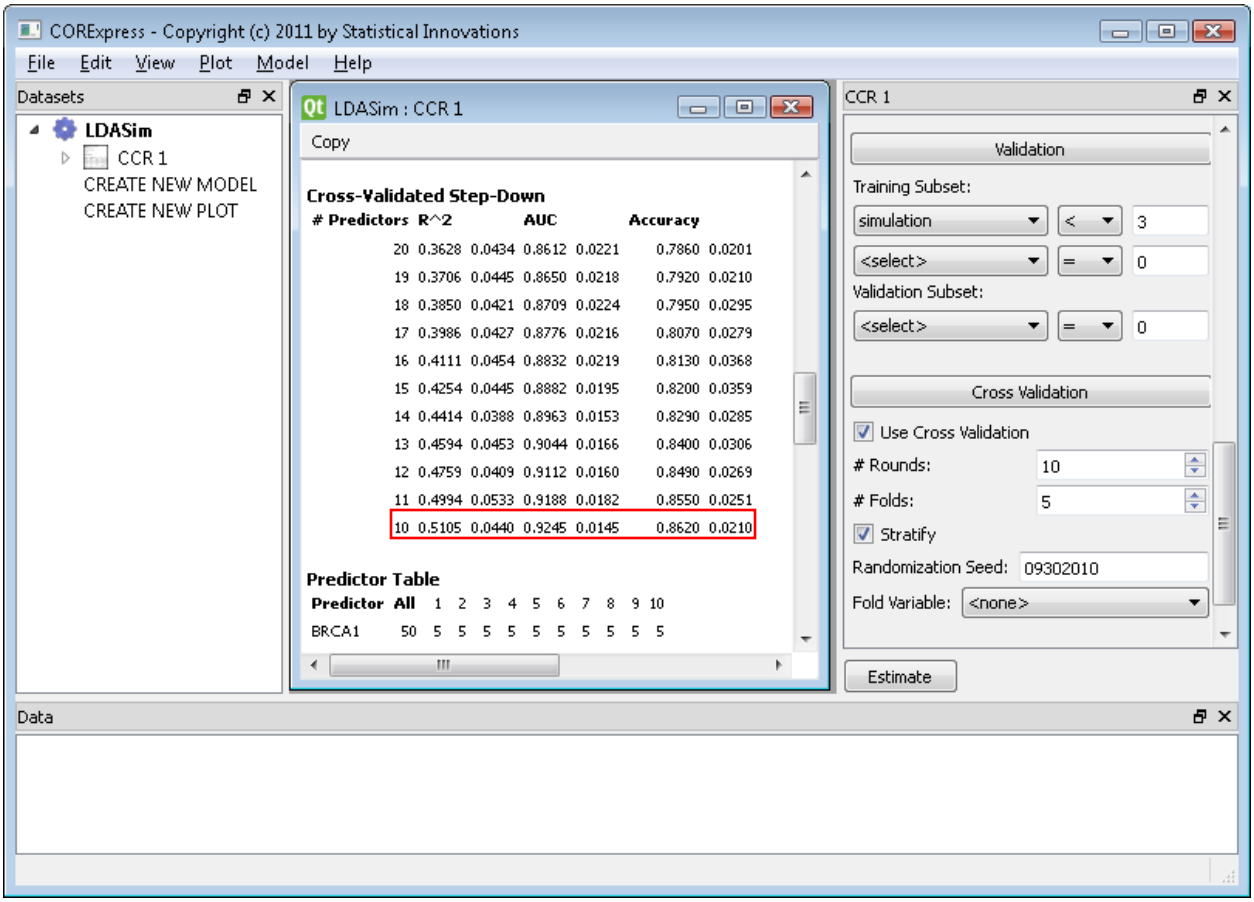


Fig. 10: Cross Validation ACC, AUC & Predictor Table Output in the Model Output Window

The cross-validation ACC and AUC (CV-ACC & CV-AUC) is located at the bottom of the CCR 1 Model Output Window along with the R^2 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-ACC. In the case of ties, the optimal number of predictors P* is taken to be the largest value for CV-AUC among those with the same highest value for CV- ACC.

Viewing the 'Optimal' Model Output:

- Click on the "LDASim : CCR 1" window in COExpress
- Scroll to the top of the window

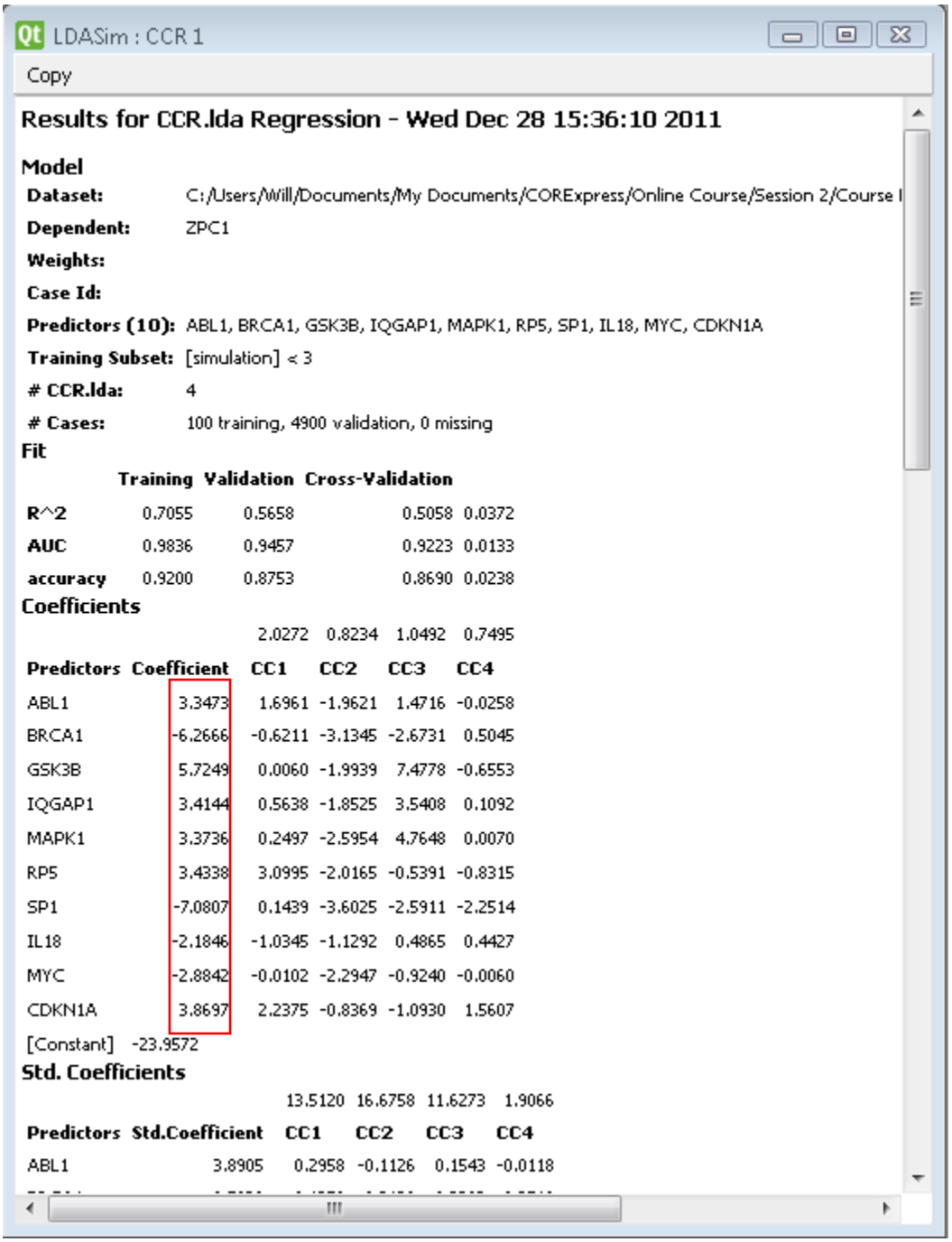
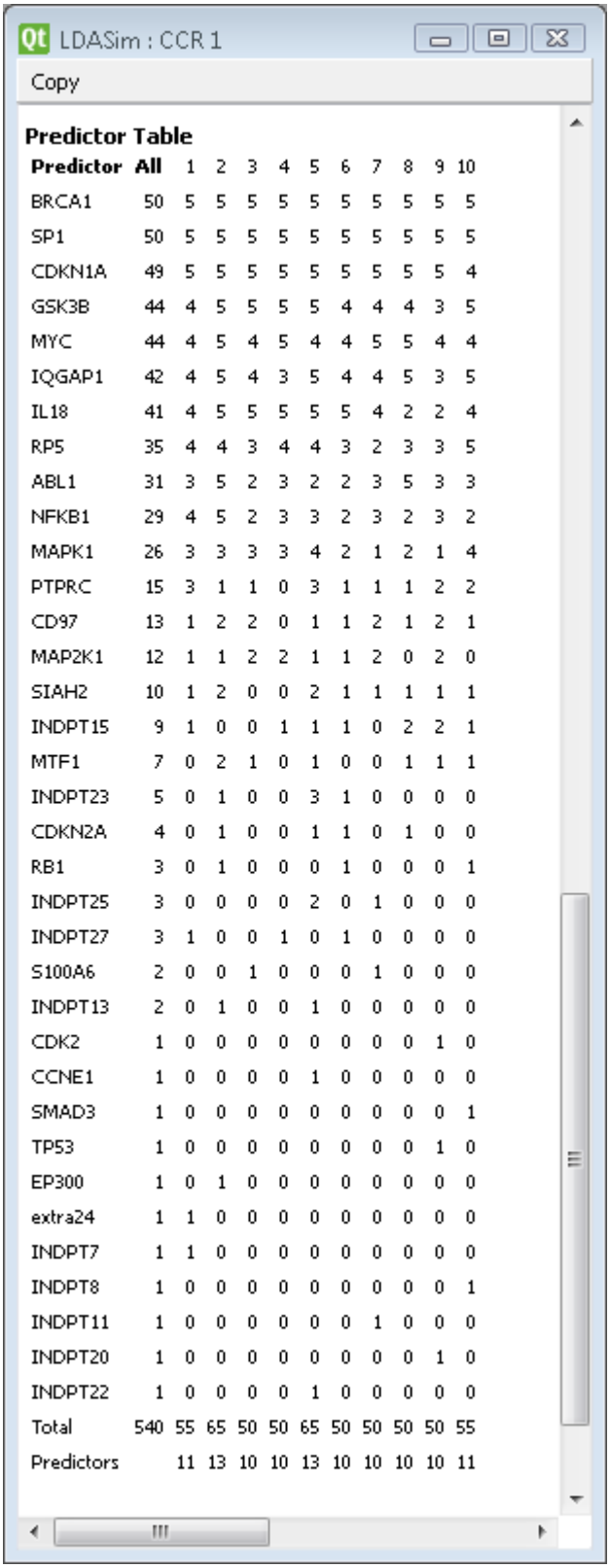


Fig. 11: Unstandardized Coefficients for K=4 in the Model Output Window
Note that for the Cross-validation, the ACC=0.8690 and for the Validation the ACC=0.8753.



Predictor Table

Predictor	All	1	2	3	4	5	6	7	8	9	10
BRCA1	50	5	5	5	5	5	5	5	5	5	5
SP1	50	5	5	5	5	5	5	5	5	5	5
CDKN1A	49	5	5	5	5	5	5	5	5	5	4
GSK3B	44	4	5	5	5	5	4	4	4	3	5
MYC	44	4	5	4	5	4	4	5	5	4	4
IQGAP1	42	4	5	4	3	5	4	4	5	3	5
IL18	41	4	5	5	5	5	5	4	2	2	4
RP5	35	4	4	3	4	4	3	2	3	3	5
ABL1	31	3	5	2	3	2	2	3	5	3	3
NFKB1	29	4	5	2	3	3	2	3	2	3	2
MAPK1	26	3	3	3	3	4	2	1	2	1	4
PTPRC	15	3	1	1	0	3	1	1	1	2	2
CD97	13	1	2	2	0	1	1	2	1	2	1
MAP2K1	12	1	1	2	2	1	1	2	0	2	0
SIAH2	10	1	2	0	0	2	1	1	1	1	1
INDPT15	9	1	0	0	1	1	1	0	2	2	1
MTF1	7	0	2	1	0	1	0	0	1	1	1
INDPT23	5	0	1	0	0	3	1	0	0	0	0
CDKN2A	4	0	1	0	0	1	1	0	1	0	0
RB1	3	0	1	0	0	0	1	0	0	0	1
INDPT25	3	0	0	0	0	2	0	1	0	0	0
INDPT27	3	1	0	0	1	0	1	0	0	0	0
S100A6	2	0	0	1	0	0	0	1	0	0	0
INDPT13	2	0	1	0	0	1	0	0	0	0	0
CDK2	1	0	0	0	0	0	0	0	0	1	0
CCNE1	1	0	0	0	0	1	0	0	0	0	0
SMAD3	1	0	0	0	0	0	0	0	0	0	1
TP53	1	0	0	0	0	0	0	0	0	1	0
EP300	1	0	1	0	0	0	0	0	0	0	0
extra24	1	1	0	0	0	0	0	0	0	0	0
INDPT7	1	1	0	0	0	0	0	0	0	0	0
INDPT8	1	0	0	0	0	0	0	0	0	0	1
INDPT11	1	0	0	0	0	0	0	1	0	0	0
INDPT20	1	0	0	0	0	0	0	0	0	1	0
INDPT22	1	0	0	0	0	1	0	0	0	0	0
Total	540	55	65	50	50	65	50	50	50	50	55
Predictors		11	13	10	10	13	10	10	10	10	11

Fig. 12: Predictor Table

One measure of importance is the absolute value of the standardized coefficients. Note that the top two with coefficients greater than four are BRCA1 and SP1 and the weakest is RP5. An alternative measure of predictor importance is given in the predictor table where importance is ranked by the number of times they were included in one of the models during the 10 rounds of 5-fold cross-validation. Note that the top 10 in the predictor table are the same as the 10 predictors in the model. Also note that BRCA1 and SP1 are again at the top of the predictor table list, but not CDKN1A is also at the top. This predictor table could be used as another criteria to select a final model.

Viewing the Training Plot:

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

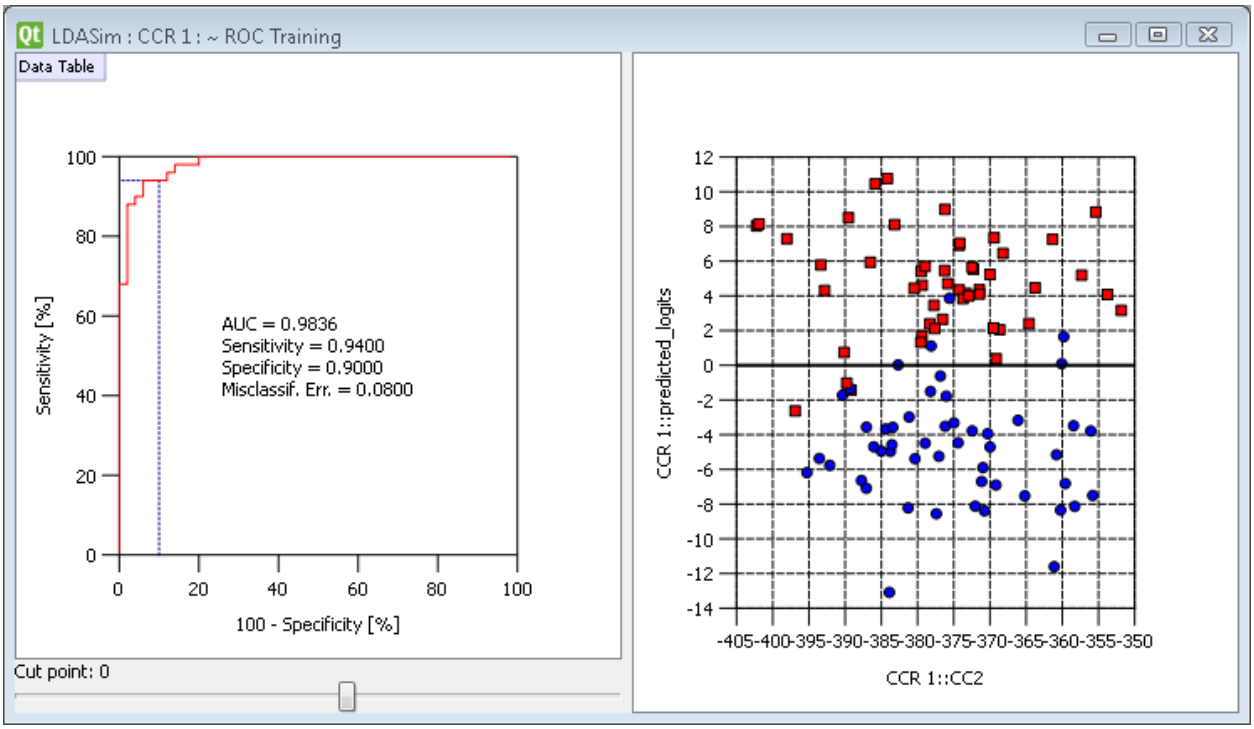


Fig. 13. Interactive ROC and Scatter Plot for the Training Data

Note that the Training-ACC of 92% reported in the Model Summary output represents correct classification of 94% of the ZPC1=1 (red dots) and 90% of the ZPC1=0 (blue dots). As you can see, 3 of the red points are below the default cutpoint, and 5 of the blue points are above the default cutpoint. If you click on one of the points in the scatterplot, it is highlighted in the data table.