Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

COREXpress® 1.0

USER'S GUIDE

Jay Magidson

Statistical Innovations

Thinking outside the brackets!™

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

This document should be cited as: Magidson, J. (2014). COREXpress User's Guide: Manual for COREXpress, Belmont, MA: Statistical Innovations Inc.

For more information about Statistical Innovations Inc. please visit our website at http://www.statisticalinnovations.com or contact us at:

Statistical Innovations Inc.
375 Concord Avenue, Suite 007
Belmont, MA 02478

e-mail: support@statisticalinnovations.com

COREXpress (patent pending) is a trademark of Statistical Innovations Inc.
Windows is a trademark of Microsoft Corporation.
SPSS is a trademark of SPSS, Inc.
Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

COREXpress User's Guide: Manual for COREXpress
Copyright © 2014 by Statistical Innovations Inc.
All rights reserved.

No part of this publication may be reproduced or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission from Statistical Innovations Inc..

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Table of Contents

1 Introduction to COREXpress and CCR 6
2 Menus 6
2.1 File 6
2.2 Edit 7
2.3 View 7
2.4 Plot 7
2.5 Model 8
2.6 Help 8
3 Windows 9
3.1 Project Window 9
3.2 Model Control Window 11
3.3 Model Output Window 12
3.4 Data Window 13
3.5 Status Bar Window 15
3.6 Plot Control Window 16
4 Data Files and Formats 17
4.1 Opening a Data File 18
4.2 Data File Formats 20
4.3 Saving Projects (Model Settings & Output) 22
5 Model Setup 23
5.1 Select the dependent variable 23
5.2 Select predictors 23
5.3 Step Down 24
5.3.1 Specifying range for # predictors 25
5.3.2 Predictor Selection Options 25
5.4 Options 27
5.4.1 Specify # Components (Max # Components) 27
5.4.2 Select Model Type 28
5.4.3 Technical Options for Logistic Regression and Survival Model Types 29

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

5.4.4 Weights 29
5.4.5 Case Id 29
5.4.6 Include Missing 30
5.4.7 NMSE 30
5.5 Specifying Training & Validation Subgroups 30
5.6 Cross-Validation (CV) Options 32
5.6.1 # Rounds: 32
5.6.2 # Folds: 32
5.6.3 Stratify: 32
5.6.4 Randomization Seed: 33
5.6.5 Fold Variable: 33
5.7 Screen 33
5.8 Estimate the Model 33
5.8.1 Canceling Model Estimation 34
5.9 Model and Output Management 34
6 Model Output and Plots 36
6.1 Output Created 36
6.1.1 Cross-Validation Predictor Graphs 36
6.1.2 Interactive Scatter Plot and ROC Curve 38
6.1.3 Model Summary Output 40
6.1.4 Cross-Validated Component Table 43
6.1.5 Cross-Validated Step-Down Table 45
6.1.6 Cross-Validation Predictor Table 47
6.2 Plots 48
6.2.1 Scatter Plot 50
6.2.2 Box Plot 52
6.2.3 Multivariate Box Plot 53
6.2.4 Histogram 54
6.2.5 ROC Plot 55
7 Generating Predictions 59
8 Summary of Correlated Component Regression (CCR) 60

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

9 Tutorials 64
9.1 Tutorial 1: Getting Started with Correlated Component Regression (CCR) in COREXpress® 64
9.2 Tutorial 2: CCR for a Continuous Dependent Variable 82
9.3 Tutorial 3: CCR for a Dichotomous Dependent Variable 111
10 Index 139

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

1 Introduction to COORExpress and CCR

Correlated Component Regression (CCR) is a regression method for predicting a dependent variable Y from a set of P predictor variables, where it is possible that P > N, N being the sample size. Different variants are available depending upon the scale type of the dependent variable, and assumptions about the predictor variables. In addition, a variant incorporating PLS regression is also available. The various model types corresponding to these variants are listed in Section 5.4.2.

A summary of CCR is provided in Section 8. For further details see:

Magidson, J. (2010) 'Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features', Proceedings of the American Statistical Association.

Magidson, J. and Wassmann, K. (2010) 'The Role of Proxy Genes in Predictive Models: An Application to Early Detection of Prostate Cancer', Proceedings of the American Statistical Association.

To get started, Section 9 presents step-by-step tutorials.

2 Menus

The Menu Bar in COORExpress has 6 general menu options: File, Edit, View, Plot, Model, and Help

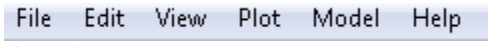


Figure 2-1: Menu Bar

2.1 File

New Project. You can start a new project and enter your data by hand or paste the data from a non-SPSS or ASCII text file source (such as Excel).

Load Dataset… Opens a data file to start a new project. COORExpress accepts input data as either an SPSS system (.sav) file or an ASCII text file (.csv, .txt, etc.). For additional information on opening data files, see Section 4.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Load Project… Opens a previously saved COREXpress project (.spp).

Save Project As… This saves the analysis settings (and associated output) that have been specified for one or more models in the active project. See Section 224.3.

Close Project. Closes the active project.

Close All Projects. Closes all open projects.

Exit. Exit the program.

2.2 Edit

Copy Plot… Allows you to copy the active plot to the clipboard as a bitmat. You will be prompted to specify the target width (pixels).

Copy Model…Allows you to copy the model output from the active Model Output window to the clipboard.

2.3 View

Workspace. Hides the Project window when unchecked.

Data. Hides the Data window when unchecked.

Control. Hides the Model Control window when unchecked.

Grid. Arranges all windows in a tile format when checked. Check again to unarranged the windows.

2.4 Plot

Add Scatter Plot. Click to add a scatter plot. For more information, see Section 6.

Add Box Plot. Click to add a box plot. For more information, see Section 6.

Add Multivariate Box Plot. Click to add a multivariate box plot. For more information, see Section 6.

Add Histogram. Click to add a histogram. For more information, see Section 6.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Add ROC Plot. Click to add a ROC plot. For more information, see Section 6.

2.5 Model

Add CCR Model. Click to add a new CCR model immediately below the last model in the active project in the Projects window.

Cancel Estimation. Click to cancel current model estimation.

2.6 Help

Register… Enter your registration code here to license the program (be sure to include all spaces).

About CORExpress... Provides general information about the program.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

3 Windows

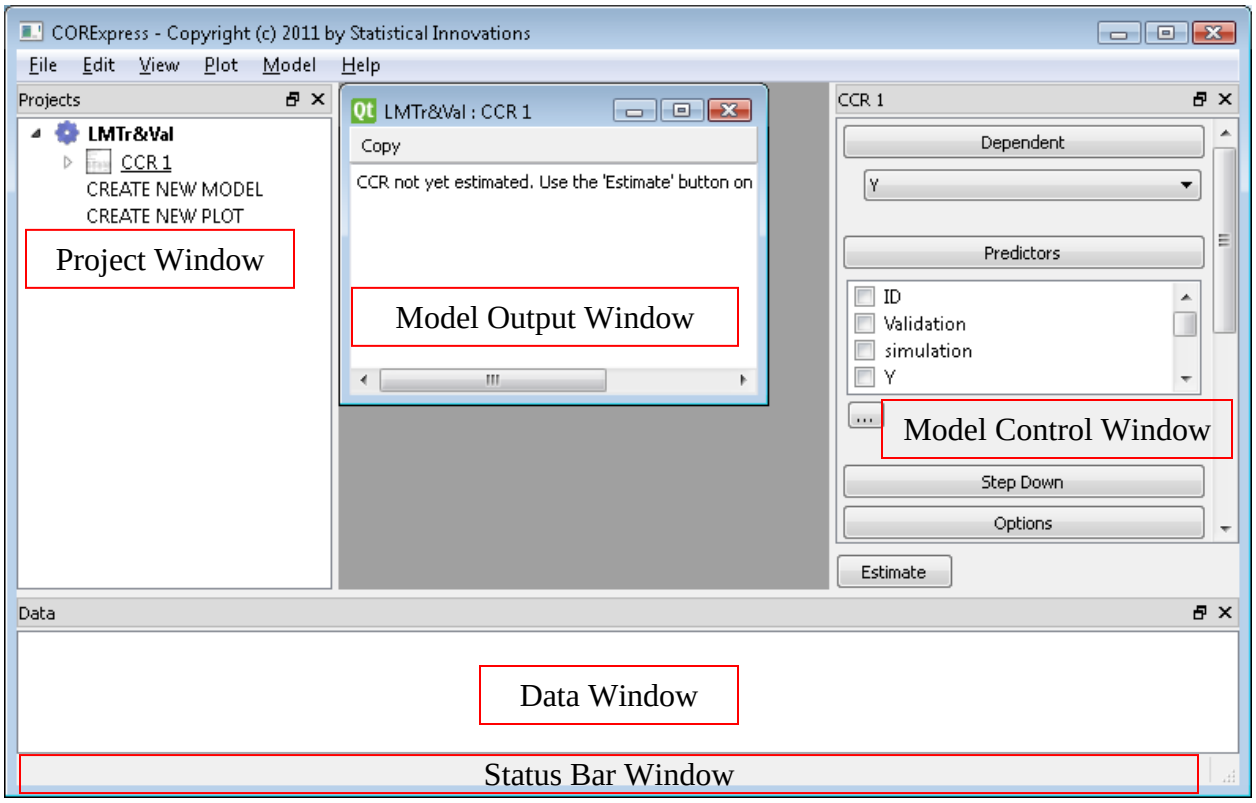


Figure 3-1: COREXpress Windows

3.1 Project Window

The Project window is located on the left-hand side of the application (See Figure 3-1). It provides a hierarchical view of the contents, consisting of the names of projects or datasets opened, the project models, and the plots generated for each model. To expand/contract any level of the hierarchy, click the ► / ▼icons.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

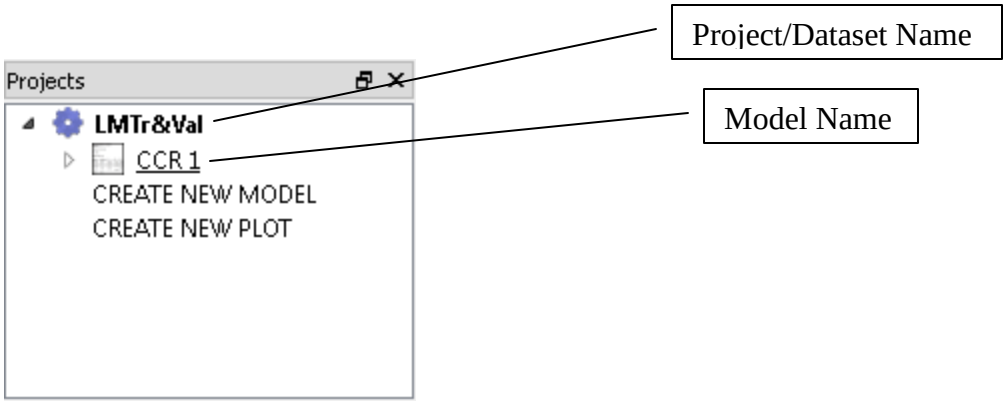


Figure 3-2: Project Window

Projects/Datasets

Right clicking on currently open projects/datasets provides the following options:

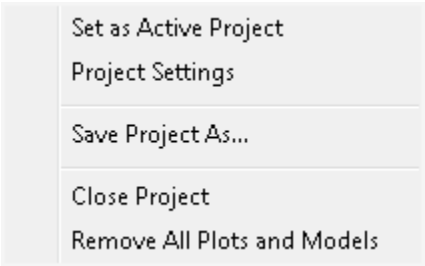


Figure 3-3: Project Options

'Set as Active Project'
Sets project as active project. The projects/datasets will now be **bold**. You can also double click on the project/dataset to set it as the active project. Set the project as active before saving and closing projects.

'Project Settings'
Opens the Data window Variable Selection dialog box. For each variable on the dataset it provides the 1) Predictor type (categorical, continuous, etc.), 2) Mean, 3) Range, 4) Count. Check the box next to the variable to view the variable for all cases in the Data window.

'Save Project As…'
Opens the Save Project Dialog window to save the project.

'Close Project'
Closes the project.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

'Remove All Plots and Models'
Deletes all previously estimated models and associated plots in the project.

CCR 1/Model Name

By default, the 1st model is assigned the name 'CCR 1', the 2nd, 'CCR 2', … . Right clicking on a model name directly below the project/dataset name provides the following options:

Figure 3-4: Model Options

'Rename'
Allows you to rename the model.

'Copy'
Click to copy the settings associated with the selected model and create a new model at the end of the list of models below the project/dataset.

'Remove'
Deletes the model and all associated model settings & plots.

3.2 Model Control Window

The Model Control window is used to select variables and set analysis options. Model specifications that are required for analysis are shown ('Dependent', 'Predictors', and 'Options') and cannot be hidden/collapsed. See Section 5 for model setup. Additional options for model development can be expanded (unhidden) by clicking on them ('Step Down', 'Validation', 'Cross Validation', and 'Screen'). See Section 5 for more details.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

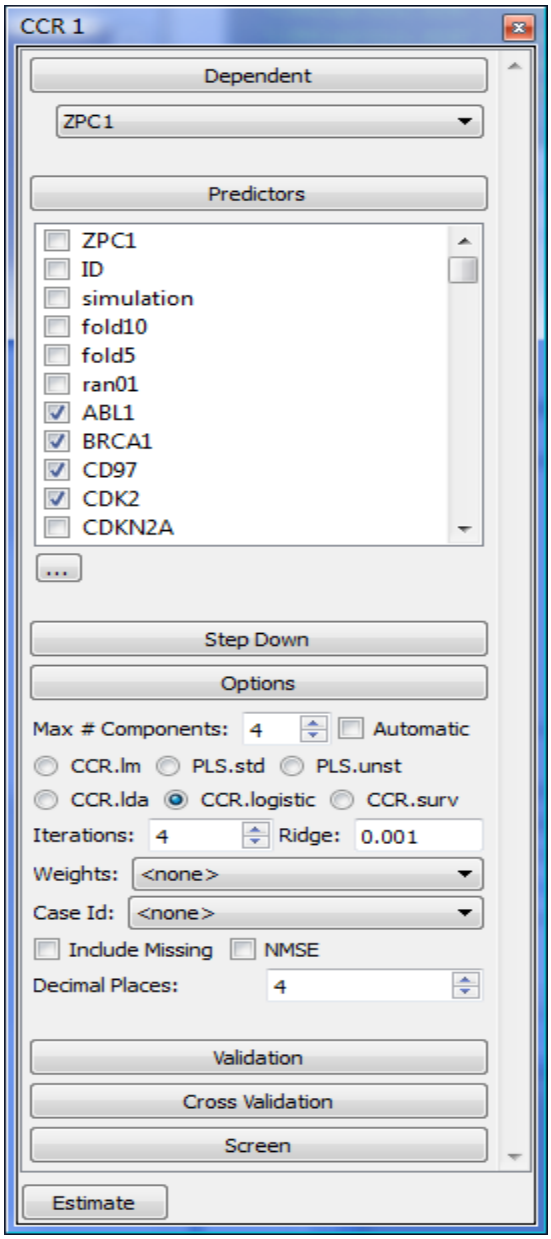


Figure 3-5: Model Control Window

3.3 Model Output Window

When a model is estimated, model summary output will appear in the Model Output window. To copy individual parts of the output, select the text using your cursor, right click, and click 'Copy' (Ctrl + C). For more details, please see Section 6.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

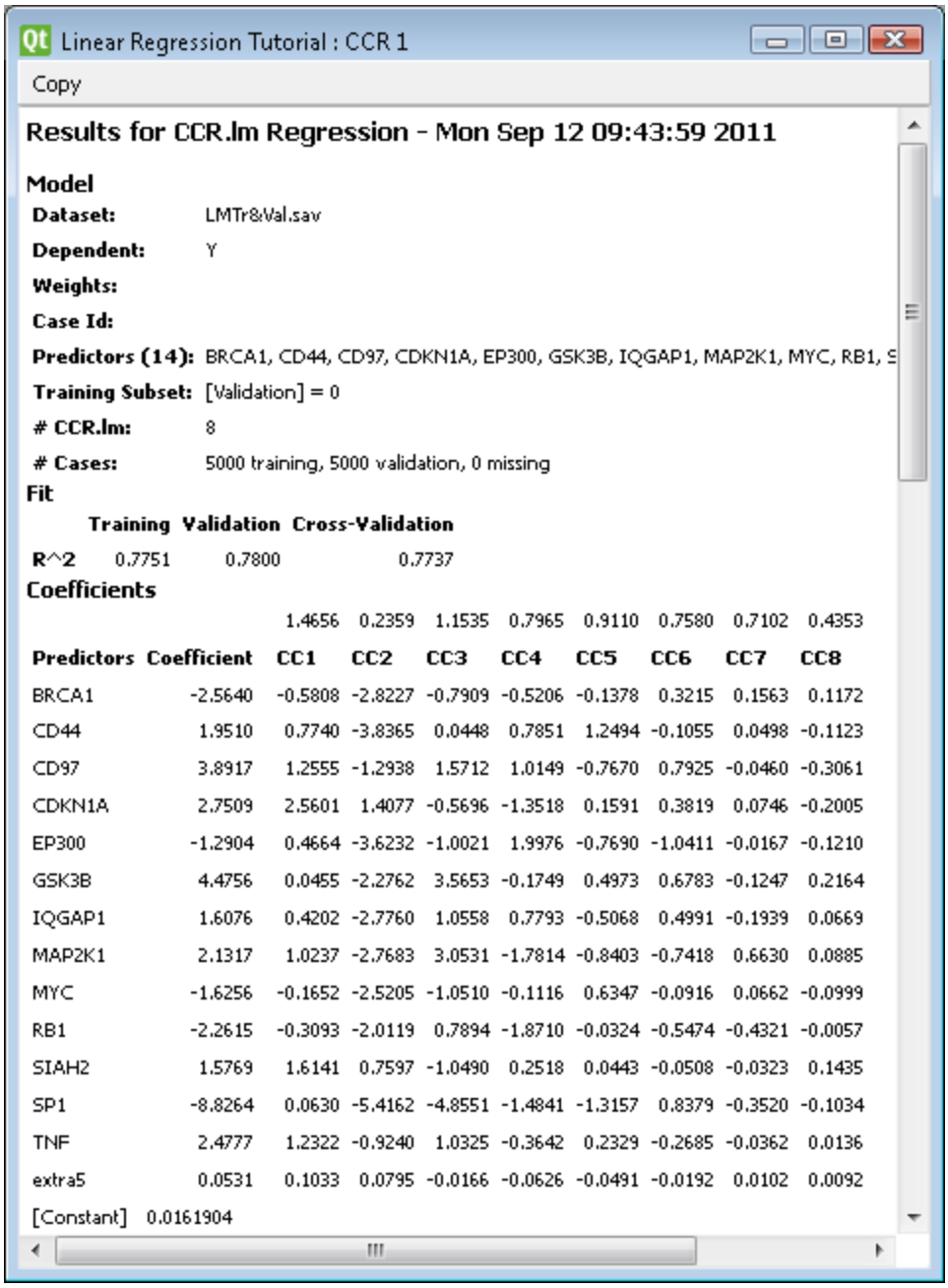


Figure 3-6: Model Summary Output

3.4 Data Window

DATA WINDOW & PROJECT SETTINGS MENU

The Data window at the bottom of COREXpress can display the values for a subset of variables included in the dataset. Variables to be displayed in the Data window can be selected from the Project Settings Menu, which includes the following information for each variable:

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- Type (continuous, categorical, etc.)
- Mean
- Range
- Count (frequency)

To open the Project Settings Menu:

- Right click on the project name in the Project window.
- Select 'Project Settings'

The following window containing descriptive summary statistics appears:

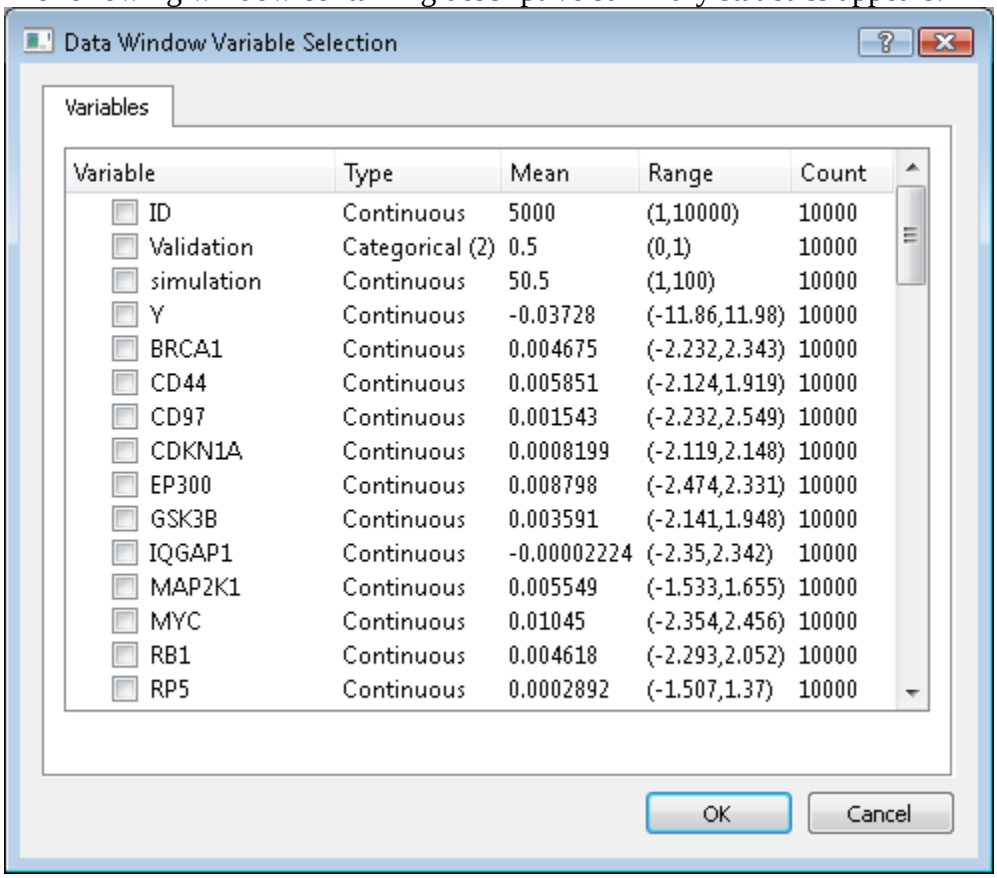


Figure 3-7: Project Settings Menu Option: Data Window Variable Selection

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

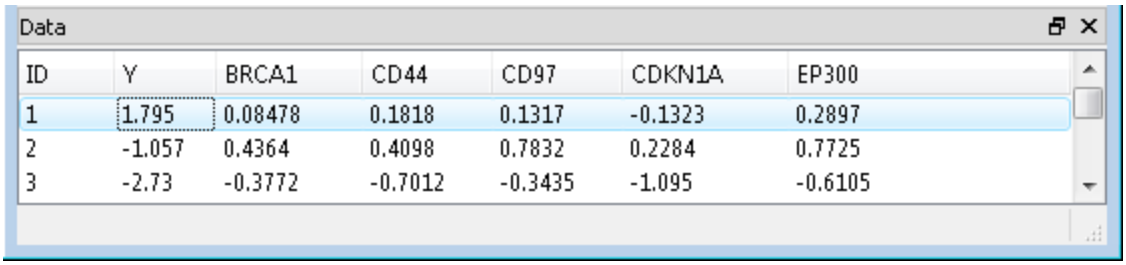


Figure 3-8: Data Window

Data Window: Sorting

In the Data window, you can sort the data by clicking on a variable name at the top of the Data window. Click once to sort ascending; click twice to sort descending. Note: sorting in the Data window will not change the order of the input data set.

Data Window: Sorting

You can change the order of appearance of the variables in the Data window by clicking on a variable name at the top of the Data window and dragging it to the desired location. Note: moving variables in the Data window will not change the order of the variables in the input data set.

3.5 Status Bar Window

The Status Bar will provide information for the following:

- When loading a dataset, the status bar will report how many cases were read.
- When estimating a model, the text 'Estimating…' will appear in the right-hand side of the status bar.
    - When applicable, status messages about Rounds, Folds, and Step-down are shown in the left-hand side during estimation.

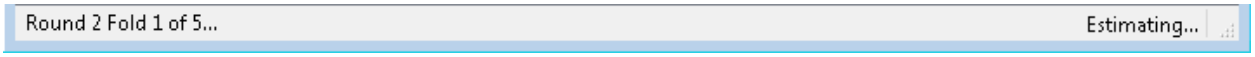


Figure 3-9: Status Bar Window

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

3.6 Plot Control Window

A variety of traditional plot options are available. To see these, double click 'CREATE NEW PLOT' from the Project window. Alternatively, click on 'Plot' from the Menu Bar at the top and select the desired plot. Once you select your desired plot, the Plot Control window will appear on the right-hand side of the program in the place of the Model Control window. To view the Model Control window, double click on the model name in the Projects window. To view the Plot Control window, double click on the plot name in the Projects window. For details, see Section 6.

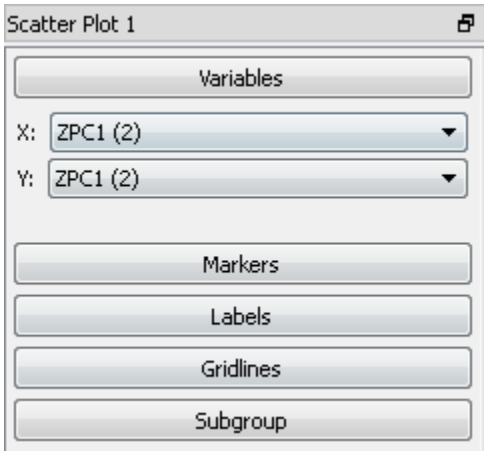


Figure 3-10: Scatter plot Control Window

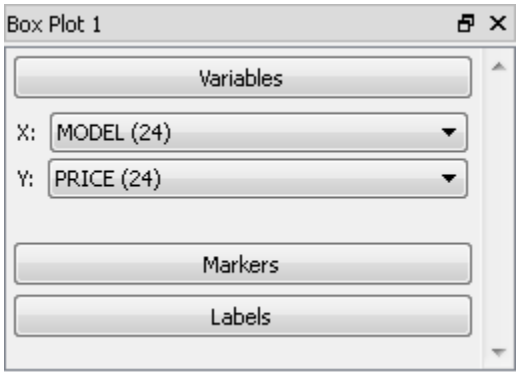


Figure 3-11: Box Plot Control Window

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

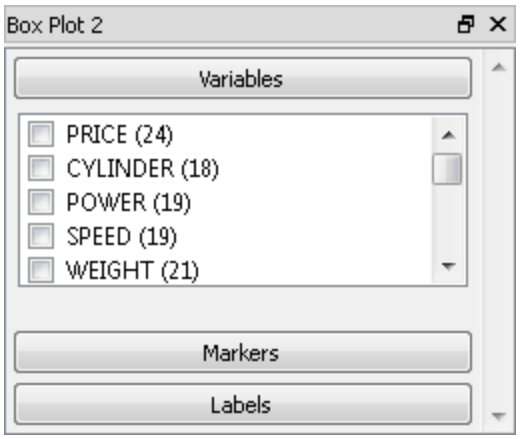


Figure 3-12: Multivariate Box Plot Control Window

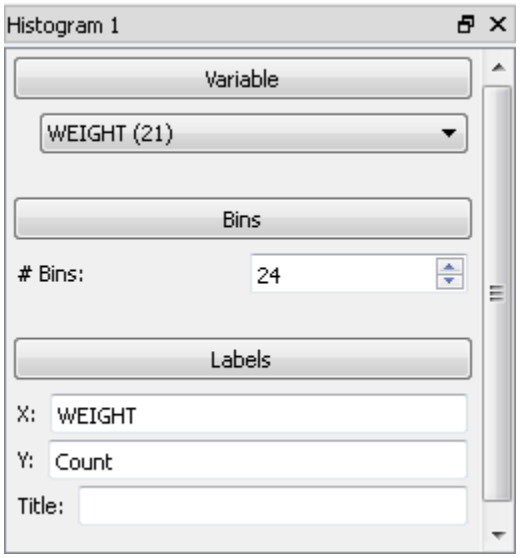


Figure 3-13: Histogram Control Window

4 *Data Files and Formats*

This Section shows how to open or close a data file and how to save projects. It also describes the alternative data file formats that can be used.

The program accepts the following input data:
1) saved as an SPSS (.sav) system file,
2) saved as an ASCII text file

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

4.1 Opening a Data File

To open data files in SPSS system file format or text file format.

To open the file, from the menus choose:

- File → Load Dataset…
- Select the name of the file you want to open and click 'Open' to load the dataset

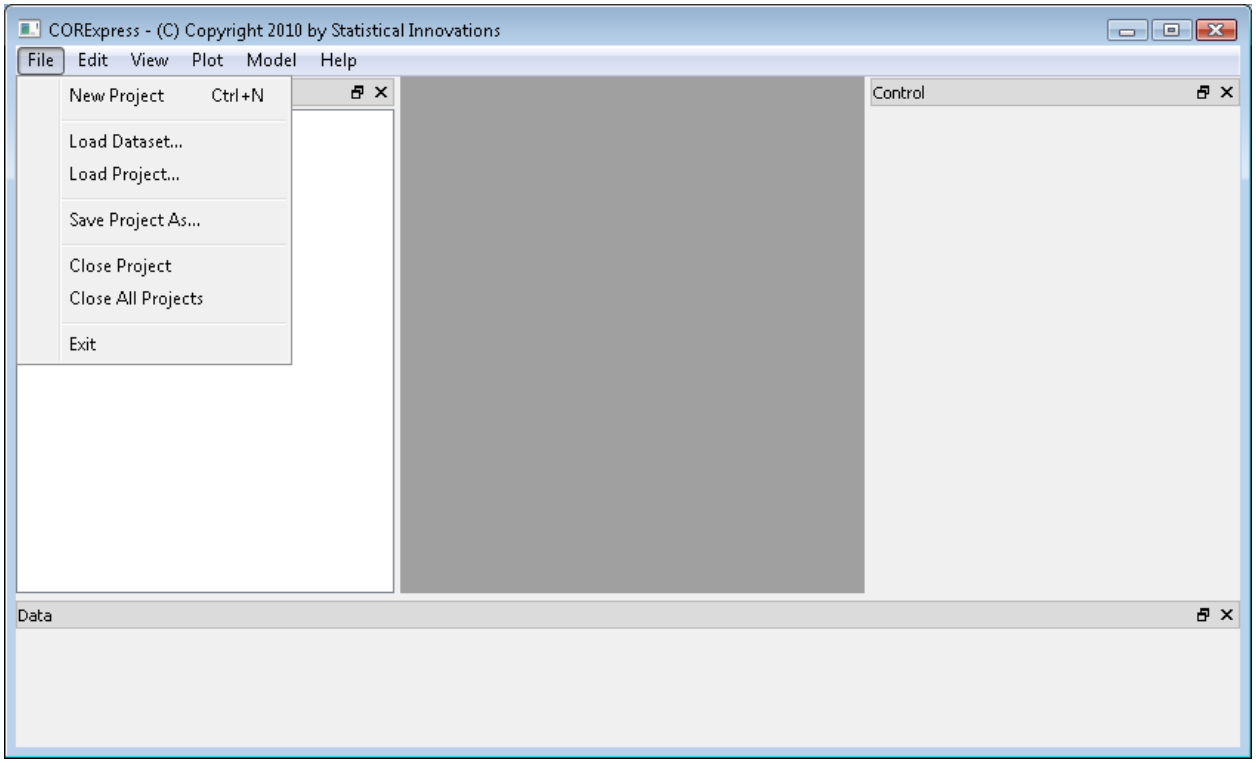


Figure 4-1: File Menu

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

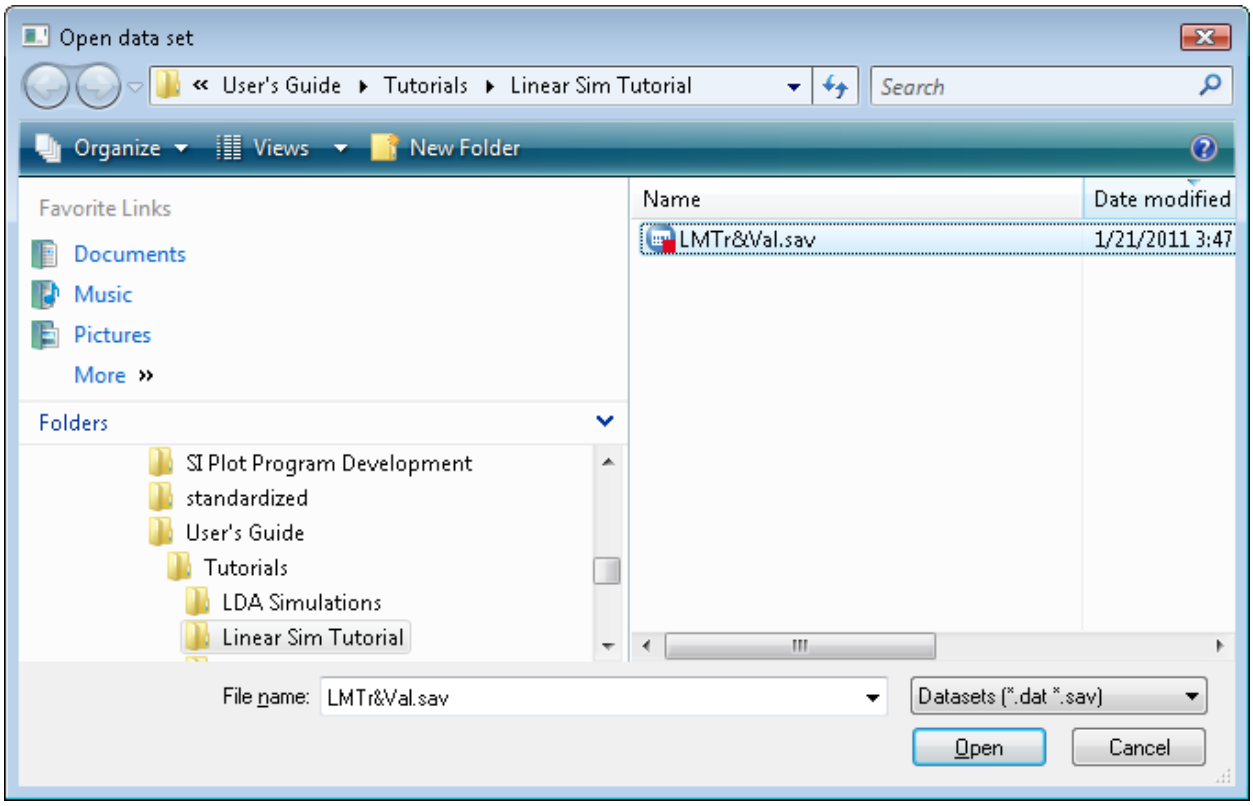
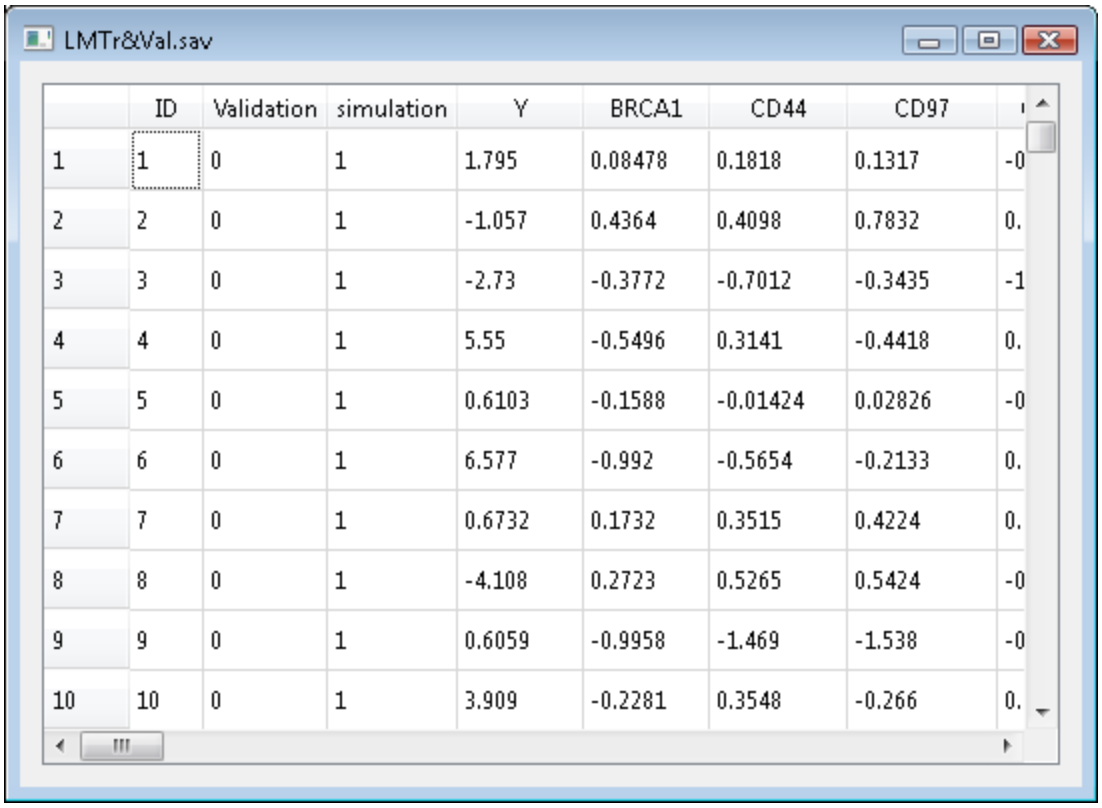


Figure 4-2: Loading a Dataset

You will now see the selected dataset loaded in the Projects window on the left. You can view the complete dataset in a new window by double clicking on the dataset name in the Project window.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)



	ID	Validation	simulation	Y	BRCA1	CD44	CD97	
1	1	0	1	1.795	0.08478	0.1818	0.1317	-0
2	2	0	1	-1.057	0.4364	0.4098	0.7832	0.
3	3	0	1	-2.73	-0.3772	-0.7012	-0.3435	-1
4	4	0	1	5.55	-0.5496	0.3141	-0.4418	0.
5	5	0	1	0.6103	-0.1588	-0.01424	0.02826	-0
6	6	0	1	6.577	-0.992	-0.5654	-0.2133	0.
7	7	0	1	0.6732	0.1732	0.3515	0.4224	0.
8	8	0	1	-4.108	0.2723	0.5265	0.5424	-0
9	9	0	1	0.6059	-0.9958	-1.469	-1.538	-0
10	10	0	1	3.909	-0.2281	0.3548	-0.266	0.

Figure 4-3: Dataset View

4.2 Data File Formats

COREXpress will load data from input files in any of the following data formats:

- SPSS system file (*.sav)
- ASCII text file (*.csv, and others)

SPSS Files

For information on creating an SPSS system file, see your SPSS manuals.

COREXpress recognizes the full 64 characters of SPSS variable names, and the 256 character limit for string values. String values are case sensitive: 'f' and 'F' are separate characters. SPSS value labels are not used. Date values are not translated from their internal representation.

Text Files

COREXpress recognizes several variants of delimiter-separated text files.

The delimiter can be the tab character, the comma, or sequence of 1 or more white space

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

characters. Comma delimited files are recognized by the use of ‘.csv’ as the file extension; the other delimiters are recognized by inspection of the data lines.

The first line of the file should have variable names.

The data for any variable may be numeric (quantitative values only) or may be a string variable, containing some or all alphabetic characters (such as 'Female', 'Male'). For numeric data, do not use commas (such as 3,634). String variables containing lower and upper case letters, such as 'f' and 'F', are distinguished as separate categories. Missing data for a numeric variable may be specified using a period ('.'). Each data record should contain exactly the same number of data elements, one for each variable name.

Below is a partial listing of a text formatted data file. This file contains one record per case. Note that records 6 and 7 are identical (0 30 0 F).

```
IMPROVE AGE TREAT GENDER
0 23 1 F
0 23 0 F
1 27 1 M
0 29 1 M
0 30 1 M
0 30 0 F
0 30 0 F
1 31 0 F
2 32 1 M
0 32 1 F
0 32 0 F
2 33 0 F
1 37 1 F
0 37 0 M
```

For data sets consisting of many observations that contain identical values on all variables, the inclusion of an optional integer value frequency count variable will reduce the number of physical records in the file. Below is a partial listing of another text formatted data file. For each data record, FREQ contains the count of observations having the specified values on each of the variables. For example, the value of FREQ for the first record, defined by the values specified as ‘1, 1, 1, 1, 1’ is 3,634.

```
BACK NECK JOINT SWELL STIFF FREQ
1 1 1 1 1 3634
1 1 1 1 2 73
1 1 1 2 1 87
1 1 1 2 2 10
1 1 2 1 1 440
```

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

1 1 2 1 2 89
1 1 2 2 1 106
1 1 2 2 2 75
1 2 1 1 1 295
2 2 2 1 2 162
2 2 2 2 1 44
2 2 2 2 2 176

4.3 Saving Projects (Model Settings & Output)

This 'File→Save Project As' menu option can be used to save the variable selections and other option settings used for one or more previously defined models in the format of a COREXpress save file (.spp). This .spp file can then be opened at a later time using the 'File→Load Project' command. The advantage of an .spp file is the ability to retrieve analysis settings and output for a particular model or series of models at a later time without having to re-specify these settings. You can also use a retrieved setting as a starting point for specifying a similar model on the same data.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

5 Model Setup

5.1 Select the dependent variable

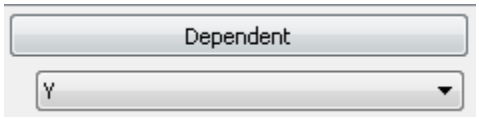


Figure 5-1: Dependent Variable Selector

By default, the dependent variable selected is the first variable on the data file. To change the selected dependent variable:

- In the Model Control window below 'Dependent', click on the drop down menu and select your desired dependent variable.

See Section 5.4.2 for details regarding the types of dependent variables accepted by the current version of COREXpress.

5.2 Select predictors

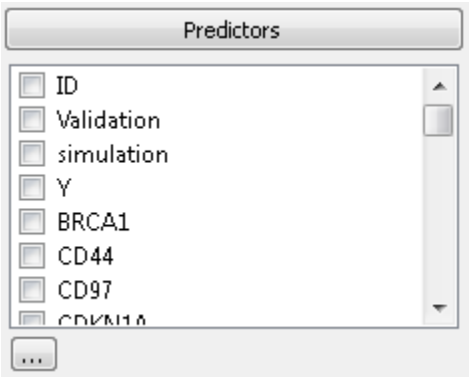


Figure 5-2: Predictor Selection

The names for all variables on the data file are listed in the Model Control window below 'Predictors'. Check the box next to the corresponding variable to select it as a predictor. To select multiple predictors at a time, click and hold on variable and move the cursor down to highlight multiple predictors. Click on the box next to any of the highlighted predictors to select all of them.

Alternatively, you can open a Predictors window to select the predictors:

- In the Model Control window below the 'Predictors' section, click the '…' button.
- The Predictors window will open.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- Click on any variable in the left box to highlight it, and click on the '>>' button to move the variable to the candidate predictors box on the right.
- To highlight multiple variables, click and hold on any variable and move the cursor down to highlight multiple predictors in the left box. Click on the '>>' button in the middle to move them to the right box as candidate predictors.

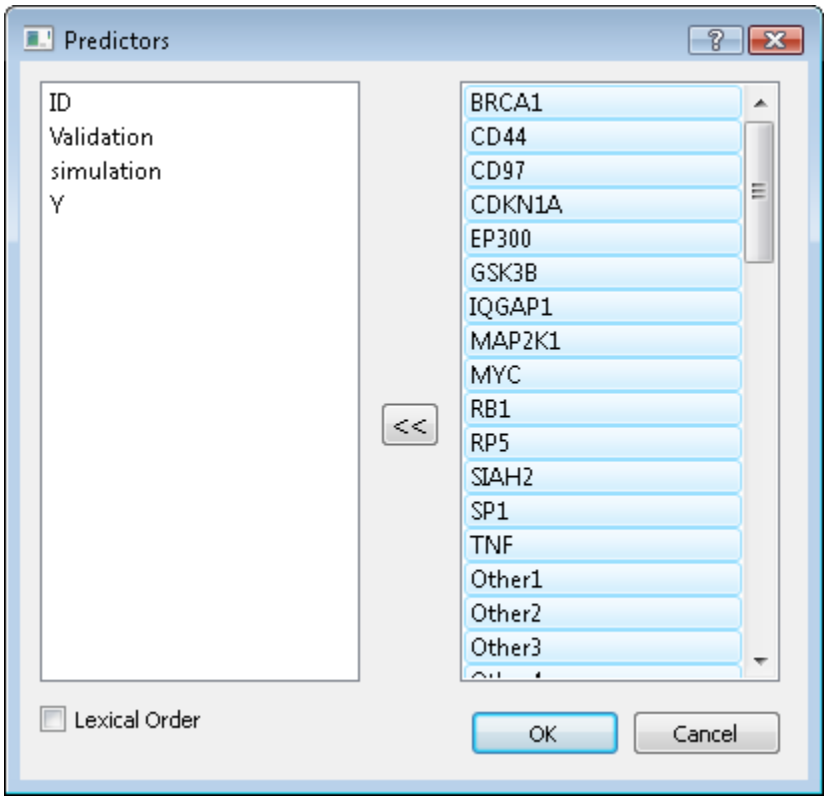


Figure 5-3: Predictor Window

5.3 Step Down

Various program features affect the selection of predictors to be included in the model. When the Step-down option is activated, the number of predictors P* is determined for a K*-component model, where either K* or the maximum value for K* is specified in the Options section. (See Max # Components and Automatic options in Section 5.4).

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

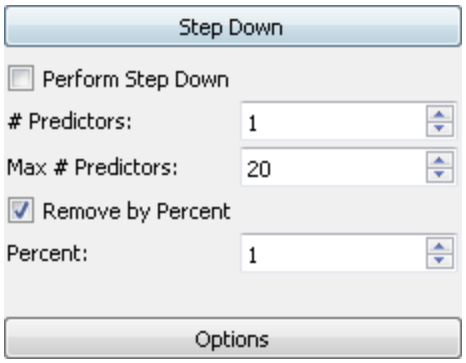


Figure 5-4: Step Down Options

5.3.1 SPECIFYING RANGE FOR # PREDICTORS

Check Box: 'Perform Step Down' Variable Selection Option

Check this box to utilize the Step-down option which yields a model containing a subset of the candidate predictors.

5.3.2 PREDICTOR SELECTION OPTIONS

'Min # Predictors' (default = 1)

You can specify the minimum number of predictors to include in the model.

'Max # Predictors' (default = 20)

You can specify the maximum number of predictors to include in the model. Enter a number greater than or equal to Min # Predictors.

Suppose you had 10,000 predictors to begin with. If you step down to say Min # = 1 predictor, but specify say Max # = 20 predictors, the final model will contain P* predictors where P* is between 1 and 20. The summary output at the bottom of the Model Output window as well as the associated graph will only report results for models containing between Min # and Max # predictors.

Note: The Cross-validation (CV) feature is used to determine the number of predictors P* to include in the model. If the CV feature is not activated, the step-down procedure yields a model with P* = Min # Predictors. For details of the CV option see Section 5.6.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Check Box: 'Remove by Percent'

In the case of large P, to reduce the amount of computational time, more than 1 predictor can be eliminated at each step. To use this option, specify the percentage of predictors to be removed at each step. The specified percentage of predictors will be removed at each step, until 100 predictors remain, at which time the step-down algorithm removes 1 predictor at a time.

By default, the percentage is set to 1%, meaning that if you had say 10,000 predictors to begin with, after 460 steps you have fewer than 100 predictors. Or, if you used 2%, after 229 steps you would be under 100 predictors. If not selected, the step-down algorithm removes 1 predictor at a time.

'Percent'

Specify the percentage of predictors to be removed at each step, until 100 predictors are left, at which time 1 predictor at a time is removed.

SELECT STEP-DOWN PREDICTORS (APPEARS ONLY AFTER A MODEL IS ESTIMATED)

Figure 5-5: Select Step-down Predictors Button

After you have estimated a model using the step-down feature, you will have the option to select only the predictors that were selected in the 'optimal' model—all previously checked predictors which were not included in the 'optimal' will be unchecked. If you check this button to select only the selected step-down predictors, the 'Perform Step Down' check box will automatically be unchecked.

Warning: It is not a valid use of cross-validation (CV) to perform CV AFTER selecting say 3 predictors using the 'Select Step-down Predictors' option, ignoring that these predictors were selected from a larger number of candidate predictors. This will yield a biased (invalid) CV-R^2. The correct way to obtain the CV-R^2 is to include all predictors, and specify that the maximum number of predictors = 3. For details of the CV option see Section 5.6.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

5.4 Options

5.4.1 SPECIFY # COMPONENTS (MAX # COMPONENTS)

Figure 5-6: Specify # Components

Max # Components: If the Automatic option is not checked, the 'Max # Components' field contains the number K of components (minimum value is 1). Under Options, click in the box to the right of 'Max # Components' to type the desired number of components. You can also click on the up and down arrows to increase or decrease the number of components by 1.

Automatic: When both the 'Automatic' and Cross-Validation options are activated, COREXpress estimates several models and provides a plot for the CV-R^2 (or NMSE) for all K-component models where K is less than or equal to the number specified in the 'Max Components' text box. Output in the Model Output window is provided for the K*-component model, where K* is the number of components associated with the model with the best CV fit statistic. The CV fit statistics depend upon the particular model type.

Note: Selection of Automatic will have no effect if the Cross-validation option is not also activated.

Note: To guard against an unreasonably large value for K being specified, prior to beginning this algorithm the value for K is reduced (as described below) to the smallest value obtained from application of the following criteria:

1) If K > N-1, K is reduced to N-1, where N is the number of observations in the estimation sample.
2) If K > P, K is reduced to P, where P is the total number of predictors
3) The following 2-step procedure is applied:
   Step 1: Initialize K'=2.
   Step 2: Estimate the K'-component model. Check to see if any of the stopping criteria a, b, or c are met (see below). If any of the criteria is met, set K to K'-1 and stop. Otherwise, increment K' to K' + 1. If K' = K stop. Otherwise return to step 2.

Criteria:

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

a) The tolerance for component S_1 in a regression of S_1 on the other components S_2, …, $S_{K'}$ exceeds 1E-6.
b) The singular value decomposition (SVD) algorithm applied to the component covariance matrix results in a singular values range that exceeds 1E+12.
c) The standard deviation for component K' exceeds 1E+10.

5.4.2 SELECT MODEL TYPE

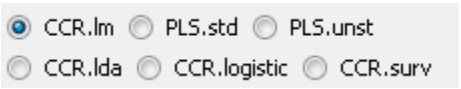


Figure 5-7: Model Type Options

CCR.lm (linear regression): continuous dependent variable, predictors assumed to be numeric (continuous, dichotomous, or discrete).

PLS.std* (PLS regression with standardized predictors): continuous dependent variable, predictors assumed to be numeric (continuous, dichotomous, or discrete).

PLS.unst* (PLS regression with unstandardized predictors): continuous dependent variable, predictors assumed to be numeric (continuous, dichotomous, or discrete).

CCR.lda (linear discriminant analysis):** dichotomous dependent variable, predictors assumed to follow a multivariate normal distribution with common variance-covariance matrix.

CCR.logistic (logistic regression):** dichotomous dependent variable, predictors assumed to be numeric (continuous, dichotomous, or discrete).

CCR-surv (survival): dichotomous dependent variable (event), with multiple records per case. Requires use of Case ID variable to identify cases. For details regarding the Case ID variable, see Section 5.4.5.

Notes:

*Unlike CCR which is scale invariant, PLS regression is scale dependent, meaning that different solutions are obtained depending upon whether predictors are maintained in their original scale (PLS.unst) or standardized to have variance 1 (PLS.std). When predictors are not all in the same units, it is generally recommended to standardize the predictors – this is done automatically as part of the PLS.std. model type option.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

** LDA and Logistic regression variants of COREXpress 1.0 are limited to dichotomous dependent variables in the current implementation. If the dependent variable has more than two values, the program automatically dichotomizes it at the mean.

5.4.3 TECHNICAL OPTIONS FOR LOGISTIC REGRESSION AND SURVIVAL MODEL TYPES

Occasionally logistic regression algorithms will not converge due to complete or quasi-complete separation between the 2 dependent variable groups. For example, with a single predictor X such that all cases in group 1 have higher X values than all cases in group 2, any regression coefficient for X will provide perfect separation of the 2 groups, and the maximum likelihood estimate for the X coefficient does not exist (it approaches infinity). Similar convergence problems may occur with near perfect separation.

To prevent such convergence problems, the CCR.logistic algorithm in COREXpress contains 2 control parameters -- the # iterations and the Ridge parameter.

5.4.3.1 Iterations

By default, the number of iterations is set to 4, which should be sufficient for most applications.

5.4.3.2 Ridge Parameter

The ridge regression penalty parameter has a default equal to .001. With no penalty (Ridge parameter = 0), the separation problems may cause nonconvergence, in which case increasing the number of iterations will yield larger and larger estimates for at least one regression coefficient. Using a sufficiently large penalty eliminates this non-convergence problem. Typically, the default parameter (.001) will be sufficiently large to prevent the non-convergence problem from occurring.

5.4.4 WEIGHTS

Individual records can be weighted differently according to the non-zero values provided in the optional Weights variable. When not specified, all records have a weight of 1. To assign a variable to be used to specify the weight, select that variable from the Weights drop down menu under Options section in the Model Control window.

5.4.5 CASE ID

If the data file contains only 1 record per case, no Case Id variable is required. For data files in which one or more cases have multiple records (e.g. repeated measures), a

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

variable must be assigned as a Case Id variable to uniquely identify each case. Use of a Case ID variable assures that all records associated with a given case are assigned to the same fold. (For further details regarding the use of folds in Cross-validation, see Section 5.6).

To assign a variable to be used to specify the Case IDs, select that variable from the Case ID drop down menu under Options section in the Model Control window.

5.4.6 INCLUDE MISSING

By default, missing data in COREXpress is handled using the 'list-wise deletion' option, which eliminates any case that has a missing value on any predictor. When the 'Include Missing' box is checked, missing data is handled by imputing the mean. Imputation is done for the training and the validation samples separately.

5.4.7 NMSE

For model types CCR.lm and CCR.pls, the Normed Mean Squared Error (NMSE) is included in a check-box as an alternative tuning criterion to the CV-R^2. NMSE is defined as the Mean Squared Error Divided by the Variance of Y. It should provide values that are greater than 0, and usually less than 1. Values greater than 1 indicate a poor fit in that the predictions (when applied to cases in the omitted folds) tend to be further from the observed Y than the baseline prediction provided by the observed mean of Y (a constant). To use CV-NMSE as the criterion to minimize, check the box to the left of 'NMSE'. If this box is not checked, the default criterion CV- R^2 will be maximized. These two criteria should give the same or close to the same solutions in most cases. CV-R^2 is computed as the square of the correlation between the predicted and observed dependent variable.

5.4.8 DECIMAL PLACES

This option controls the number of decimal places displayed for the results of model estimation. A value of 4 is used by default.

5.5 Specifying Training & Validation Subgroups

The Validation section of the Model Control window allows splitting the cases into Training and Validation samples, where only the cases selected to be in the training sample are used to estimate the model. To use this option, specify a criteria for defining the 'Training Subgroup'. By default, the cases not selected for the training subgroup constitute the 'Validation Subgroup'. This default option can be overridden by also defining a 'Validation Subgroup'. If both a training and validation subgroups are defined, cases meeting neither selection criteria are not used for model development nor for model validation.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

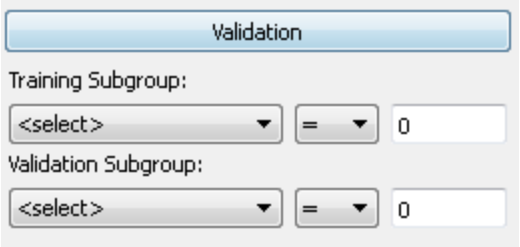


Figure 5-8: Training & Validation Subgroup Options

Selecting the Training Subgroup

By default, if no Training or Validation Subgroup is specified, all cases on the dataset will be used for analysis. If a Training Subgroup (an 'analysis' sample) is specified, only these cases will be used to develop the model. By default, all cases not specified in the Training Subgroup will be used as the Validation Subgroup. The Validation Subgroup cases are not used at all during model development, but are used to independently test the performance of the model developed on the Training ('analysis) Subsample.

- In the Model Control window, click on 'Validation' and options will appear for selecting training and validation sample cases.
- Under 'Training Subgroup', click on the '<select>' drop down menu and specify a variable to be used to select the subsample.
- Click on the '=' drop down drop down menu choose from the following operators:
  - = (equal to)
  - ~= (not equal to)
  - < (less than)
  - > (greater than)
  - <= (less than or equal to)
  - >= (greater than or equal to)
- Click in the Training Subgroup numeric box and specify number to be used to specify the subgroup.
- You can use multiple variables to specify the training subsample. To use additional variables, click on '<select>' in the 2nd row and continue to specify the subsample

The 'Validation Subsample' can be specified in the same manner as the Training Subsample.

By default, when the Cross-validation option is used, if a Training Subsample is specified, all other records on the data file are used for the Validation Subsample. Cross-validation cannot be performed when a Validation Subgroup is specified.

After estimating a model, the model output will report the number of cases used in the Training Subsample and the number of cases in the Validation Subsample.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

5.6 Cross-Validation (CV) Options

CV statistics are used to determine the number of predictors P* in a model when the Step-down option is activated, and to determine the number of components K* when the Automatic option is used.

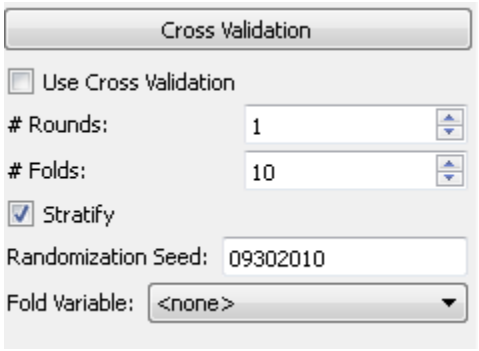


Figure 5-9: Cross-Validation Options

Use Cross Validation Box: Check this box to use the Cross Validation option

5.6.1 # ROUNDS:

By default, R = 1 round of M-fold CV is performed. When 2 or more rounds are requested, standard errors are also provided among the CV fit statistics.

5.6.2 # FOLDS:

By default, cases are allocated among M = 10 folds. R fold assignment variables are appended to the data file. To view these variables, see Section 4.1). Generally speaking, it is recommended that the number of folds be between 5 and 10, and if possible, to divide evenly into the sample size N. Thus, for example, if N = 24, M might be selected to be 6 or 8.

Note: The fold assignments will differ if you change the order of the cases on your datafile.

5.6.3 STRATIFY:

When using the cross-validation option, COREXpress assigns cases randomly to each fold, and if the 'Stratify' option is selected, cases in each dependent variable category will be uniformly distributed (or as close as possible to uniform) among each fold.

Note that after estimating CCR.lm, COREXpress automatically removes the checkmark from the Stratify CV option, which is not applicable to model types CCR.lm or PLS.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

5.6.4 RANDOMIZATION SEED:

By default, when the Cross-Validation box is checked, the program will divide the cases into folds using the default random seed (09302010). Changing the seed alters the fold assignments. The seed is interpreted as a character string and not as a number: it can contain any arbitrary characters.

5.6.5 FOLD VARIABLE:

Use the 'Fold Variable' option to determine fold assignments based on a 'fold' variable that you already have on your data file. Otherwise, COREXpress will automatically generate a fold variable, which can be seen on the datafile viewer (scroll to the right) by double clicking on the dataset from the Projects window. A fold variable contains positive integer values 1, 2, …, M where M = # folds.

5.7 Screen

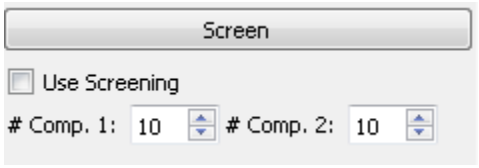


Figure 5-10: Screen Options

You can select the number of non-zero loadings for component #1 (C1), and also set an upper limit on the number of non-zero loadings for component 2 (C2). For example, if C1=5 and C2=4, component 1 will have 5 non-zero loadings and component 2 will have between 5 and 9 non-zero loadings. The 5 predictors that have non-zero loadings on component 1 will also always be allowed to have non-zero loadings on component 2 along with the 4 predictors having the largest magnitude for the component 2 loadings. (In the rare case that these 4 predictors are among the 5 with non-zero loadings on component 1, the total number of predictors with non-zero loadings on component 2 will only be 4).

5.8 Estimate the Model

Once you have specified your model, to estimate the model you may:

- Click 'Estimate' (located at the bottom of the Model Control window)

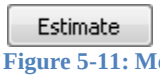


Figure 5-11: Model Estimator Button

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

If you opened a previously saved .spp file, you must click 'Estimate' and re-estimate the model(s) in order to view any previously saved Plot output.

5.8.1 CANCELING MODEL ESTIMATION

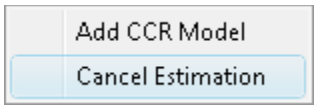


Figure 5-12: Model Menu

You can cancel model estimation once it has begun estimating. Once estimation has been canceled, you can edit the model specifications and estimate a new model. Note that once an estimation is canceled, it cannot be resumed from the cancelation point—the estimation must restart from the beginning.

Once the model estimation has begun,

- To cancel the estimation procedure, click 'Model' from the Tool Bar and then click 'Cancel Estimation'.

Restarting Model Estimation

- To restart model estimation, click **'Estimate'**.

5.9 Model and Output Management

Opening a New Model

After clicking 'CREATE NEW MODEL', COREXpress automatically generates a new 'CREATE NEW MODEL' option in the Projects window (it will be directly below the last specified model for a data file).

To setup a new model,

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- Double click on 'CREATE NEW MODEL' in the Projects window under the datafile

Model setup options will appear in the Model Control window for the new model.

OR..

- Click 'Model' from the menu options
- Click 'Add CCR Model'

Model setup options will appear in the Model Control window for the new model.

CHANGING MODEL NAMES

To customize the name for your model:

- Right click the model name (for example, 'CCR 1') in the Projects window
- Click 'Rename' from the new option menu
- Type the desired model name
- Hit 'Enter' on your keyboard

Model names can be edited before or after the model is estimated. The new model name will now appear at the top of the associated Model Output window, Model Control window, and plots.

DELETING MODELS

To delete a previously estimated model:

- Right click the model name (for example, 'CCR 1') in the Projects window
- Click 'Remove' from the new option menu

DELETING PROJECTS

To delete a project and all associated models and output:

- Double click on the desired data file name (or associated models) in the Projects window to activate the project
- Click 'File' → 'Close Project'
- You will be prompted to close the project. Click 'Yes' to close the project or 'No' to return to the project.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

DEFINING A NEW MODEL

Once you have estimated at least one model, there are several ways to specify a new model on the same data file.

For any previously estimated model, double click on the model name (e.g. CCR1…CCR2, etc.) from the Projects window to make the model active. The Model Control window will now show model options for the active model. You can change any model settings in the Model Control window and click 'Estimate' to re-estimate the model based on the new model settings.

6 Model Output and Plots

6.1 Output Created

When the Cross-validation Option and the Step-down Options are used, depending upon the model specified, a CV-ACC/CV-AUC (cross-validated accuracy and Area Under the ROC Curve), CV-R^2, or CV-AUC Plot is generated and opened in a new window.

6.1.1 CROSS-VALIDATION PREDICTOR GRAPHS

For model types CCR.lda and CCR.logistic, the CV-Predictor graphs plots CV-AUC and CV-ACC based on the specified K-component model as a function of the number of predictors P ranging from the specified 'Max # Predictors' down to the specified 'Min # Predictors' (See Section 5.3 for Step-down Option details).

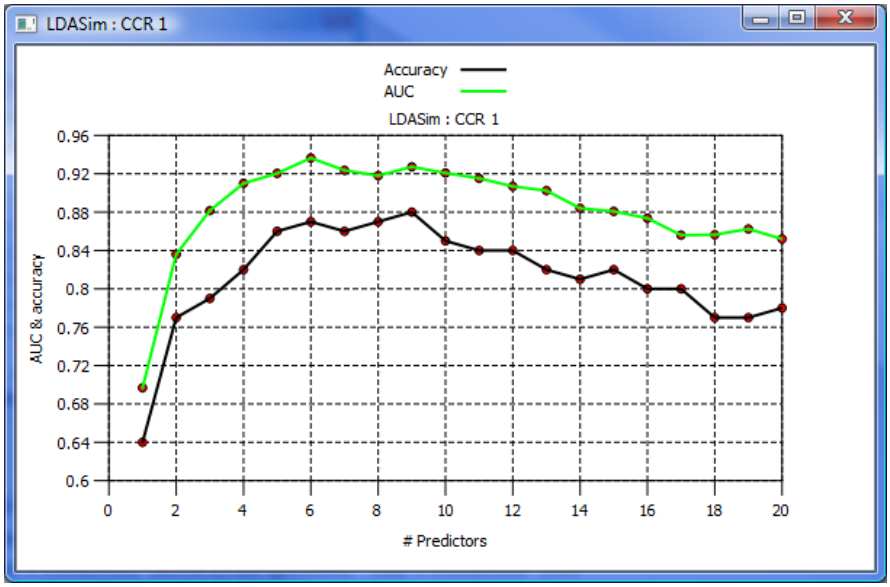


Figure 6-1: CV-AUC/ACC Plot

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

For model types CCR.lm and CCR.pls, the CV-R^2 graph corresponds to the cross-validation R^2 based on the specified K-component model as a function of the number of predictors P ranging from the specified ‘Max # Predictors‘ down to the specified ‘Min # Predictors‘ (See Section 5.3 for Step-down Option details).

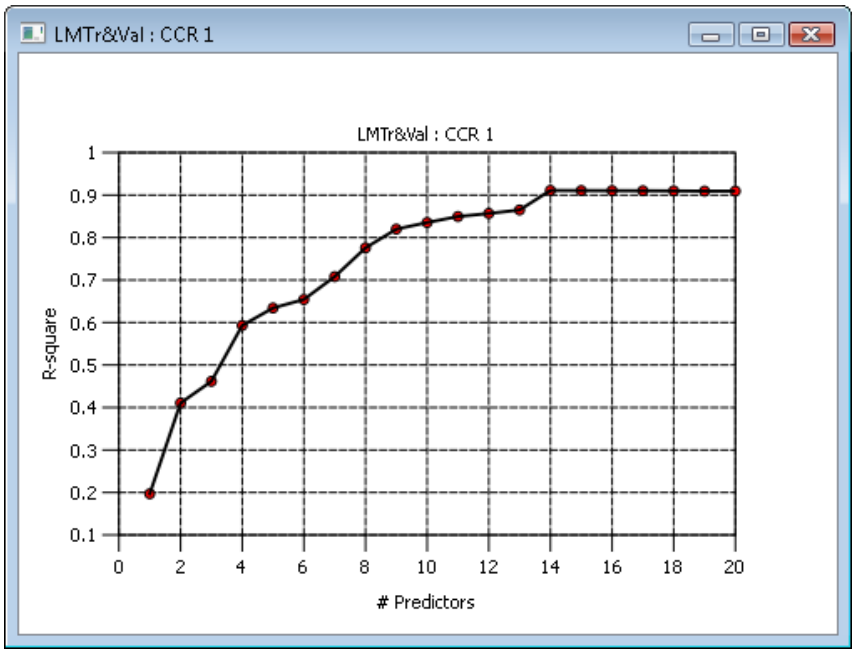


Figure 6-2: CV-R^2 Plot

For model type CCR.surv, the CV-AUC graph corresponds to the cross-validation AUC based on the specified K-component model as a function of the number of predictors P ranging from the specified ‘Max # Predictors‘ down to the specified ‘Min # Predictors‘ (See Section 5.3 for Step-down Option details).

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

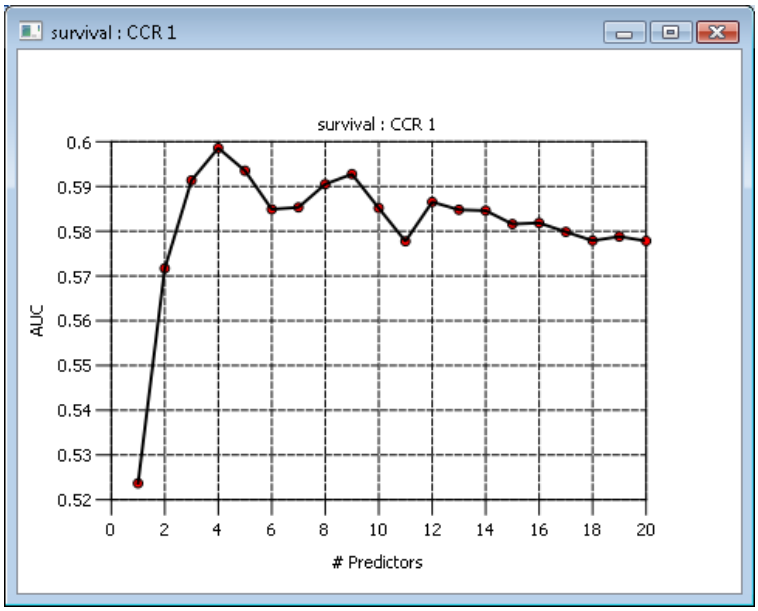


Figure 6-3: AUC Plot

6.1.2 INTERACTIVE SCATTER PLOT AND ROC CURVE

For model types CCR.lda, CCR.logistic, and CCR-surv, two interactive plots are available -- one for the training subgroup, the other for the validation subgroup.

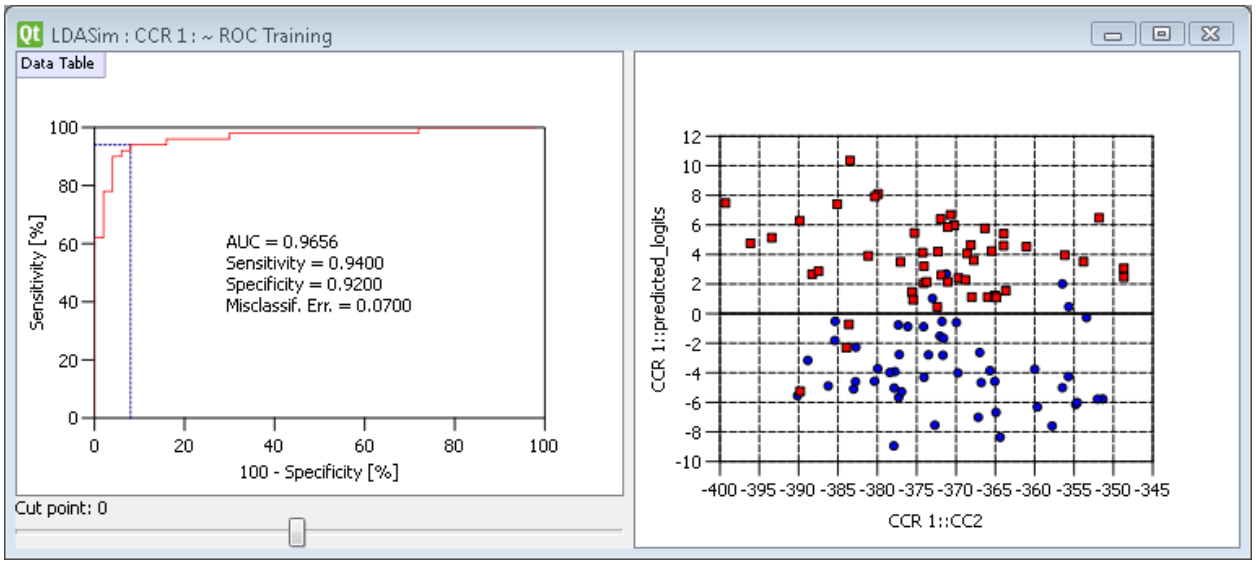


Figure 6-4: Training Dataset ROC & Scatter Plot

Each point on the red ROC curve corresponds to a particular logit cut-point depicted by a horizontal reference line in the associated scatter plot. 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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

horizontal equal-probability reference line in the scatter plot) are predicted to be in group 1 (category 1 of the dependent variable), those below the cut-point being predicted to be in group 2.

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.

The slider, located in the Model Control window beneath the Cutpoint box can be used to see how the sensitivity and specificity changes with different cut-points. To change the cut-point:

- Position the curser on the slider, left-click and move it to the right to raise the reference line to a higher cut-point or to the left to lower the reference line to a lower cut-point.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

6.1.3 MODEL SUMMARY OUTPUT

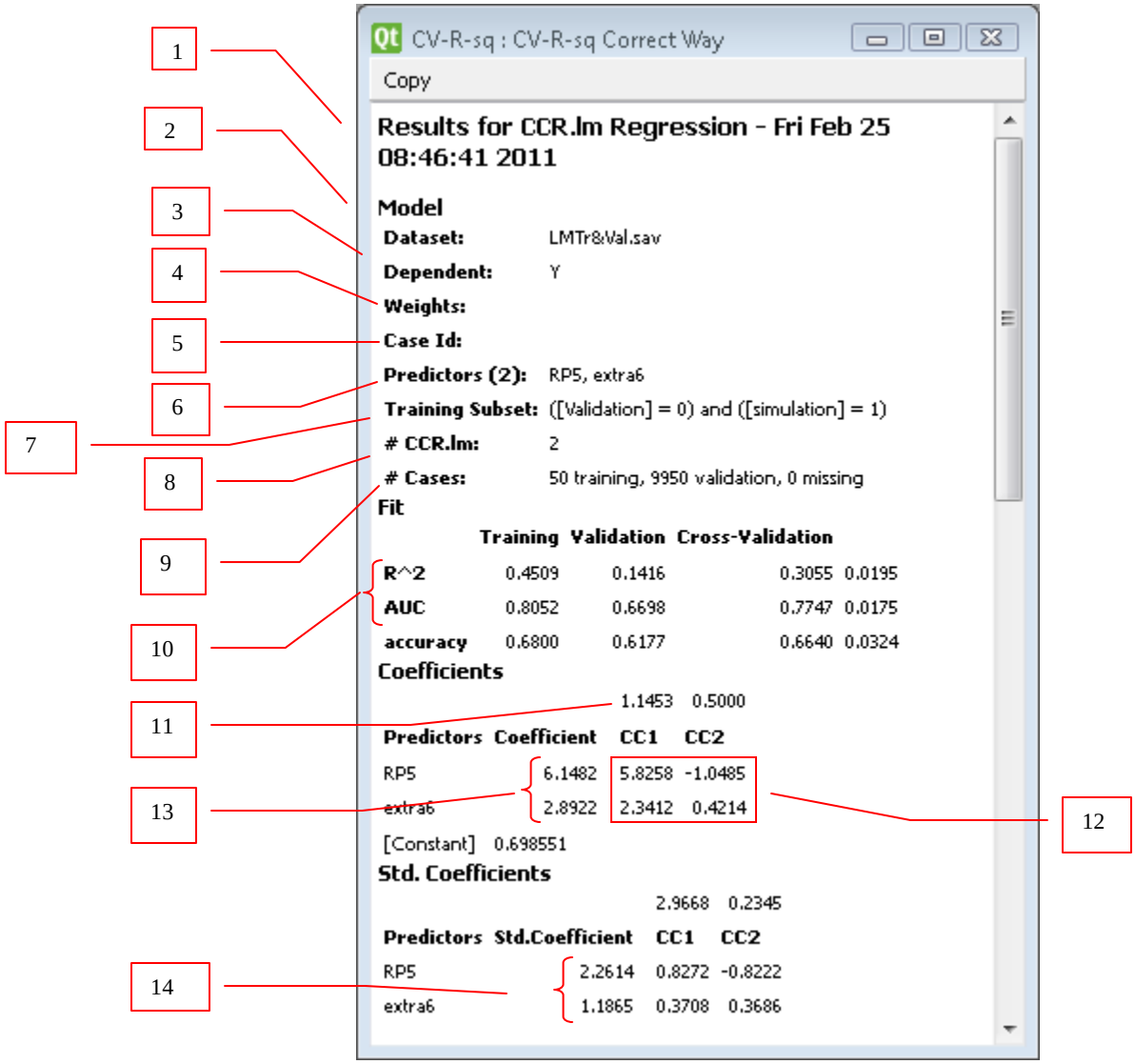


Figure 6-5: Model Summary Output

1. Type of Model and Date Estimated

This section of the output states the type of model specified (CCR.lm, CCR.lda, etc.) and the date that it was estimated.

2. Dataset Used

This section of the output includes the name of the dataset used to estimate the model. It may also include the directory path where the dataset is stored on your computer.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

3. Dependent Variable

Shows the dependent variable used for estimation of the model.

4. Optional Weight Variable

If a case weight is used, that variable is listed here.

5. Optional Case ID

If a case ID is used, that variable is listed here.

6. # Predictors (P) Included in Model

The number of predictors in the final model will be shown in parenthesis. A list of the names of the final predictors in the model is shown to the right of the # predictors.

7. Optional Subgroup for Training Data

If a training subgroup is defined, the criteria used to define it are shown here. If a validation subgroup is defined, the criteria used to define it are shown here.

8. # Components (K) Used

9. # Cases in Training Sample & Validation Sample (if Validation subgroup is defined)

10. R^2, AUC, and ACC for Training Sample and Optionally for Validation & Cross Validation

If a subgroup of cases is specified to be used as the training sample (see Section 5.5.) these output statistics are also displayed for the validation subgroup in a column labeled 'Validation'.

If the 'Cross Validation' option is used, additional output is displayed in a column labeled 'Cross-Validation'.

For model types CCR.lm and PLS, if the 'NMSE' box is checked, normed mean squared error (NMSE) is reported in addition to R^2. NMSE is defined as the Mean Squared Error (MSE) divided by the variance of the dependent variable. For the NMSE reported in the 'Validation' column, the variance of the dependent variable is computed based on the validation sample. For the NMSE reported in the 'Training' and 'Cross-validation' columns, the variance of the dependent variable is computed based on the training sample.

The CV- R^2 reported in the model summary output is the average of the CV-R^2(P*.r) across the rounds. For round r, the OPTIMAL NUMBER OF PREDICTORS P*.r, is determined for that round, and an average is computed of these CV- R^2(P*.r). Thus, the CV-R^2 listed among the fit statistics at the top is always at least as large as the one shown in the Cross-Validated Step-down

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

table associated with P* predictors, which is the one used to select the number of predictors to display in the output.

Std. Error for Cross-validated R^2, AUC, and ACC

If more than one round is specified in a Cross Validation, the standard error is reported to the right of the associated output statistics.

Note: For model type PLS.std, the predictors are standardized to have variance 1. In this case, if Cross Validation is performed, during the cross-validation procedure, the predictors are re-standardized within each fold subgroup.

11. Component Weights

The predicted value for Y can be computed as a linear function of the components where these weights are applied to the components. For example,

Y hat = 0.698551 + 1.1453 * CC1 + 0.5 * CC2 (1)

See Section 8 for detailed definitions of the components and component weights.

12. Predictor Loadings

Unstandardized loadings are used to compute the components. For example, from the CCR.lm model output shown in Figure 6-5 (see above) is:

CC1 = 5.8258 * RP5 + 2.3412 * extra6 (2.1)

CC2 = -1.0485 * RP5 + 0.4214 * extra6 (2.2)

For CCR.lda and CCR.logistic, CC1 and CC2 are in logit units.

See Section 8 for detailed definitions of the relationships between the predictors and components, as quantified by the predictor loadings.

13. Unstandardized Regression Coefficients

Unstandardized regression coefficients are used to predict the dependent variable Y. For example, from the CCR.lm model output shown in Figure 6-5 (see above),

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

using (2.1) and (2.2) to substitute for CC1 and CC2 in (1), we obtain:

Y hat = 0.698551 + 6.1482 * RP5 + 2.8922 * extra6 (3)

where 6.1482 = 1.1453 * 5.8258 + 0.5 * (-1.0485)

and 2.8922 = 1.1453 * 2.3412 * +0.5 * 0.4214 as shown in Equation A.4 (see Section 8)

for CCR.lda and CCR.logistic, predictions are for logit(Y).

For PLS.std, predictors are standardized by dividing by their standard deviation. The unstandardized regression coefficients reported are for the *standardized* predictors.

14. Standardized Regression Coefficients

Standardized regression coefficients are used to assess the importance of the predictors, predictors with the highest magnitude being the most important. Each standardized regression coefficient equals the corresponding unstandardized coefficient multiplied by the ratio std(X_g)/std(Y), where 'std' denotes standard deviation. For further details, see Section 8.

For PLS.std, predictors are standardized by dividing by their standard deviation, so that std(X_g) = 1 for each predictor g =1,2,…,P. The standardized regression coefficient in this case equals the corresponding unstandardized coefficient reported divided by std(Y).

6.1.4 CROSS-VALIDATED COMPONENT TABLE

Optional output produced when 'Automatic' option is used with the specification of a maximum number of components (see Section 5.4.1).

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

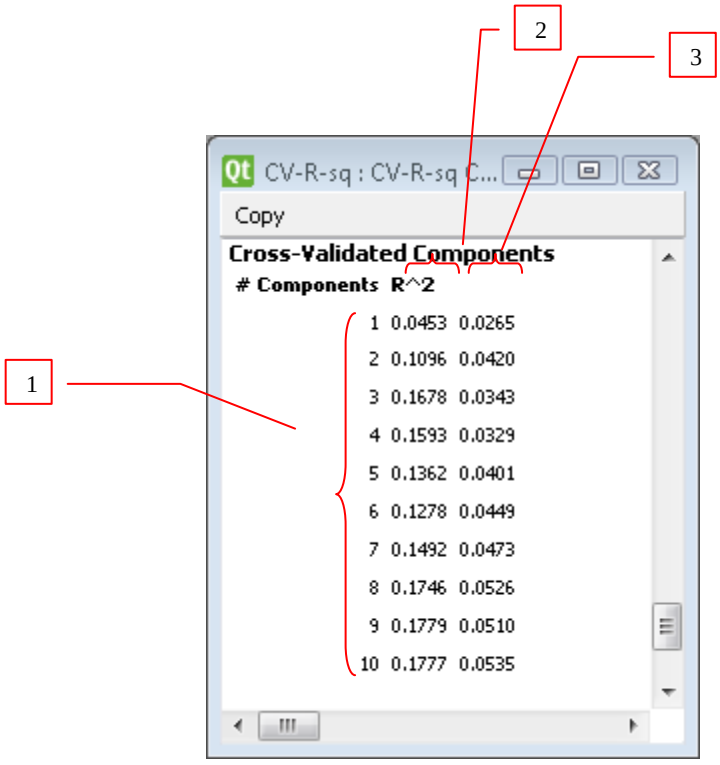


Figure 6-6: Cross-Validated Component Table

1. Number of Components (Column 1)

2. CV-R^2 (Column 2)

3. Standard Error of CV-R^2 (Column 3): Optional output produced when R>1 rounds of M-folds are requested.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

6.1.5 CROSS-VALIDATED STEP-DOWN TABLE

Optional output produced when 'Step-down' option is used with the specification of a maximum number of components (see Section 5.4.1).

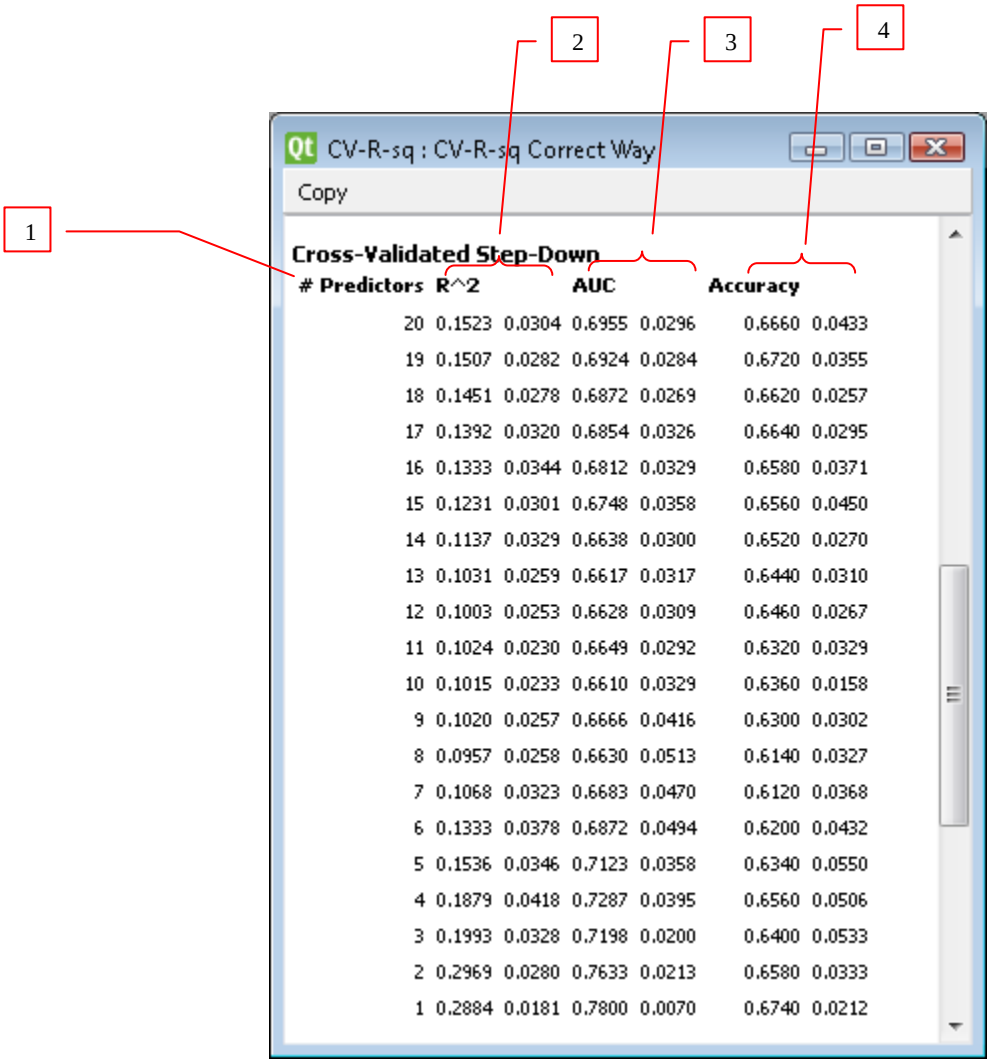


Figure 6-7: Cross-Validated Step-Down Table

1. Number of Predictors

2. R^2 and Standard Error: Standard Error is optional output produced when R>1 rounds of M-folds are requested.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

$CV\text{-}R^2(P) = Avg(r)\ CV\text{-}R^2_r(P)$

$CV\text{-}R^2(P^*.r) = Avg(r)\ Max[P]\ CV\text{-}R^2_r(P)$

The 2 CV-R^2 statistics measure different things.

The CV- R^2 reported in the Cross-validated Step-down table, called CV-R^2 (P), is the average CV- R^2(P) across all rounds, for models containing that number of predictors P.

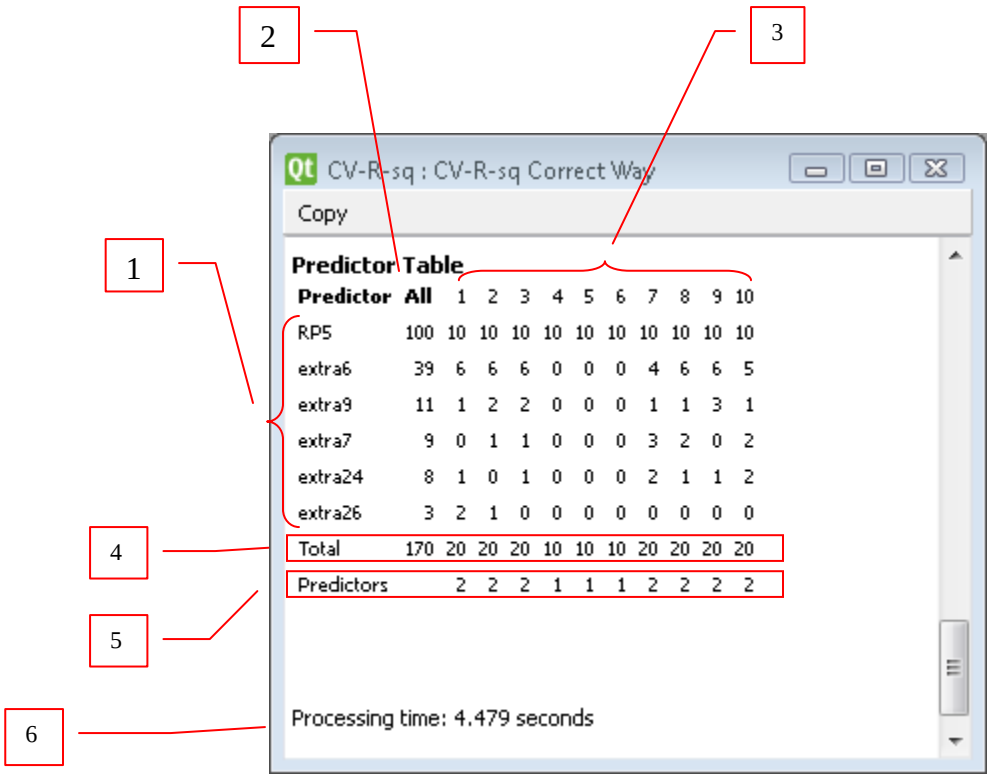
The CV- R^2 reported in the model summary output is the average of the CV- R^2(P*.r) across the rounds. For round r, the optimal number of predictors P*.r, is determined for that round, and an average is computed of these CV- R^2(P*.r). Thus, the CV-R^2 in the model fit statistics output displayed at the top is always at least as large as the one in the Cross-validation Step-down table displayed at the bottom. The CV-R^2 in the Cross-validation Step-down table is the one used by default to select the number of predictors.

3. AUC and Standard Error: Standard Error is optional output produced when R>1 rounds of M-folds are requested.

4. Accuracy and Standard Error: Standard Error is optional output produced when R>1 rounds of M-folds are requested.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

6.1.6 CROSS-VALIDATION 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
extra6	39	6	6	6	0	0	0	4	6	6	5
extra9	11	1	2	2	0	0	0	1	1	3	1
extra7	9	0	1	1	0	0	0	3	2	0	2
extra24	8	1	0	1	0	0	0	2	1	1	2
extra26	3	2	1	0	0	0	0	0	0	0	0
Total	170	20	20	20	10	10	10	20	20	20	20
Predictors		2	2	2	1	1	1	2	2	2	2

Figure 6-8: Cross-Validated Predictor Table

1. Predictors

2. Total for Each Predictor: The number of times the predictor is selected in a model where 1 of the M-folds is eliminated (maximum = M).

3. Round #: r = 1, 2, …, R

4. Total: Sum of the totals for a given round (= M * P_r)

5. Optimal # of Predictors Selected in Each Round (P_r)

6. Processing Time

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Cross-Validated Step-Down

Renaming, Copying, and Deleting Models and their Associated Output/Plots

Right clicking on a model directly below the project/dataset in the Project window provides the following options:

'Rename'
Allows you to rename the model.

'Copy'
Click to copy the model settings and create a new model at the end of the list of models below the project/dataset.

'Remove'
Deletes the model and all associated model settings & plots.

6.2 Plots

A wide range of traditional plot options are available. To see these, double click 'CREATE NEW PLOT' from the Project window. Alternatively, click on Plot from the Menu Bar at the top and select the desired plot. Once you select your desired plot, the Plot Control window will appear on the right-hand side of the program in place of the Model Control window.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

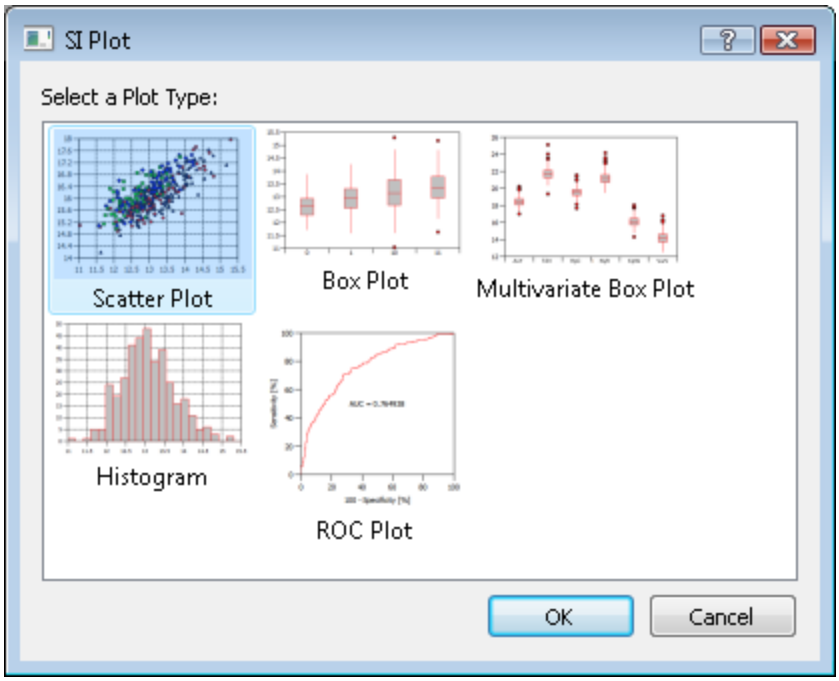


Figure 6-9: Create New Plot Options

In particular:

Select the **Scatter Plot** option to construct additional scatter plots for the training data only, for the validation data only, for any selected subgroup 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 **Multivariate Box Plot** to compare the distributions of 2 or more variables.

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

Select the **ROC plot** option to customize a new ROC plot.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

6.2.1 SCATTER PLOT

To add a Scatter Plot, from the menu bar click 'Model'→ Add Scatter Plot. On the right-hand side of the application, the Scatter Plot Control window will appear with options to specify a scatter plot.

Figure 6-10: Scatter Plot Control Window

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Variables: Specify the X- and Y-axis variables to be plotted. All changes will be updated automatically.

Markers: You can change the color, shape, and size of the markers plotted. Using the 'Var:' drop down menu, you can specify a variable to distinguish the markers in the plot. All changes will be updated automatically.

Labels: You can edit the labels for the X- and Y-axis. You can also include a title for the plot. All changes will be updated automatically.

Gridlines: Uncheck the option boxes to hide the X or Y gridlines.

Subgroup: You can choose to only show a subgroup of your data in the plot. By default, if subgroup is specified, all cases on the dataset will be shown in the plot. If a subgroup is specified, only these cases will be shown in the plot.

- In the Scatter Plot Control window, click on 'Subgroup' and options will appear for selecting a subgroup.
- Click on the '<select>' drop down menu and specify a variable to be used to select the subsample.
- Click on the '=' drop down drop down menu choose from the following operators:
  - = (equal to)
  - ~= (not equal to)
  - < (less than)
  - > (greater than)
  - <= (less than or equal to)
  - >= (greater than or equal to)
- Click in the Subgroup numeric box and specify number to be used to specify the subgroup.
- You can use multiple variables to specify the training subsample. To use additional variables, click on '<select>' in the 2nd row and continue to specify the subsample

Update: After specifying a subgroup, click 'Update' to view the changes in the plot.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

6.2.2 BOX PLOT

To add a Box Plot, from the menu bar click 'Model'→ Add Box Plot. On the right-hand side of the application, the Box Plot Control window will appear with options to specify a box plot.

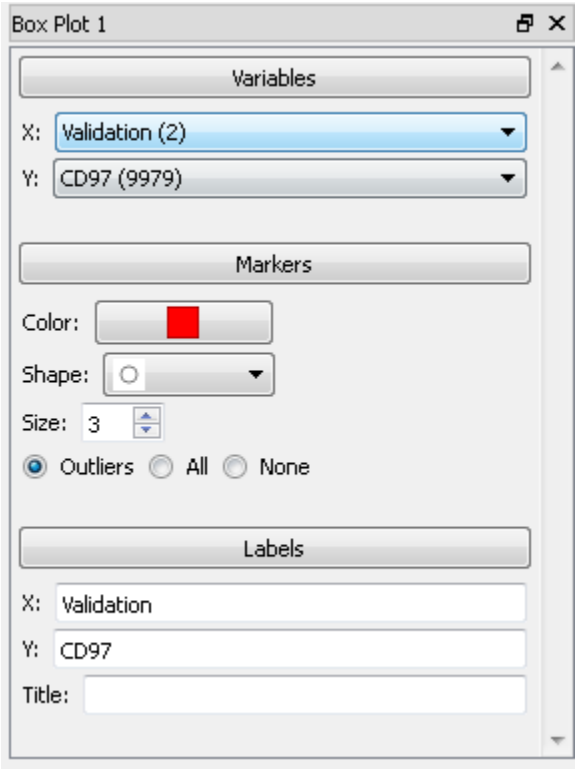


Figure 6-11: Box Plot Control Window

Variables: Specify the X- and Y-axis variables to be plotted. All changes will be updated automatically.

Markers: You can change the color, shape, and size of the markers plotted. Using the 'Var:' drop down menu, you can specify a variable to distinguish the markers in the plot. All changes will be updated automatically.

Labels: You can edit the labels for the X- and Y-axis. You can also include a title for the plot. All changes will be updated automatically.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

6.2.3 MULTIVARIATE BOX PLOT

To add a Multivariate Box Plot, from the menu bar click 'Model'→ Add Multivariate Box Plot. On the right-hand side of the application, the Multivariate Box Plot Control window will appear with options to specify a box plot.

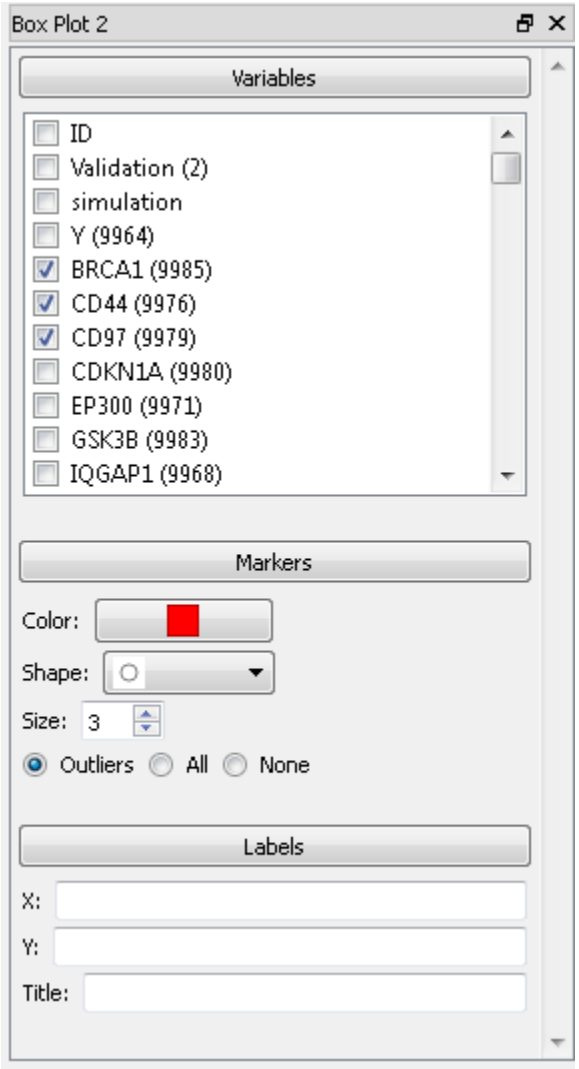


Figure 6-12: Multivariate Box Plot Control Window

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Variables: To select a variable, check the box next to the variable name. All changes will be updated automatically.

Markers: You can change the color, shape, and size of the markers plotted. Using the 'Var:' drop down menu, you can specify a variable to distinguish the markers in the plot. All changes will be updated automatically.

Labels: You can edit the labels for the X- and Y-axis. You can also include a title for the plot. All changes will be updated automatically.

6.2.4 HISTOGRAM

To add a Histogram, from the menu bar click 'Model'→ Add Histogram. On the right-hand side of the application, the Histogram Control window will appear with options to specify a histogram.

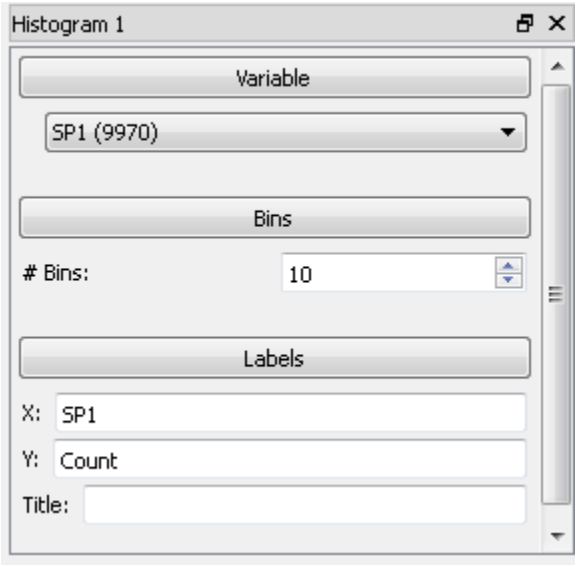


Figure 6-13: Histogram Control Window

Variable: Use the drop down menu to select the variable to be shown in the histogram. All changes will be updated automatically.

Bins: In the '# Bins' numeric box, you can specify the number of bins. All changes will be updated automatically.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Labels: You can edit the labels for the X- and Y-axis. You can also include a title for the histogram. All changes will be updated automatically.

6.2.5 ROC PLOT

To add a ROC Plot, from the menu bar click 'Model'→ Add ROC Plot. On the right-hand side of the application, the ROC Plot Control window will appear with options to specify a ROC Plot.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

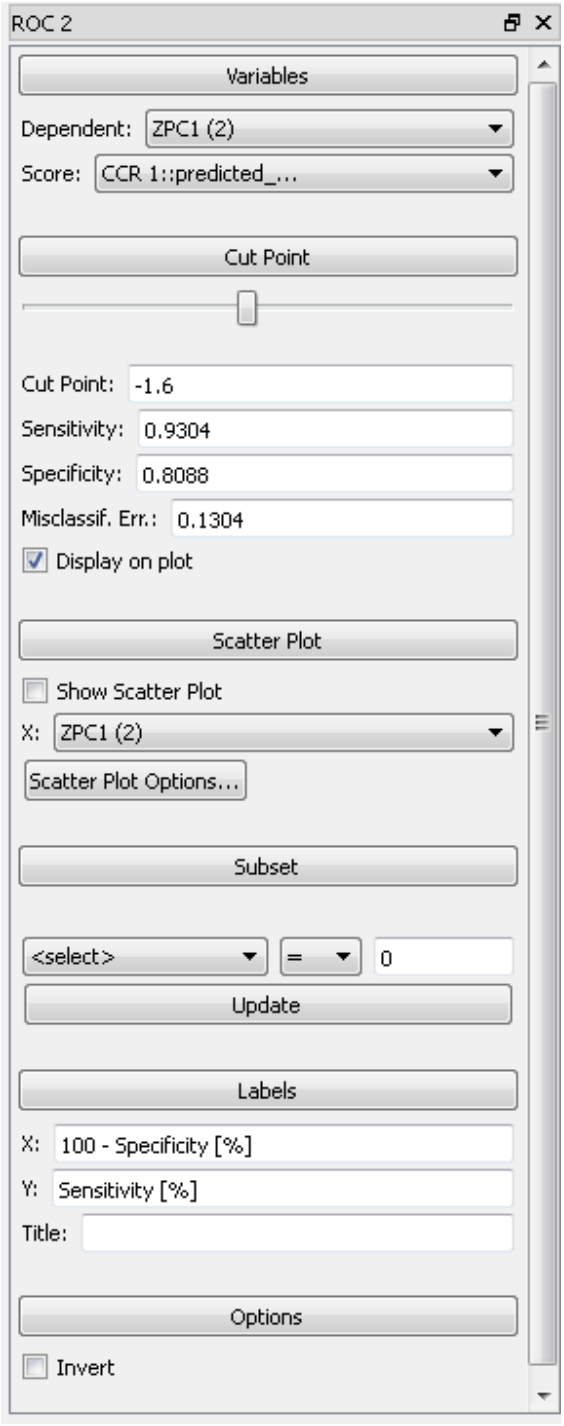


Figure 6-14: ROC Control Window

Variables

Dependent: Use the drop down menu to select the dependent variable.

Score: Use the drop down menu to select the score variable.

All changes will be updated automatically.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Cutpoint

Sliding Cutpoint Bar: Position the curser on the slider, left-click and move it to the right to raise the reference line to a higher cut-point or to the left to lower the reference line to a lower cut-point.

Cut Point: By default, the cut point will be 0. You can type a specific cut point in the 'Cut Point' numeric box.

Sensitivity: The sensitivity (correct classification rate for dependent variable group 1) numeric box cannot be edited. The sensitivity will automatically update as the cut point is changed.

Specificity: The specificity (correct classification rate for dependent variable group 2) numeric box cannot be edited. The specificity will automatically update as the cut point is changed.

Misclassif. Err.: The misclassification error equals 1 - overall correct classification rate. The numeric box cannot be edited. The misclassification error will automatically update as the cut point is changed.

Display on plot (checkbox): Not yet operational.

Scatter Plot

Show Scatter Plot (checkbox): Check the box to show a corresponding scatter plot.

X: Use the drop down menu to specify the X-axis variable for the scatter plot.

Scatter Plot Options…: Click to view the scatter plot options. See description of Scatter Plot above for more details.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

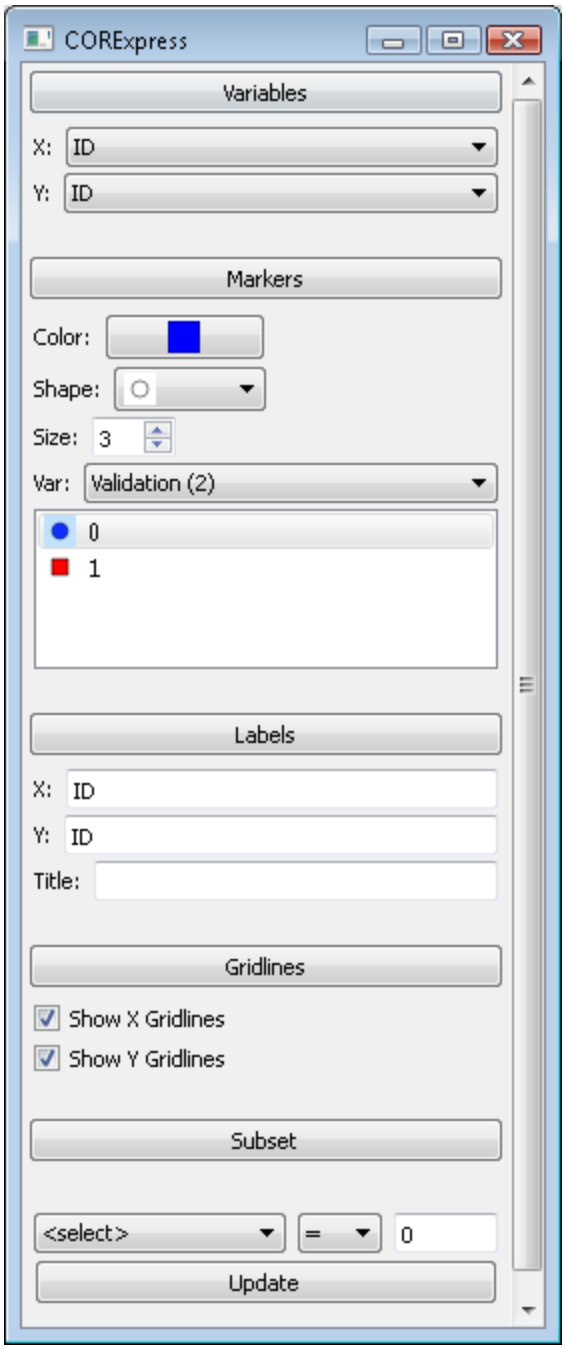


Figure 6-15: Scatter Plot Options

For Scatter Plot Options details, see Scatter Plot description above.

Subgroup: You can choose to only show a subgroup of your data in the plot. By default, if subgroup is specified, all cases on the dataset will be shown in the plot. If a subgroup is specified, only these cases will be shown in the plot.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- In the Scatter Plot Control window, click on 'Subgroup' and options will appear for selecting a subgroup.
- Click on the '<select>' drop down menu and specify a variable to be used to select the subsample.
- Click on the '=' drop down drop down menu choose from the following operators:
  - = (equal to)
  - ~= (not equal to)
  - < (less than)
  - > (greater than)
  - <= (less than or equal to)
  - >= (greater than or equal to)
- Click in the Subgroup numeric box and specify number to be used to specify the subgroup.
- You can use multiple variables to specify the training subsample. To use additional variables, click on '<select>' in the 2nd row and continue to specify the subsample

Update: After specifying a subgroup, click 'Update' to view the changes in the plot.

Labels: You can edit the labels for the X- and Y-axis. You can also include a title for the plot. All changes will be updated automatically.

Options:

Invert (checkbox): Check the box to invert the ROC curve.

7 Generating Predictions

The dataset view was described earlier in Section 3.4. Following model estimation, the predictions for the dependent variable can be appended to the dataset window as follows:

If the dataset window is currently open, close it. Then, reopen it by double clicking on the datafile name in the Projects window. Predictions now appear as additional variables, for both the training and validation cases. Predictions for the validation cases are based on the coefficients reported in the Model Summary Output window, which are based on the training cases.

For example, the prediction of the dependent variable Y obtained from the CCR.lm model for which the output is shown in Figure 7-1 (see below) is:

Y hat = 0.698551 + 6.1482 * RP5 + 2.8922 + extra6

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

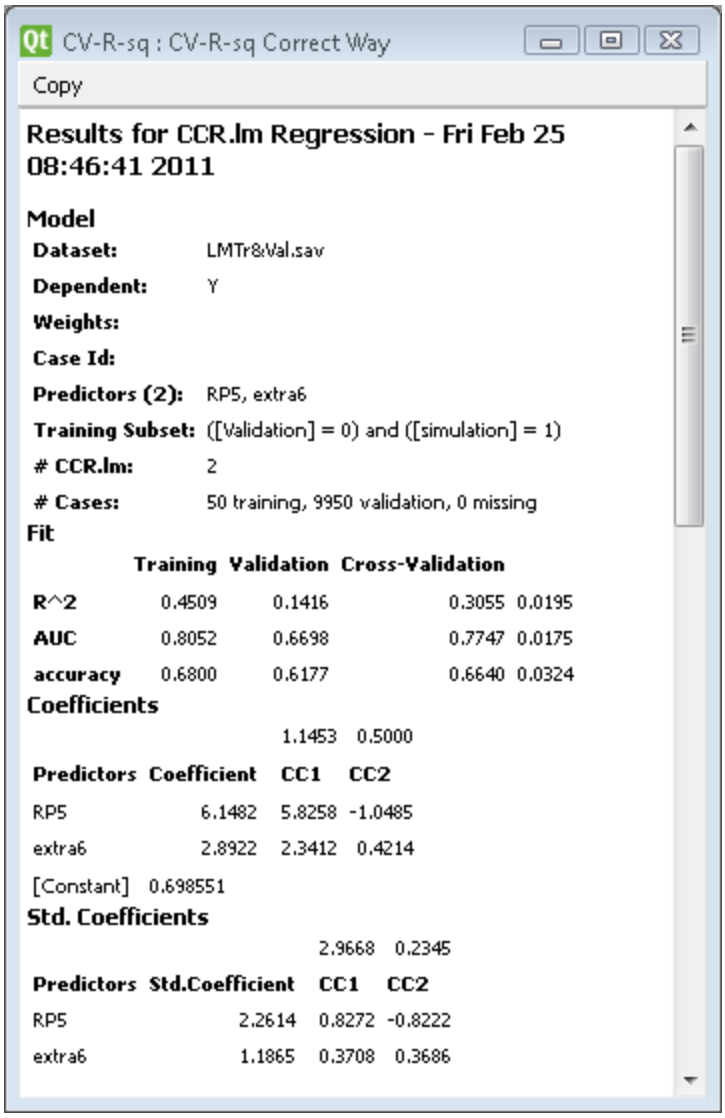


Figure 7-1: Model Output Summary

8 *Summary of Correlated Component Regression (CCR)*

The regression procedures available in the Correlated Component Regression (CCR) module produce reliable predictions from data with P correlated explanatory (X) variables, where multicollinearity may exist and P can be greater than the sample size N. The procedures are based on Generalized Linear Models (GLM). As an option, the CCR step-down variable selection option may be activated to exclude irrelevant Xs.

The linear part of the model is a weighted average of K components $S = (S_1, S_2, \ldots, S_K)$ where each component itself is a linear combination of the predictors. When the dependent variable Y is *continuous*, these procedures provide a reliable alternative to traditional *linear* regression

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

methods, where components may be correlated (CCR-LM procedure), or restricted to be uncorrelated with component weights obtained by PLS regression techniques (CCR-PLS). For Y *dichotomous*, these procedures produce a reliable alternative to *Logistic* regression (CCR-Logistic) and *linear discriminant analysis* (CCR-LDA).

Traditional maximum likelihood regression methods, which employ no regularization at all, can be obtained as a special case of these models when K=P (the *saturated* model). Regularization, inherent in the CCR procedures, reduces the variance (instability) of prediction and also lowers the mean squared error of prediction when the predictors have moderate to high correlation. The smaller the value for K, the more regularization is applied. Typically, K will be less than 10 (quite often K = 2, 3 or 4) regardless of P. M-fold cross-validation techniques are available to determine the amount of regularization to apply.

When the CCR step-down option is activated with M-fold cross-validation, the optimal number of predictors, P*, will also be determined. Output includes a table of predictor counts, reporting the number of times each predictor is included in the final model. This can be used as an alternative measure of variable importance (Tenenhaus, 2010), as a supplement to the standardized regression coefficients. Additional options can limit the number of predictors to be included in the model.

The 4 regression procedures in the XLSTAT-CCR module differ according to the assumptions made about the scale type of the dependent variable Y (continuous vs. dichotomous), and the distributions (if any) assumed about the predictors.

Linear regression (CCR-LM, PLS)

Predictions for the dependent variable Y based on the linear regression model are obtained as follows:

$$\hat{Y} = S(S'DS)^{-1}S'DY \qquad \text{(A.1)}$$

where D is a diagonal matrix with case weights as the diagonal entries.

With K=2 components we have:

$$\hat{Y} = \alpha + b_{1.2}S_1 + b_{2.1}S_2 \qquad \text{(A.2)}$$

where $b_{1.2}$ and $b_{2.1}$ are the component weights, the components being given by:

$$S_1 = \sum_{g=1}^{P} \lambda_{g.1} X_g \quad \text{and} \quad S_2 = \sum_{g=1}^{P} \lambda_{g.2} X_g$$

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

and $\lambda_{g.1}$ and $\lambda_{g.2}$ are component coefficients (loadings) for the gth predictor on components S_1 and S_2 respectively.

The component weights and loadings are obtained from traditional OLS regression. By substitution we get the reduced form expression:

$$\hat{Y} = \alpha + \sum_{g=1}^{P} (b_{1.2}\lambda_{g.1} + b_{2.1}\lambda_{g.2}) X_g \quad \text{(A.3)}$$

where

$$\beta_g = b_{1.2}\lambda_{g.1} + b_{2.1}\lambda_{g.2} \quad \text{(A.4)}$$

is the (regularized) regression coefficient for predictor Xg.

Regardless which linear regression model (CCR-LM, or PLS) is used to generate the predictions, when the number of components K equals the number of predictors P, the results are identical to those obtained from traditional least squares (OLS and WLS) regression. Traditional least squares regression produces unbiased predictions, but such predictions may have large variance and hence higher mean squared error than regularized solutions (K < P). Thus, predictions obtained from the CCR module are typically more reliable than those obtained from a traditional regression model.

Procedures CCR-LM and PLS assume that the dependent variable Y is continuous:

- CCR-LM is invariant to standardization and also allows the components to be correlated (recommended)

- PLS produces different results depending upon whether or not the predictors are standardized to have variance 1. By default, the PLS 'standardize' option is activated.

Logistic Regression (CCR-Logistic) and Linear Discriminant Analysis (CCR-LDA)

Logistic regression is the standard regression approach for analyzing a dichotomous dependent variable. Both Linear and Logistic regression are GLM (Generalized Linear Models) in that a linear combination of the explanatory variables ('linear predictor') is used to predict a function of the dependent variable. In the case of linear regression, the mean of Y is predicted as a linear function of the X variables. For logistic regression, the logit of Y is predicted as a linear function of X.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

$$Logit(Y) = \alpha + b_{1.2}S_1 + b_{2.1}S_2$$

which in reduced form yields:

$$Logit(Y) = \alpha + \sum_{g=1}^{P}(b_{1.2}\lambda_{g.1} + b_{2.1}\lambda_{g.2})X_g$$

Logit(Y), defined as the natural logarithm of the probability of being in category 1 of Y divided by the probability of being in category 2 of Y, can easily be transformed to yield the probability of being in either category. For example, the conditional probability of being in group 1 can be expressed as:

Prob(Y=1|X) = Exp[Logit(Y|X)] / [1 + Exp[Logit(Y|X)]

Thus, the logistic regression model is a model for predicting the probability of being in a particular group. Linear Discriminant Analysis is another model used commonly to obtain predicted probabilities for a dichotomous Y:

- CCR-LDA assumes that the X variables follow a multivariate normal distribution within each Y group, with different group means but common variances and covariances.

- CCR-Logistic makes no distributional assumptions.

Depending upon which method is selected, CCR-LM, CCR-LDA, or CCR-Logistic, in the case where P < N, setting K = P yields the corresponding (saturated) regression models:

Module CCR-LM (or PLS) is equivalent to OLS regression (for K = P)

Module CCR-Logistic yields Logistic regression (for K = P)

Module CCR-LDA yields Linear Discriminant Analysis (for K = P)

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

9 Tutorials

9.1 Tutorial 1: 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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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).

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

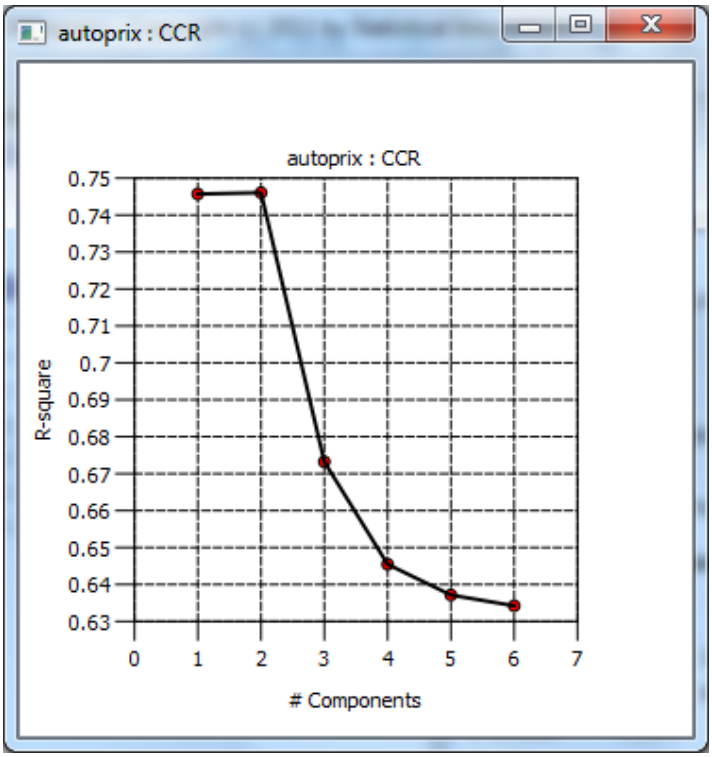


Fig. 1. Cross-Validation Component (CV-R^2) Plot

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

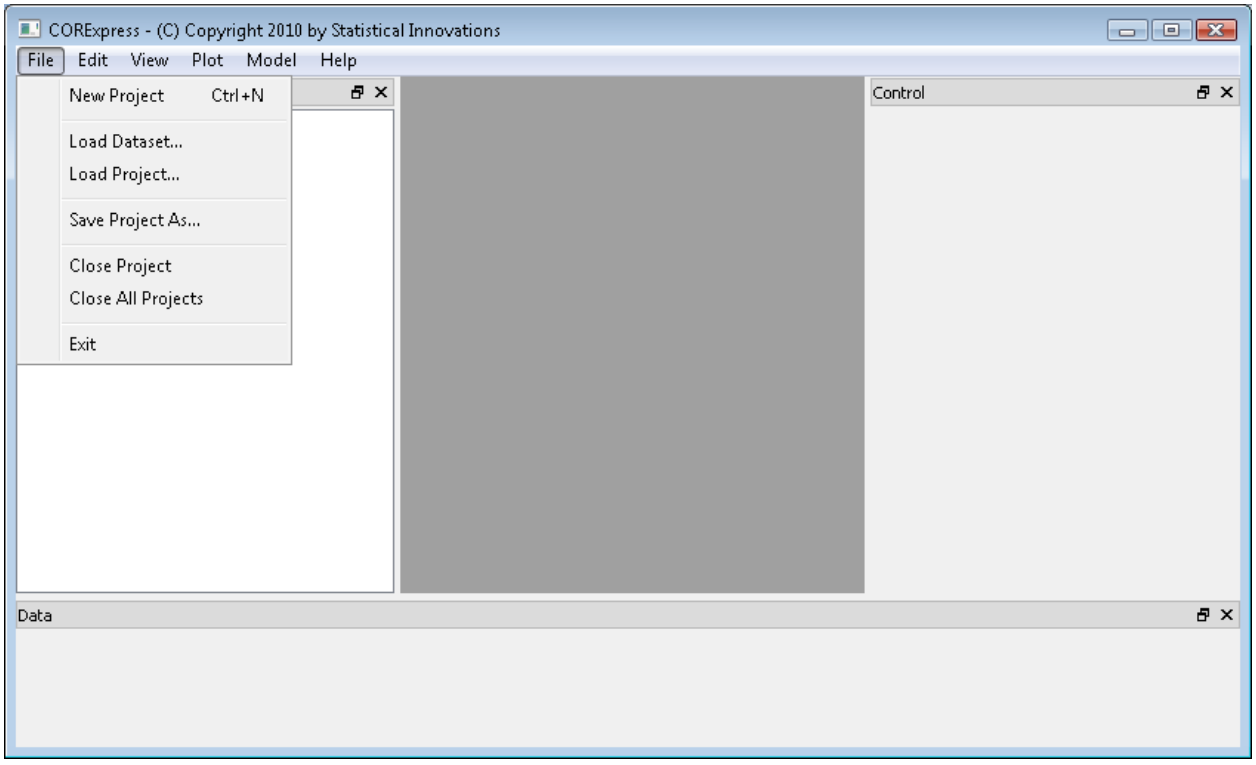


Fig. 2: File Menu

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

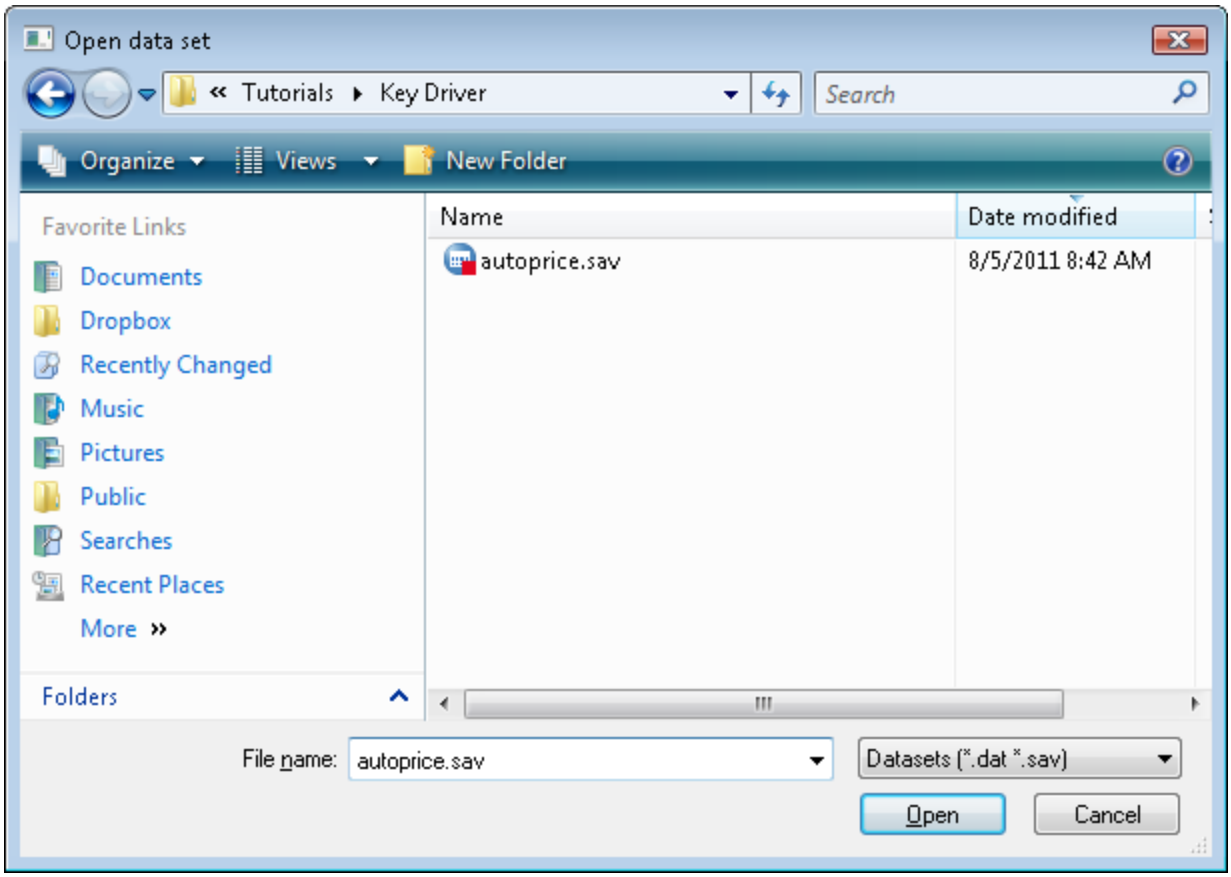


Fig. 3: Loading a Dataset

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

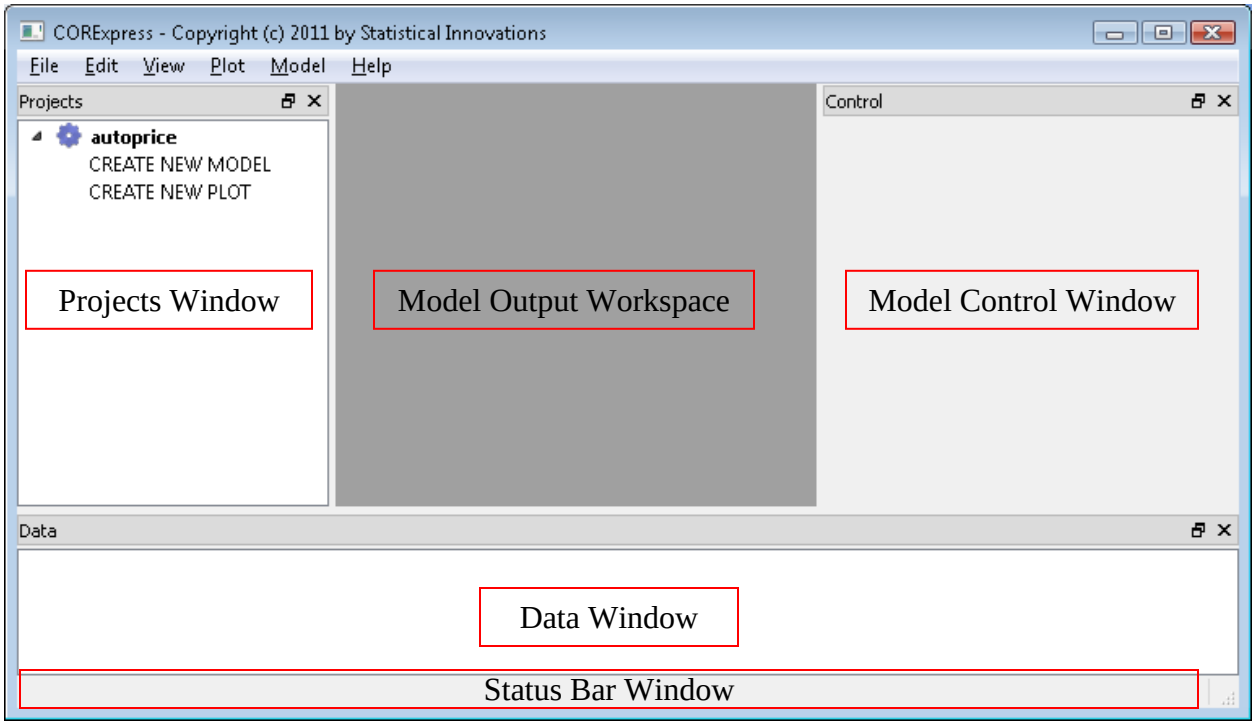
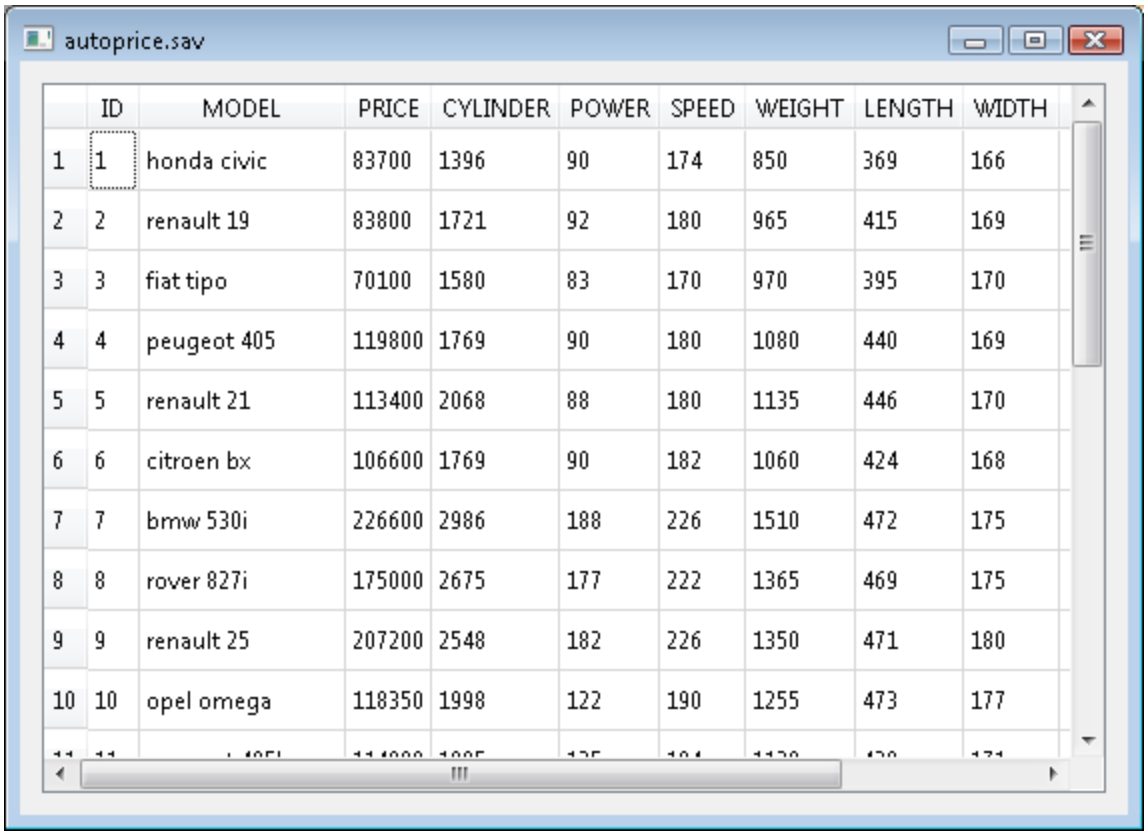


Fig. 4: COREXpress Windows

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).

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)



	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: COREXpress 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:

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- 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.

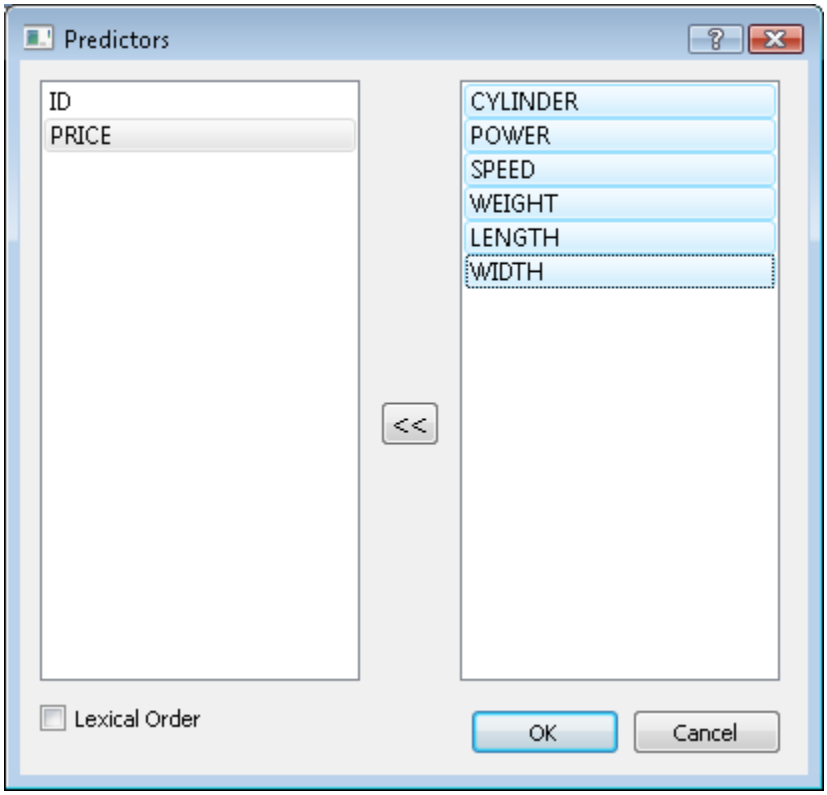


Fig. 6: Predictor Window

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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:

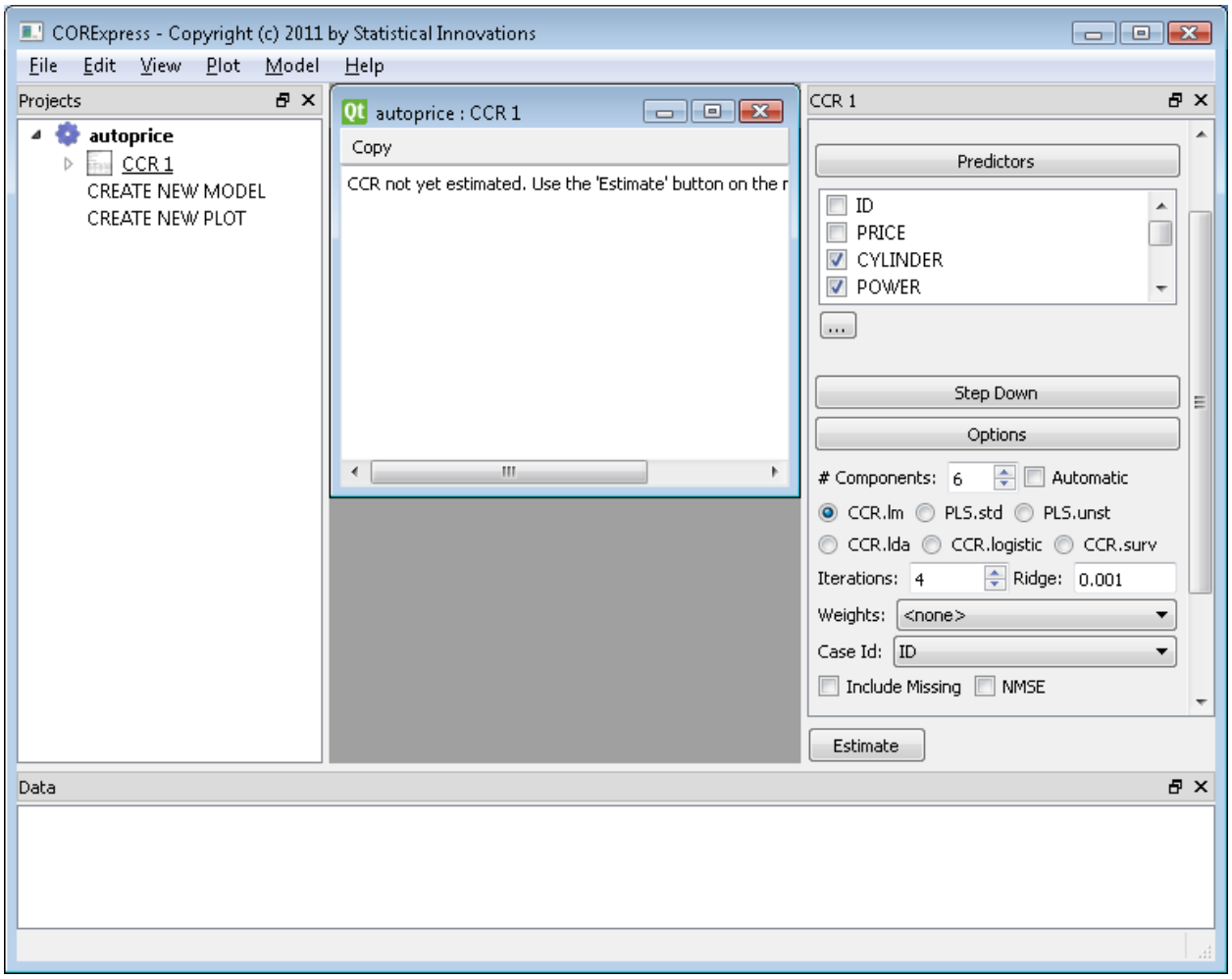


Fig. 7: Control Window

Estimate the Specified Model:

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

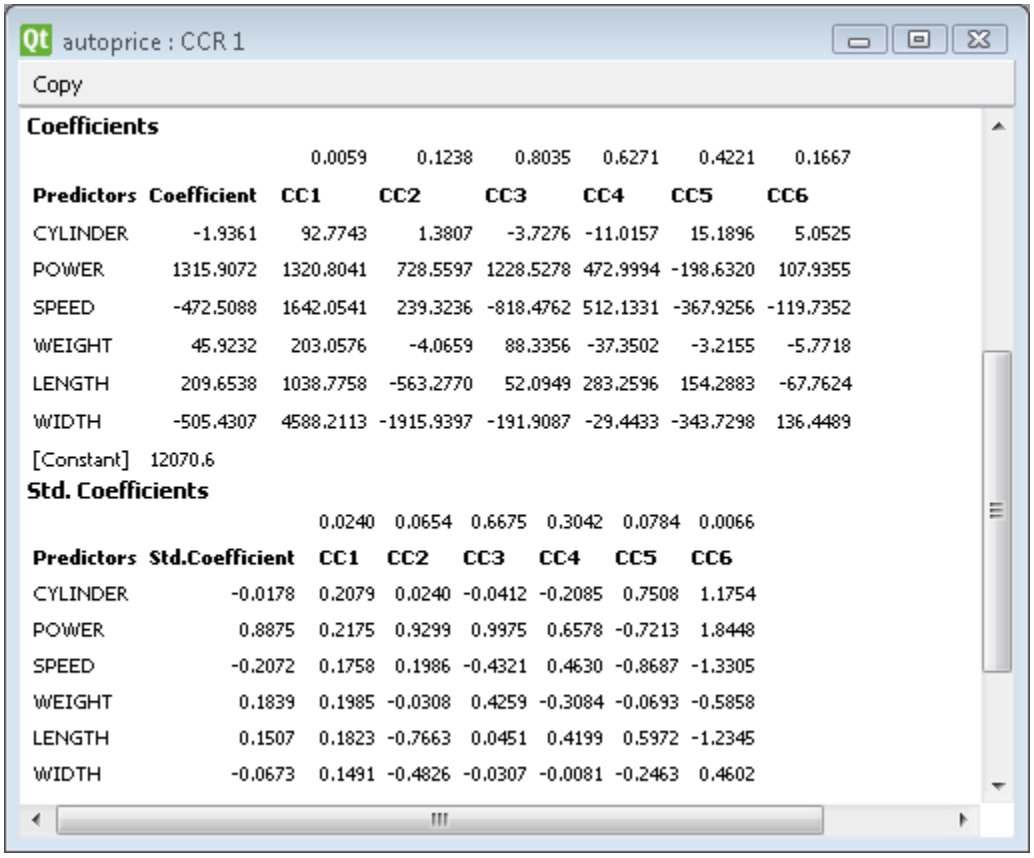


Fig. 8. Coefficient estimates obtained from the 6-component (saturated) CCR model

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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 ($K \leq 6$). 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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

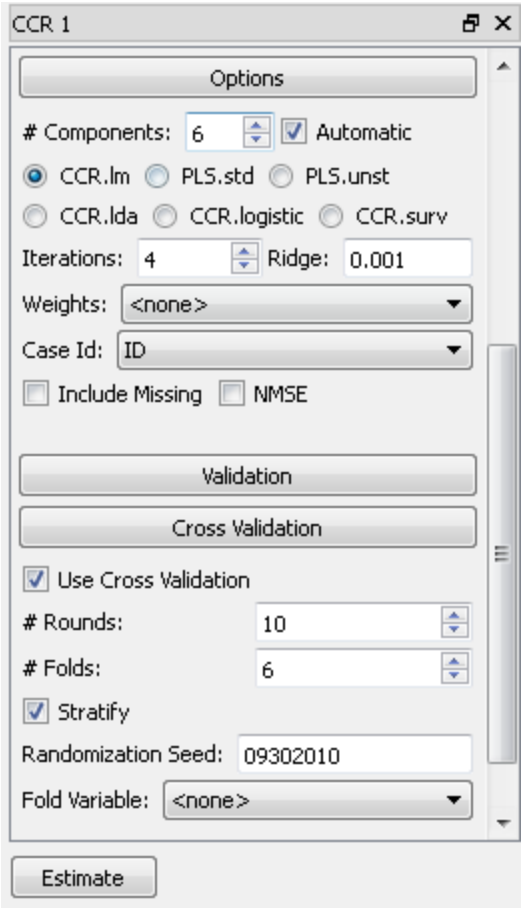


Fig. 9. Model Control Window with Automatic activated and Cross-validation specified

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

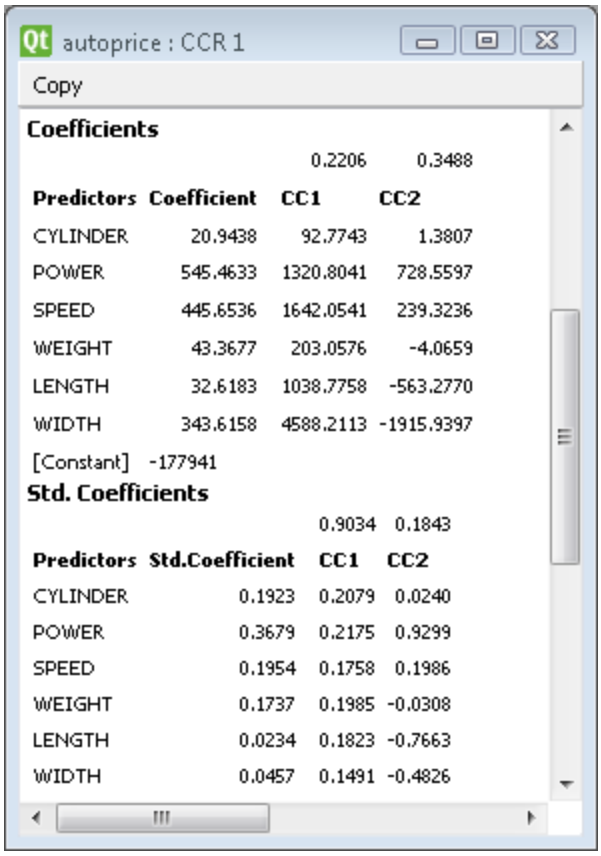


Fig. 10. Model Output

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

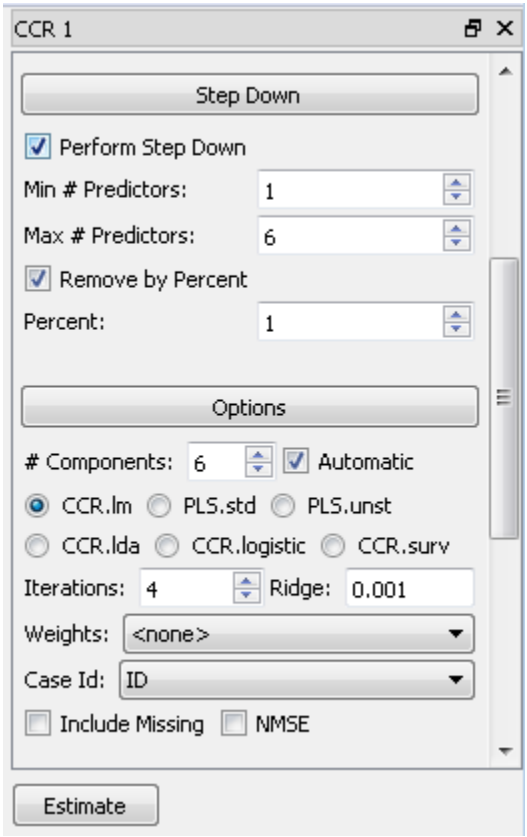


Fig. 11. Model Control Window with Step-down options specified

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

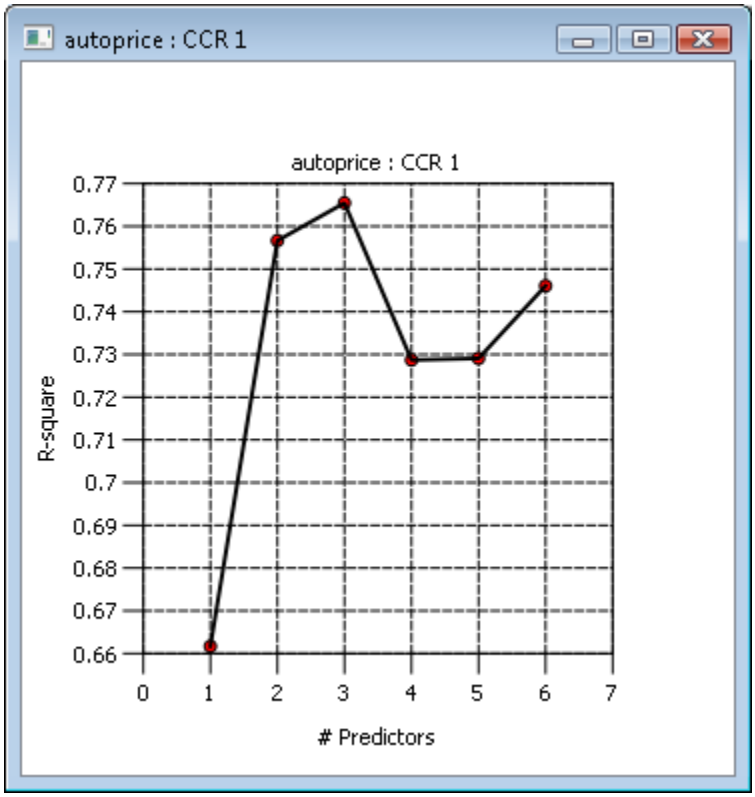


Fig. 12.

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

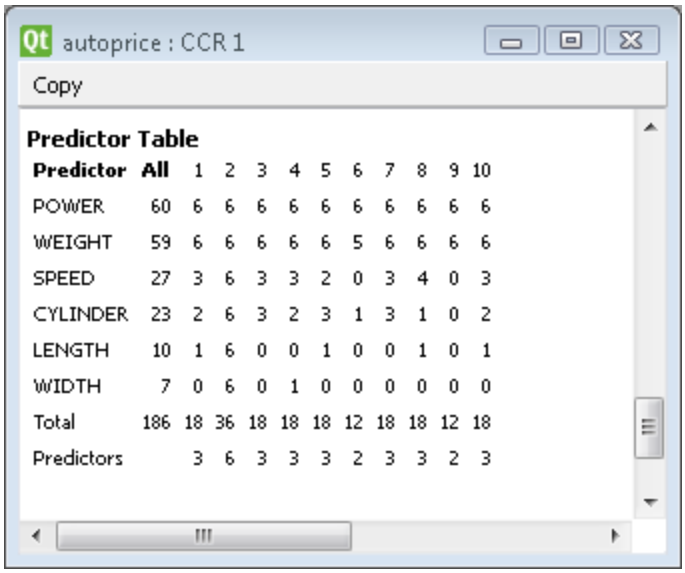


Fig. 13.

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

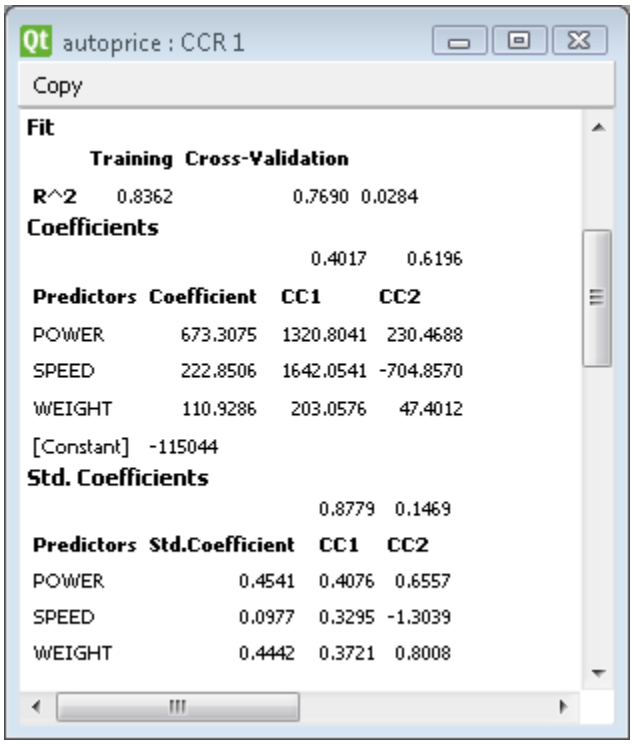


Fig. 14

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

General Discussion and Additional Tutorials

Key driver regression attempts to ascertain the importance of several key explanatory variables (predictors) $X_1, X_2, \ldots, X_P$ 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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

9.2 Tutorial 2: CCR 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 1A, which illustrates the use of COREXpress with P=56 predictors with a large sample of size N=5,000, and Part 1B, 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 1A 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 1B 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 1A 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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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 1B, in order to provide comparisons that are comparable to those in Part 1A 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 1B 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 1B.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Tutorial 2A: 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.

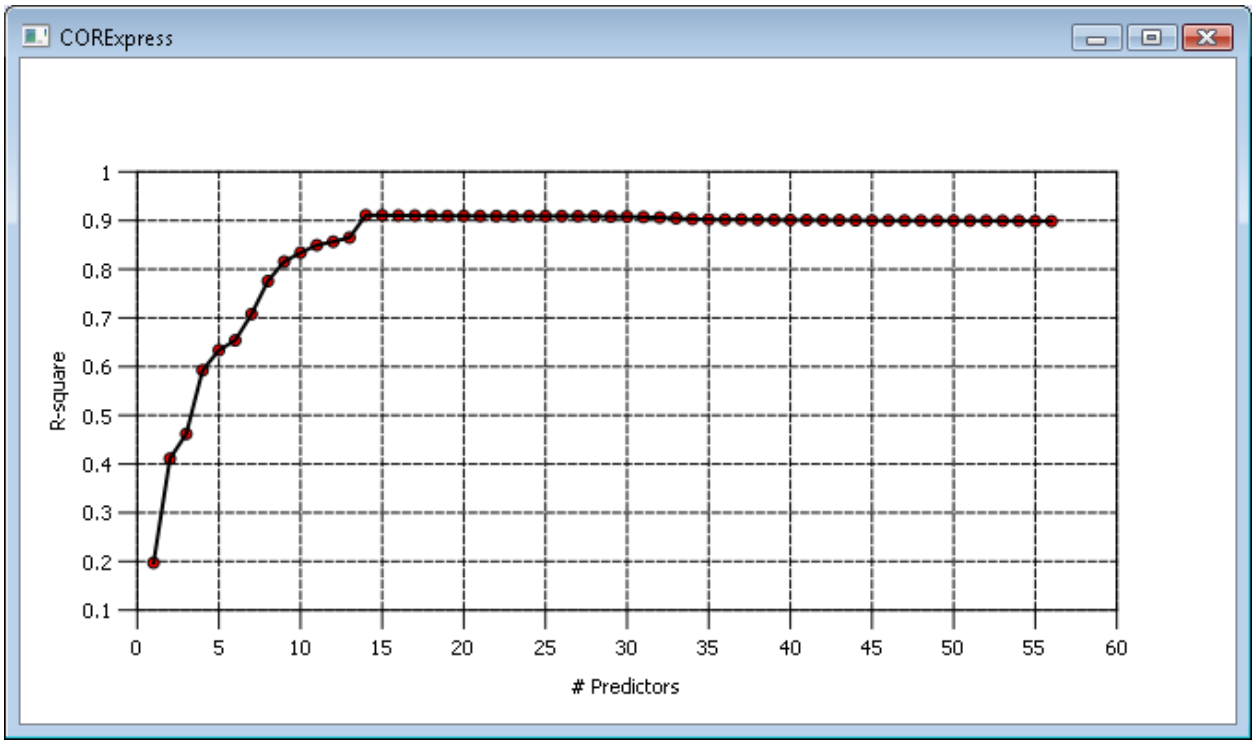


Fig. 1. CV-R^2 Plot of CV-R^2 for K=8. The maximum value for CV-R^2 occurs with P=14 predictors included in the model.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

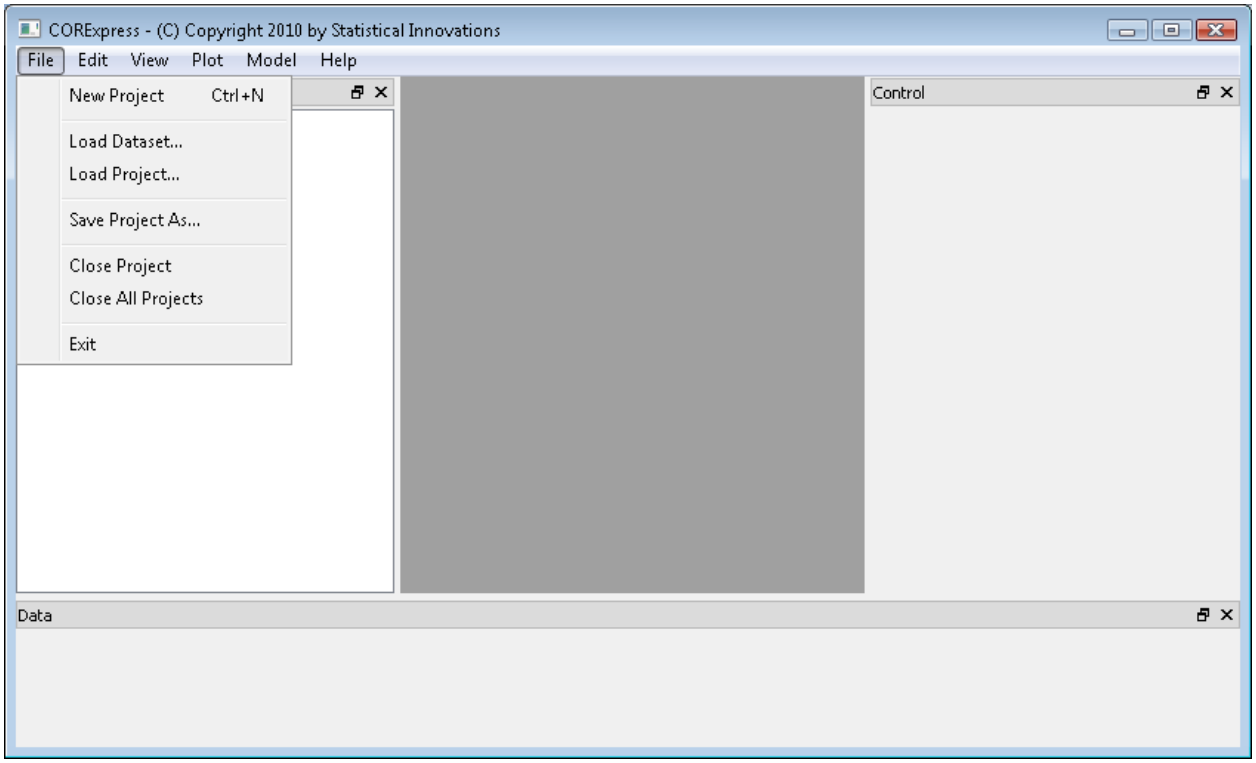


Fig. 2: File Menu

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

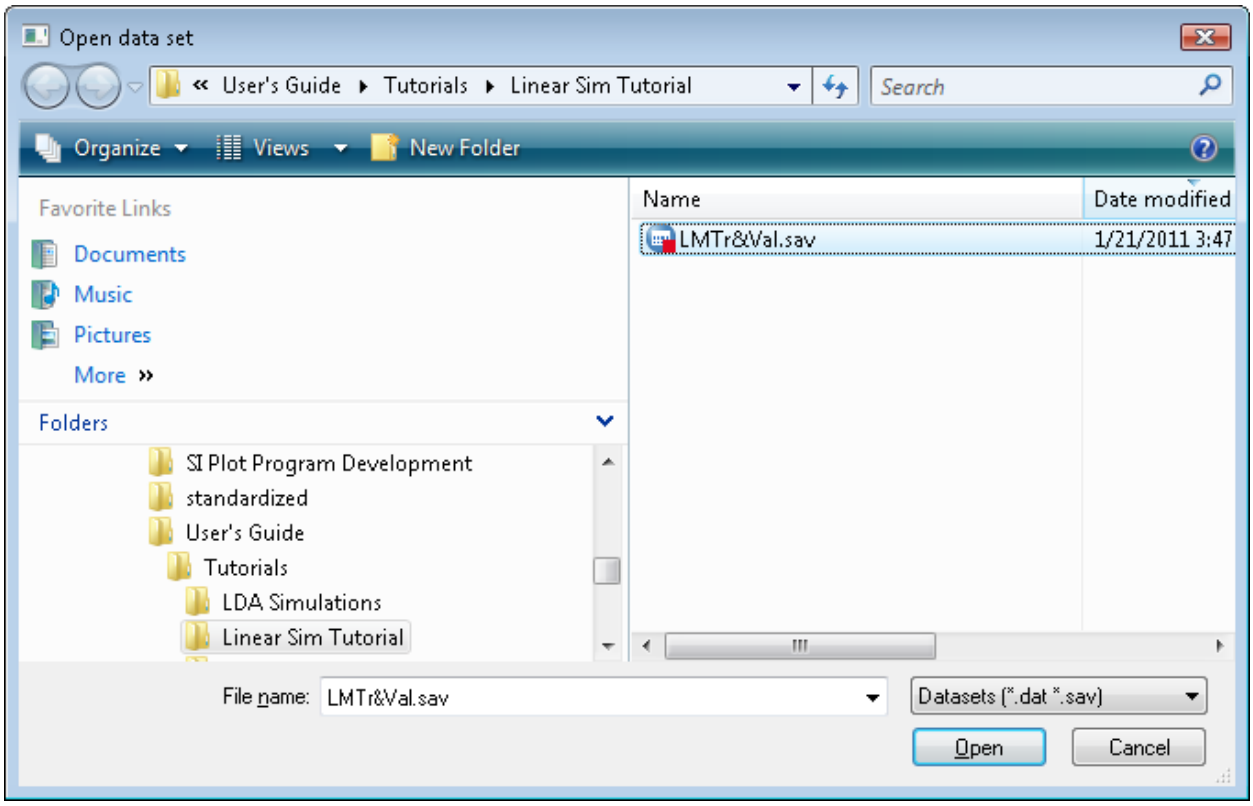


Fig. 3: Loading a Dataset

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

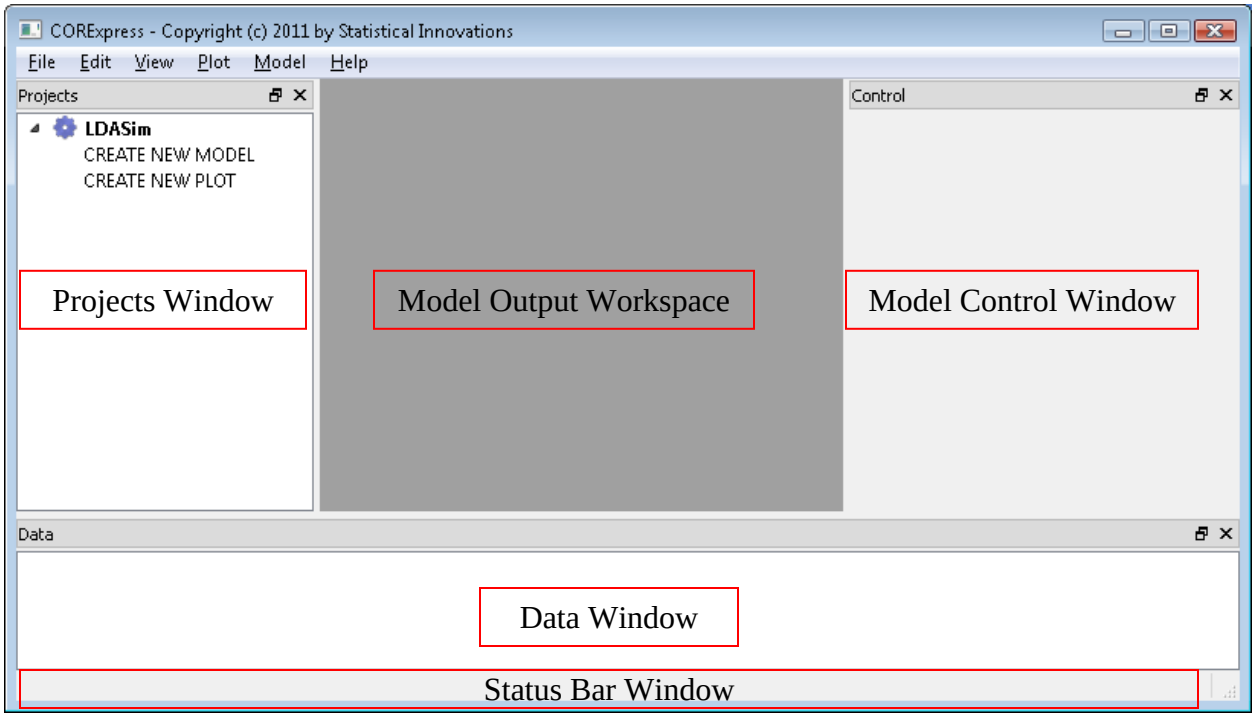
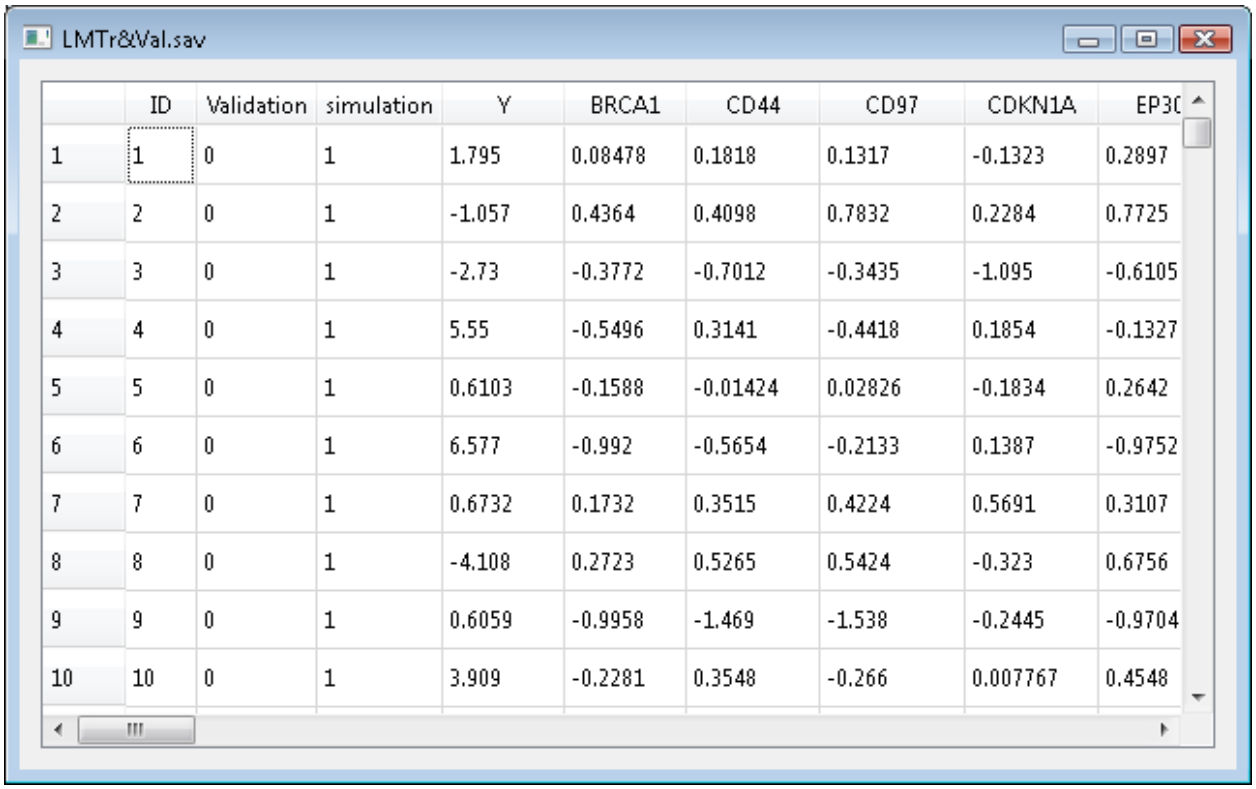


Fig. 4: COREXpress Windows

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)



	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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- 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.

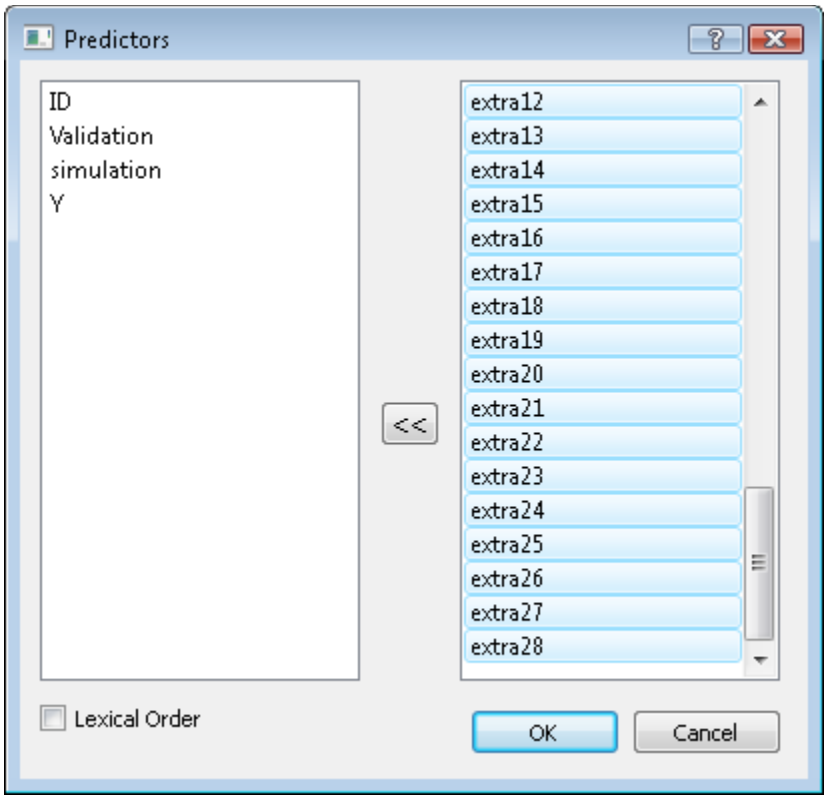


Fig. 6: Predictor Window

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Your Control window should now look like this:

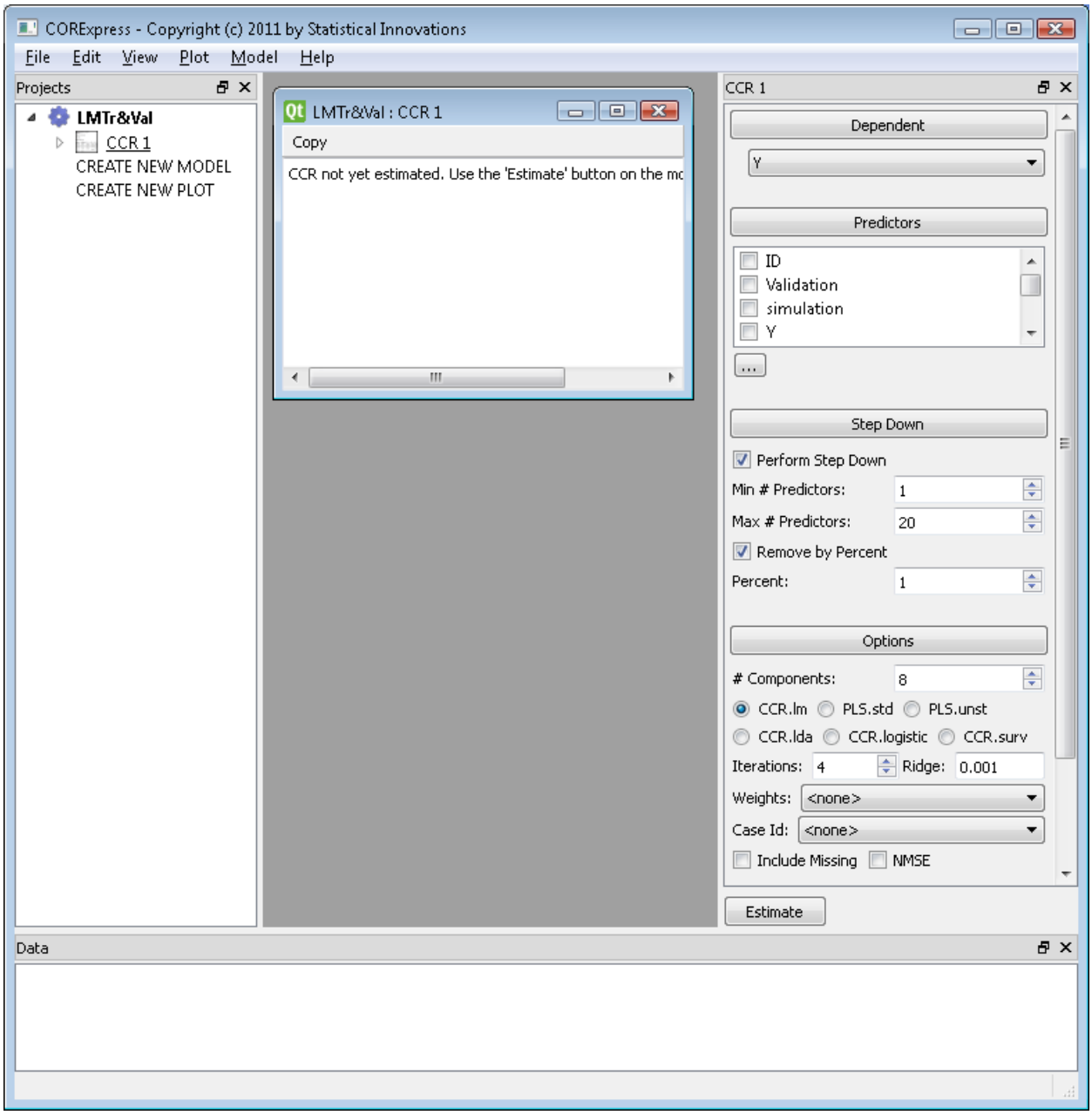


Fig. 7: Control Window

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Your Control window should now look like this:

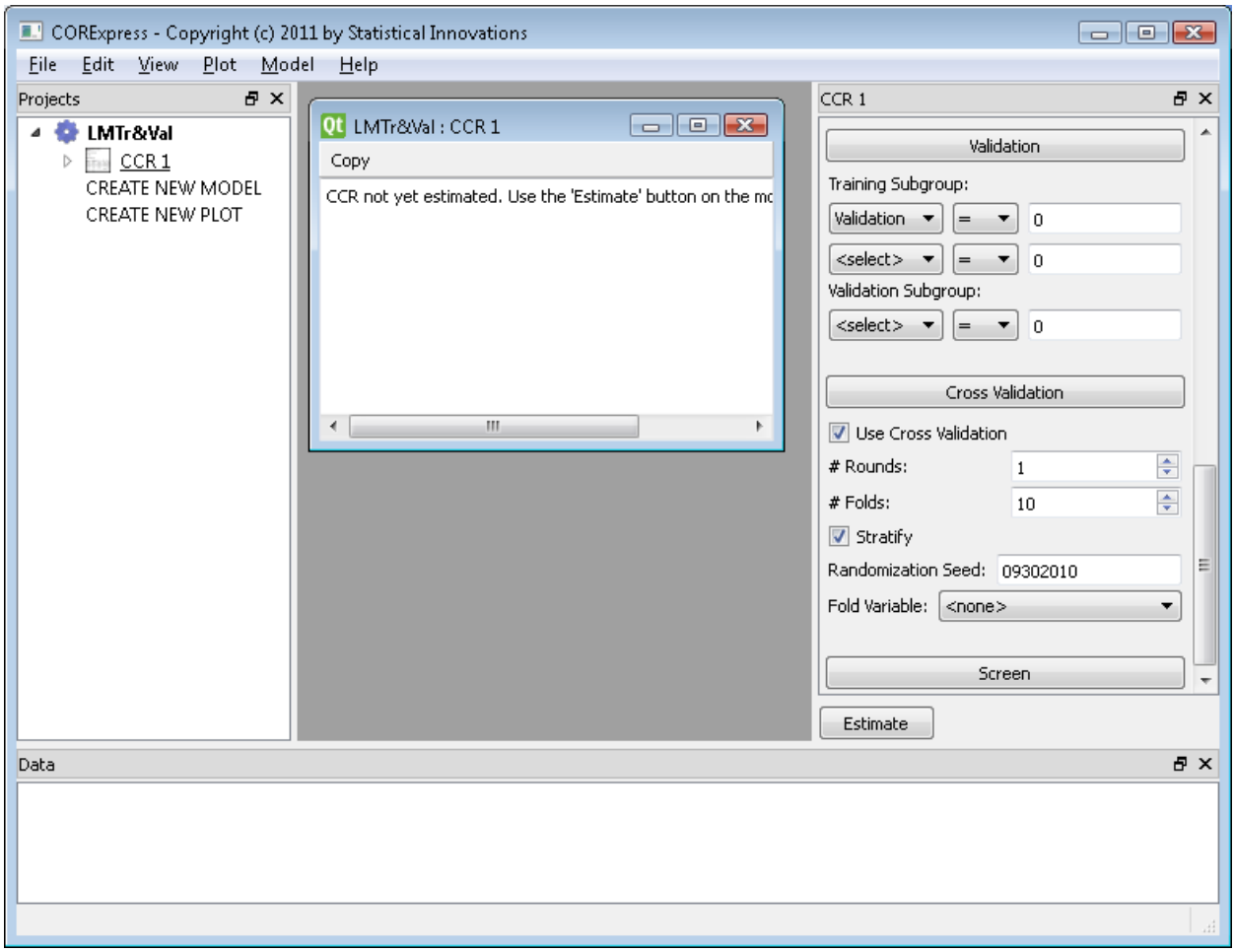


Fig. 8: Control Window

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.

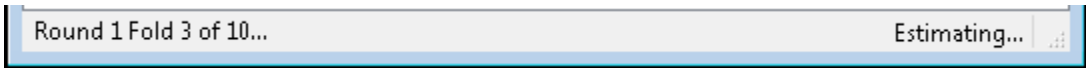


Fig. 9: COREXpress Status Bar

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

View Model Output

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Viewing CV-R^2 Plot:

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

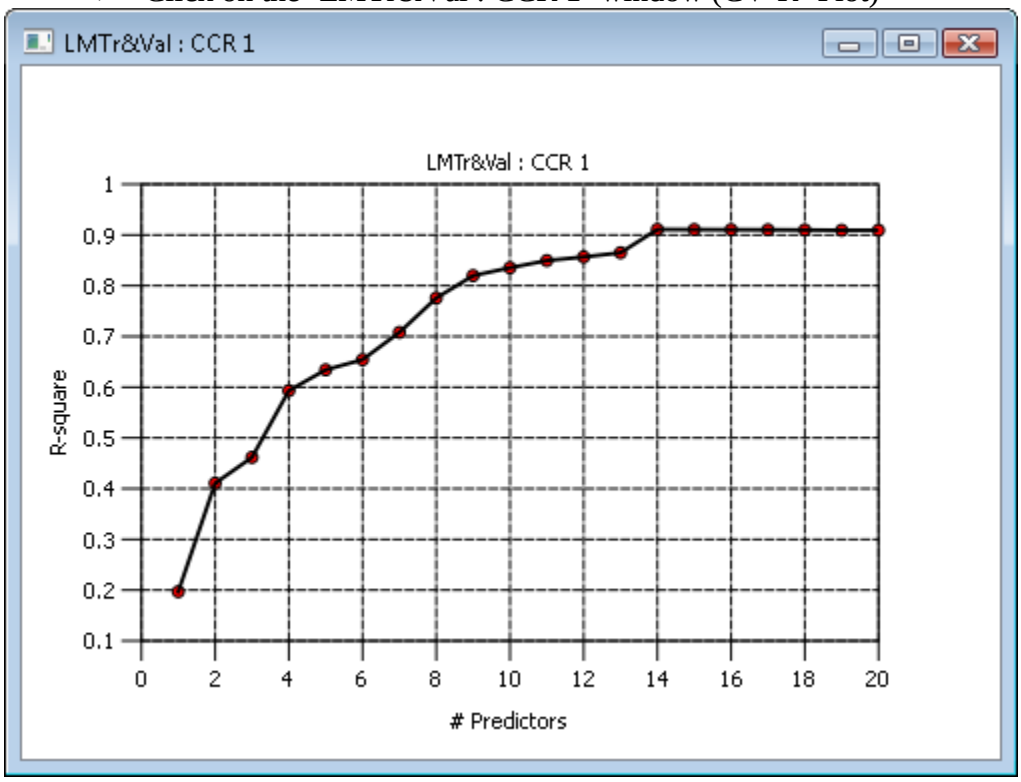


Fig. 10: CV-R^2 Plot

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

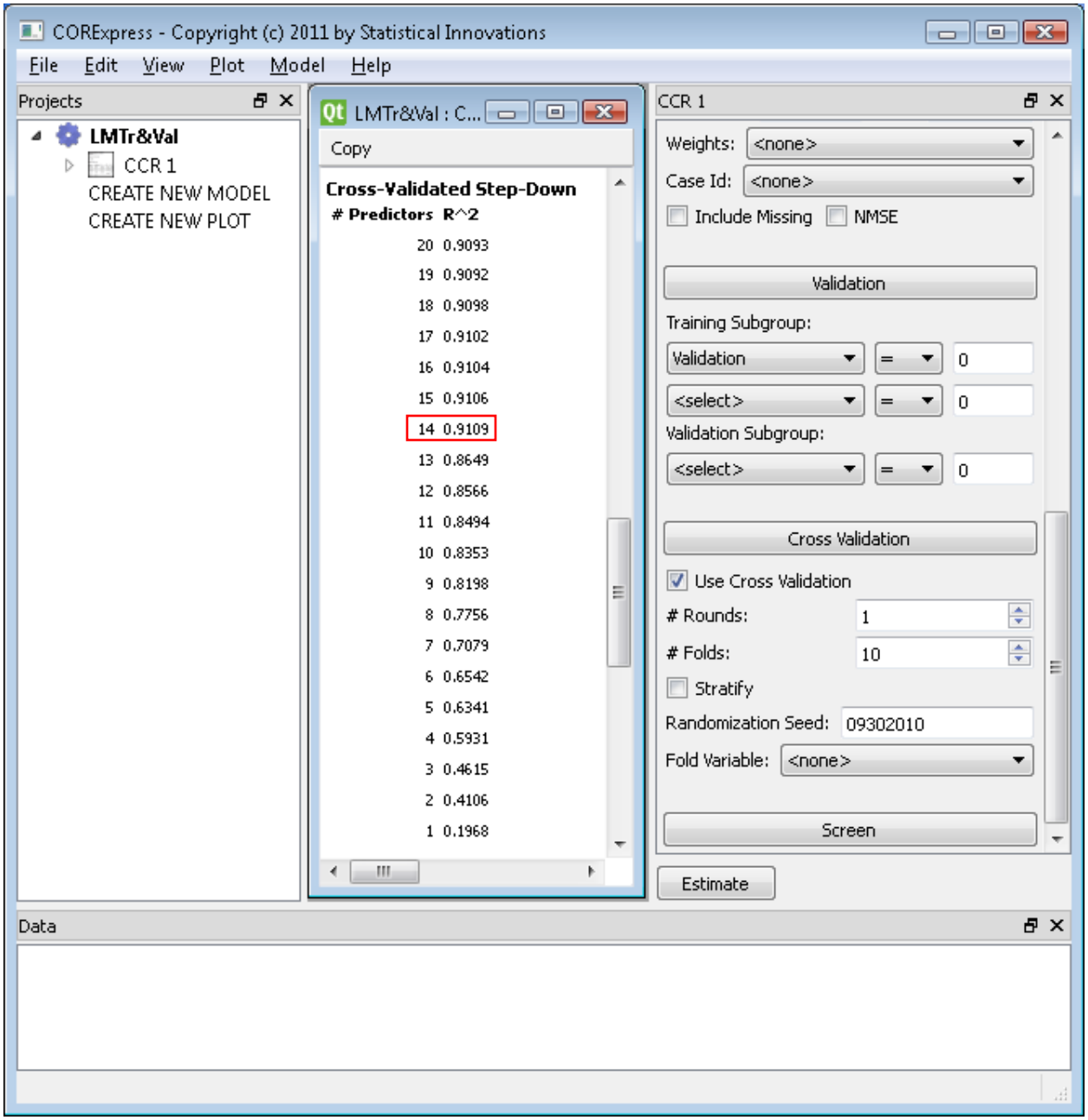


Fig. 11: Cross Validation R^2 Output in the Model Output Window

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Viewing the 'Optimal' Model Output:

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

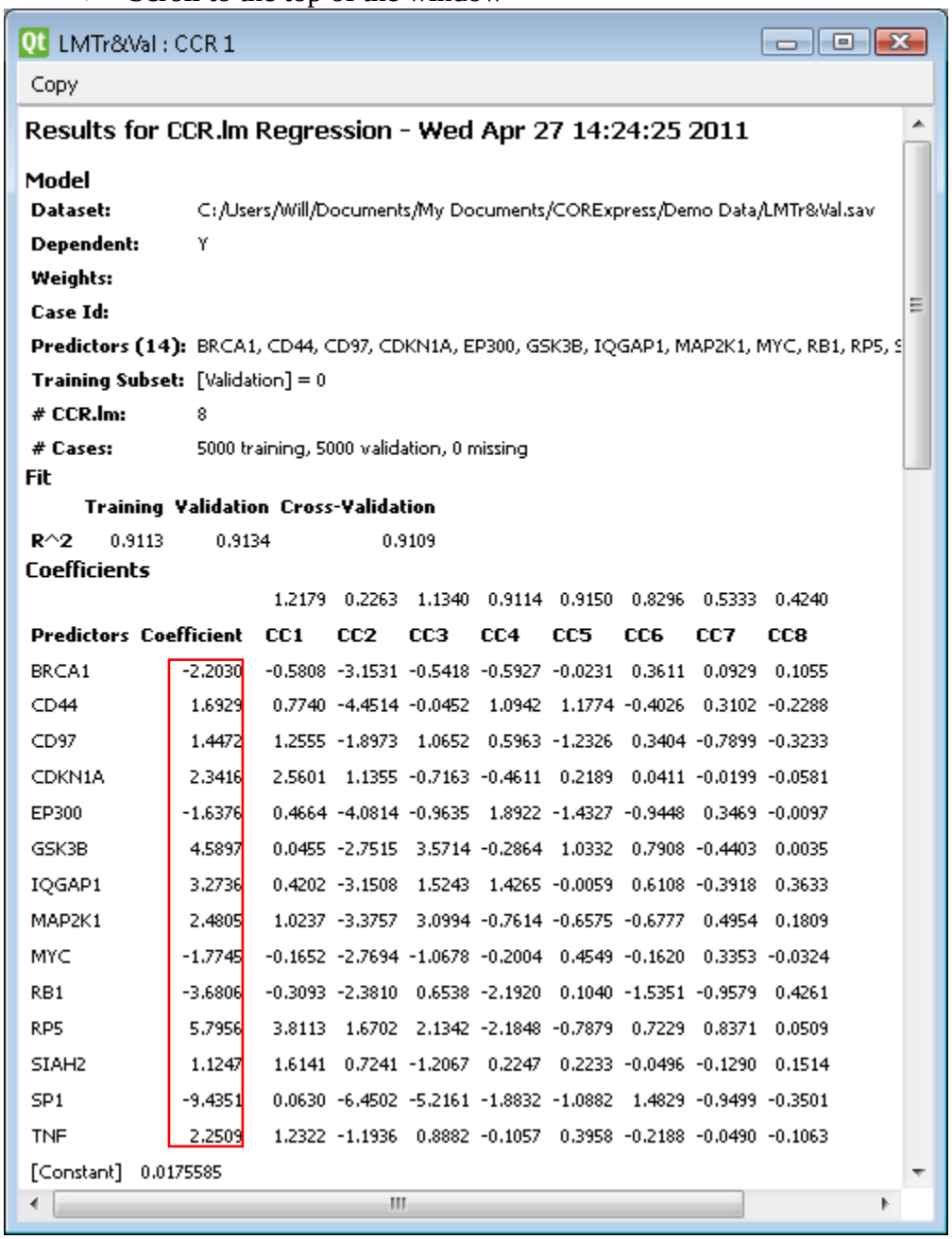


Fig. 12: Unstandardized Coefficients for K=8 in the Model Output Window
Note that for the Training, the R^2=0.9113 and for the Validation the R^2=0.9134.

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.)

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

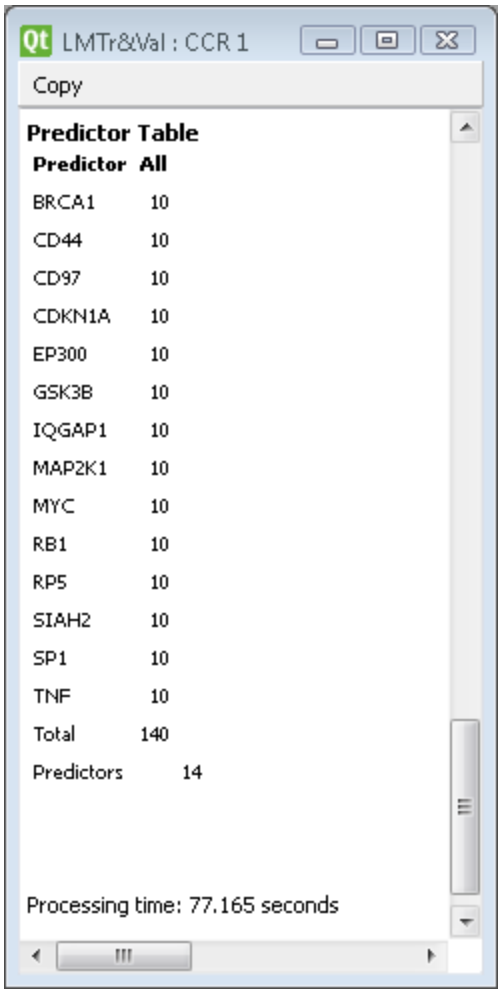


Fig. 13. Predictor Counts Table

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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'

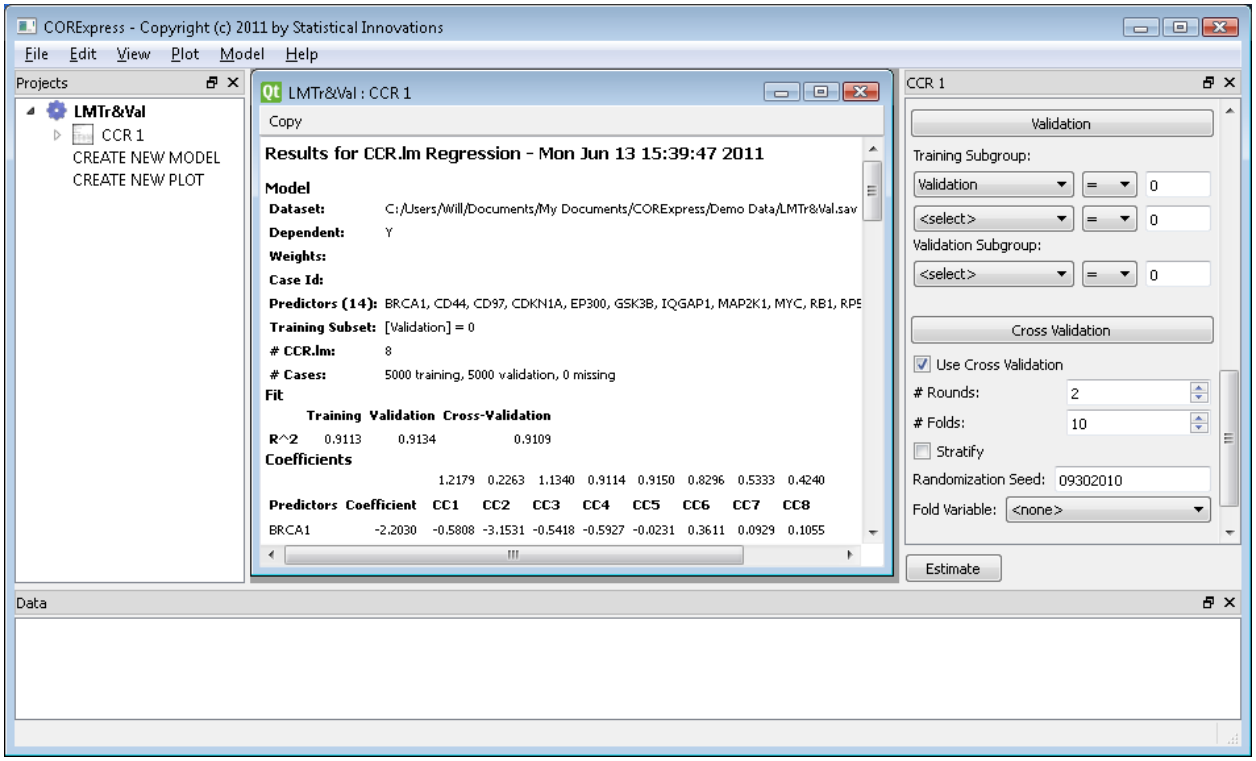


Fig. 14: Control Window

- Click 'Estimate'

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

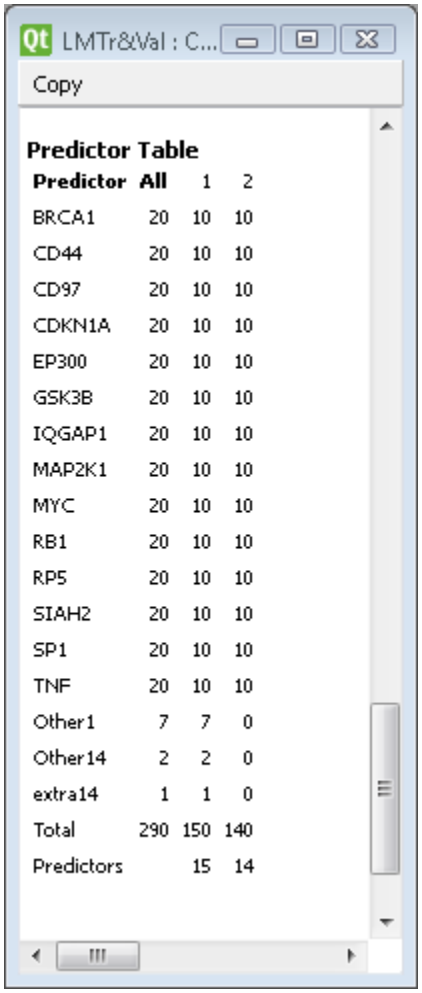


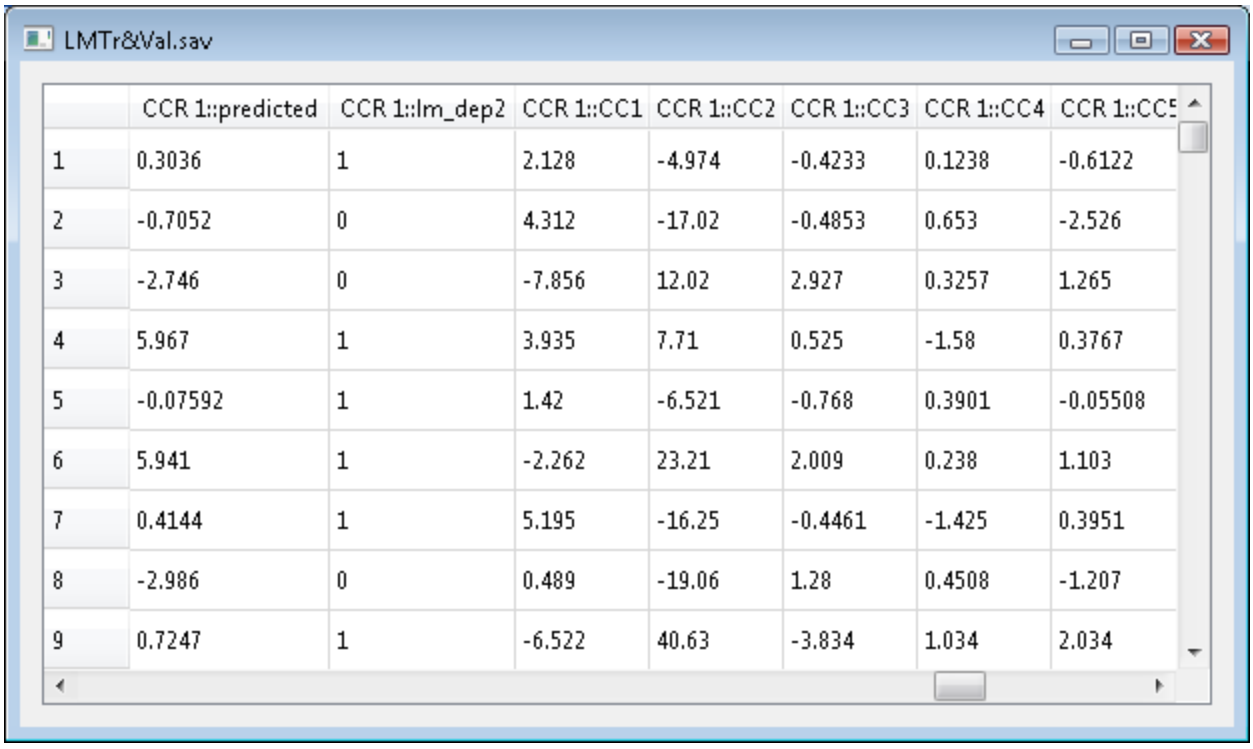
Fig. 15: Predictor Counts Table for K=10-component model with 2 rounds of 10-folds

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)



	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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

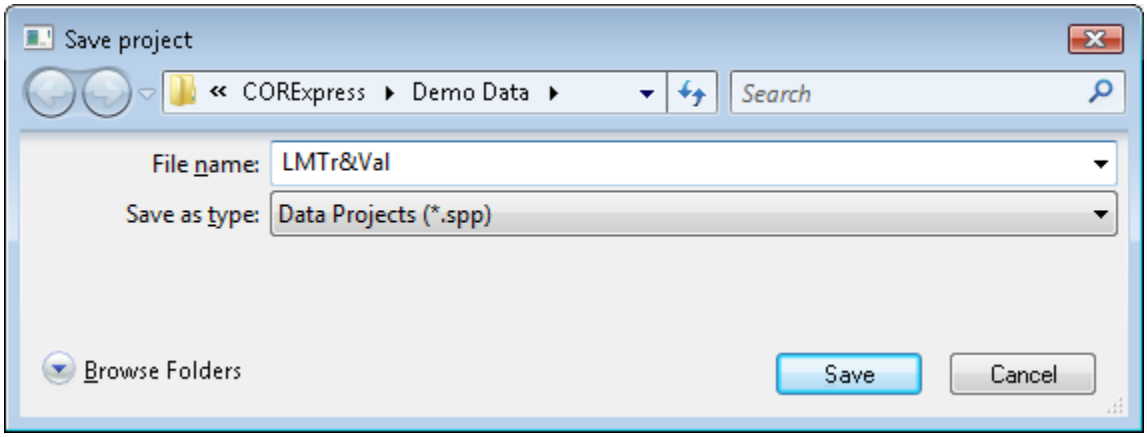


Fig. 17: Saving a Current Project

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Tutorial #2B: 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

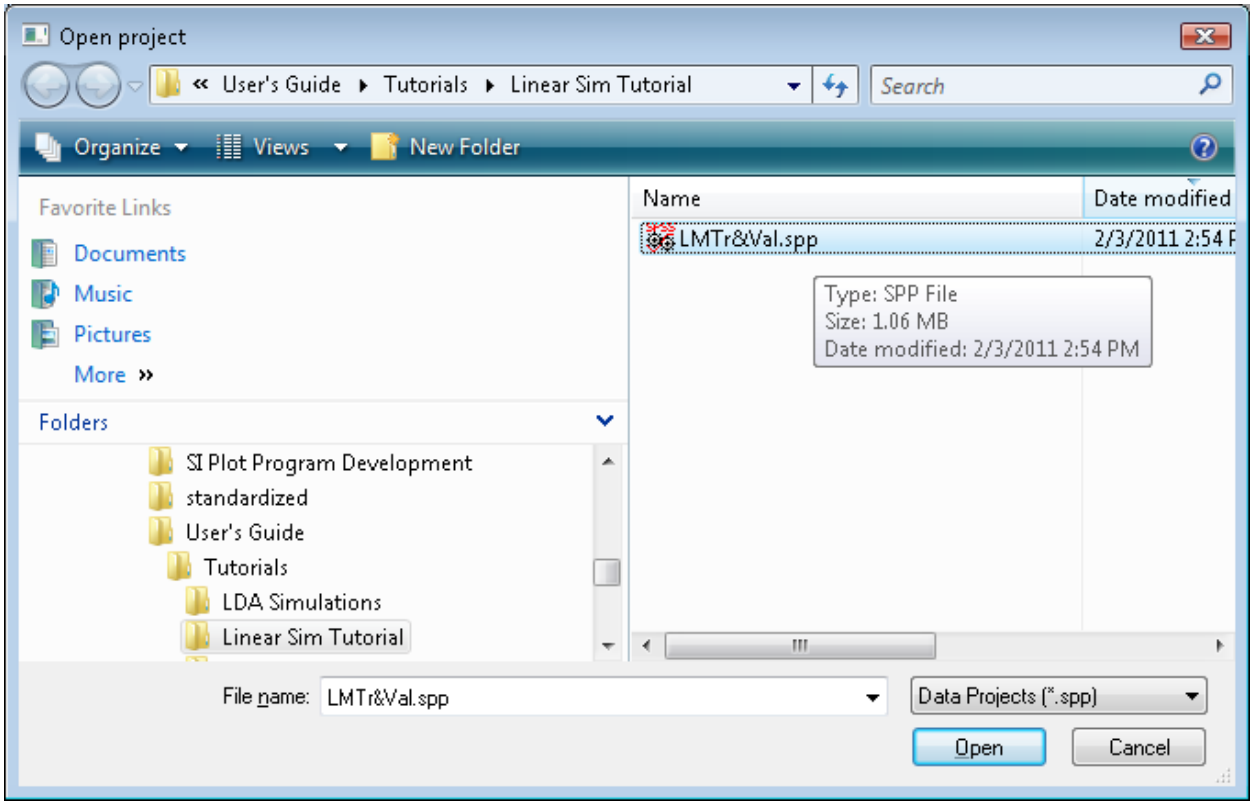


Fig. 18: Loading a previously saved project

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- 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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Your program should now look like this:

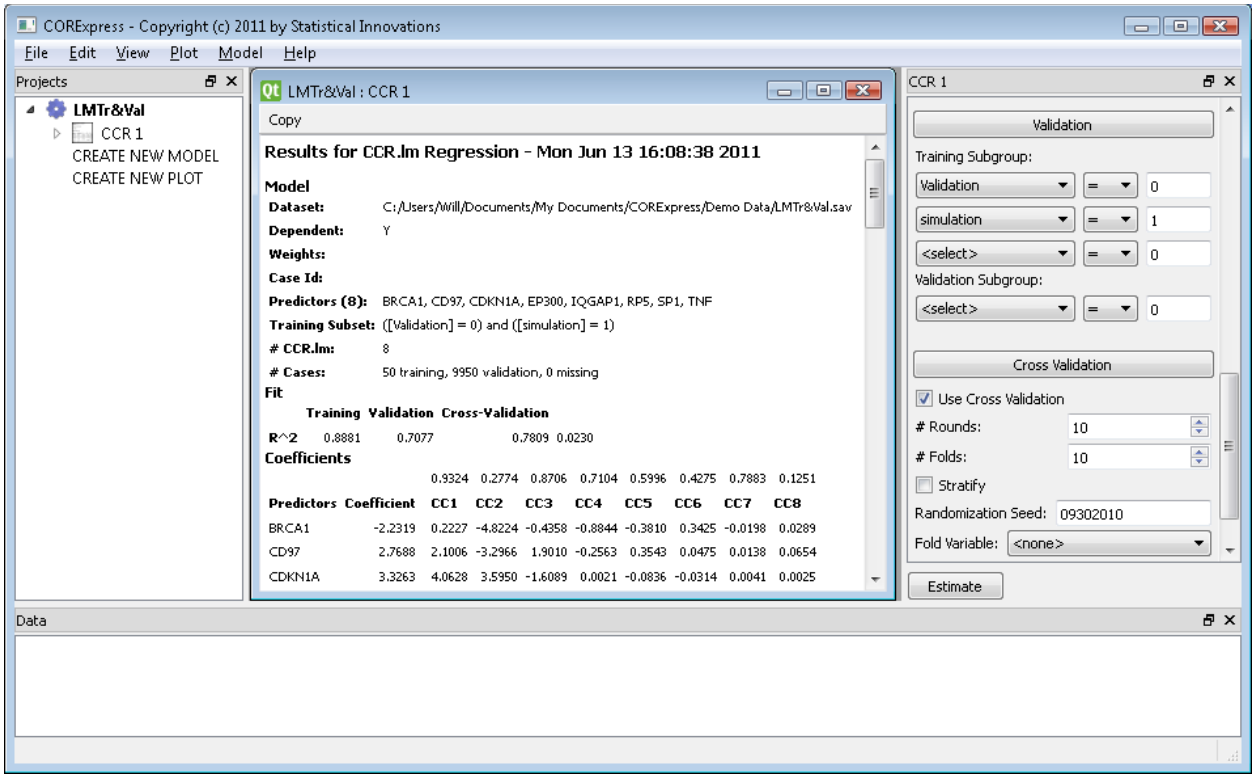
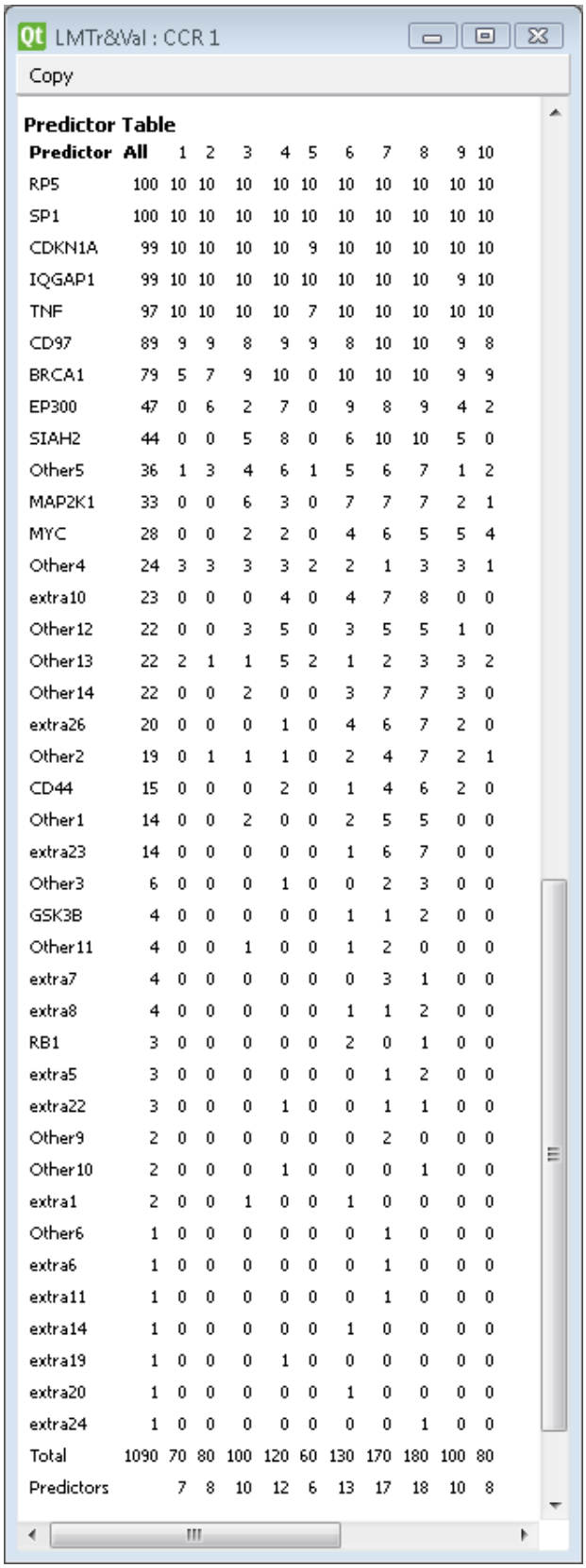


Fig. 19: Control Window

- 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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)



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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

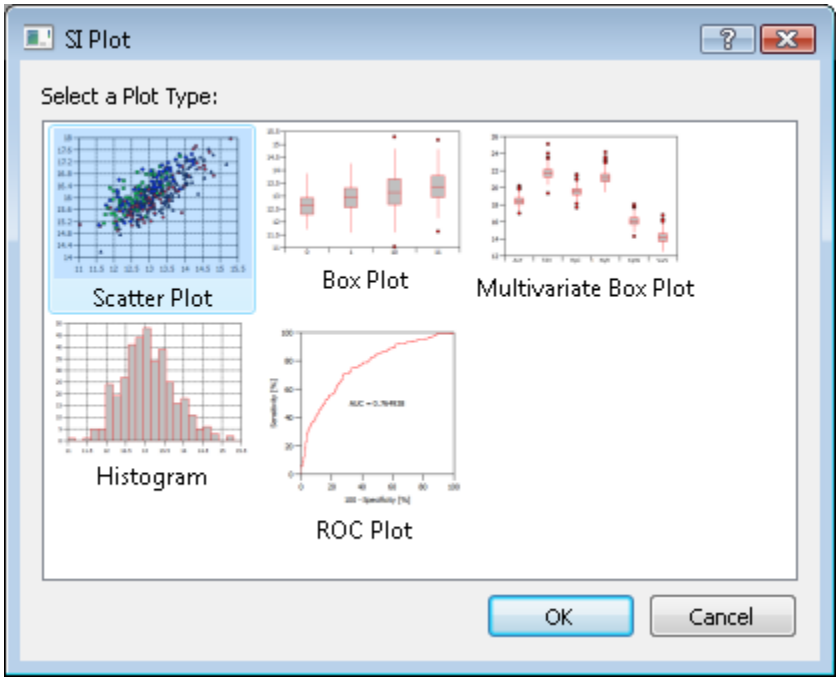


Fig. 21. Create New Plot Options

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'

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

The following window appears:

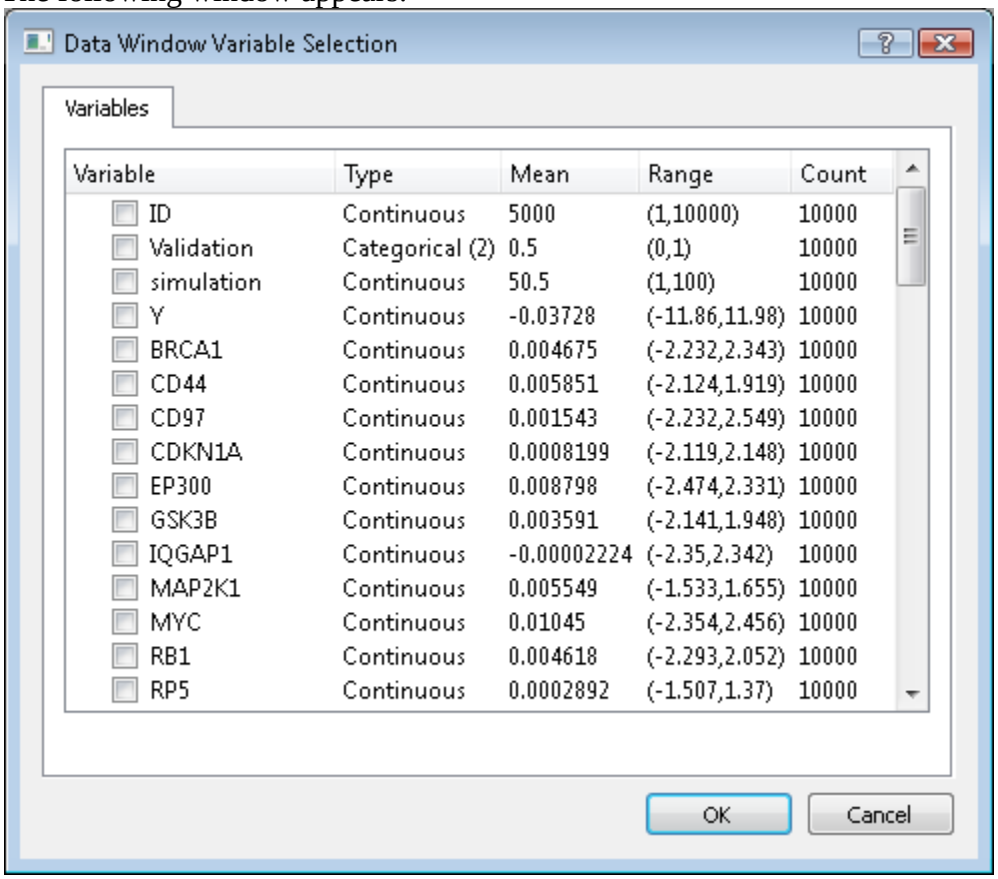


Fig. 22. Project Settings

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

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

9.3 Tutorial 3: CCR 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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

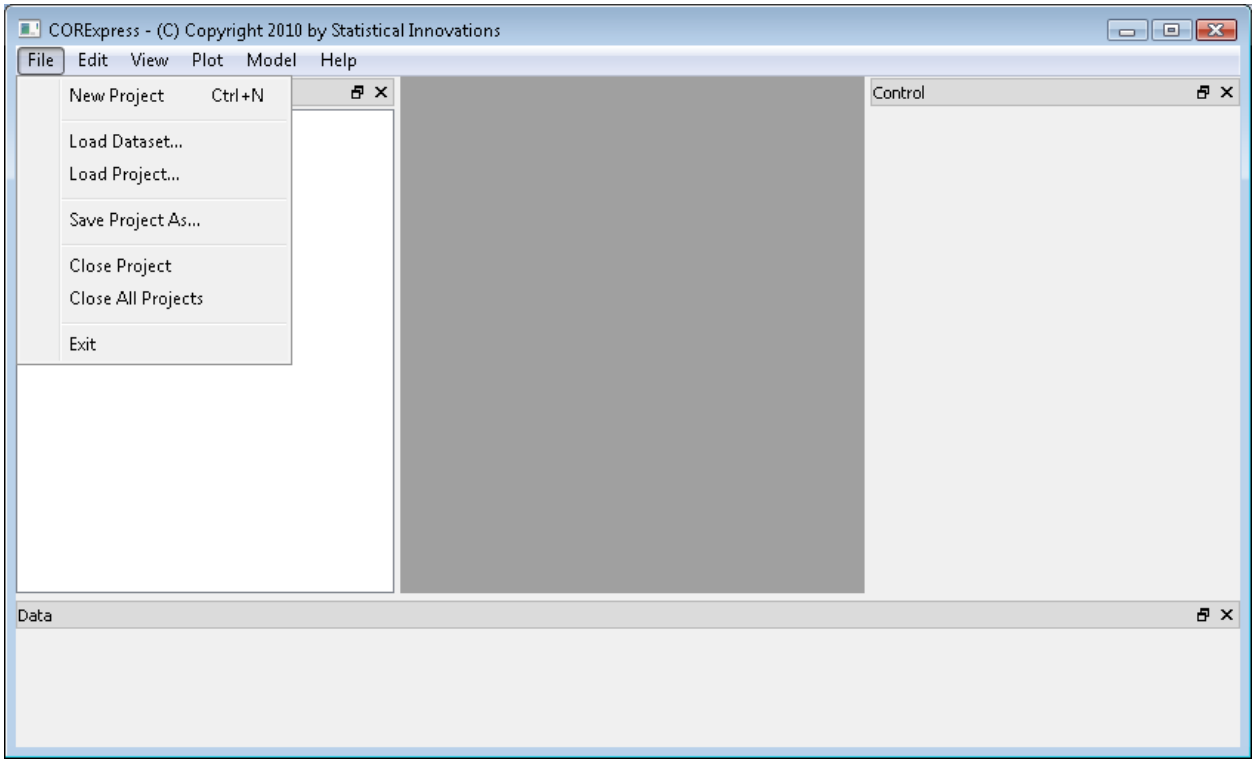


Fig. 1. File Menu

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

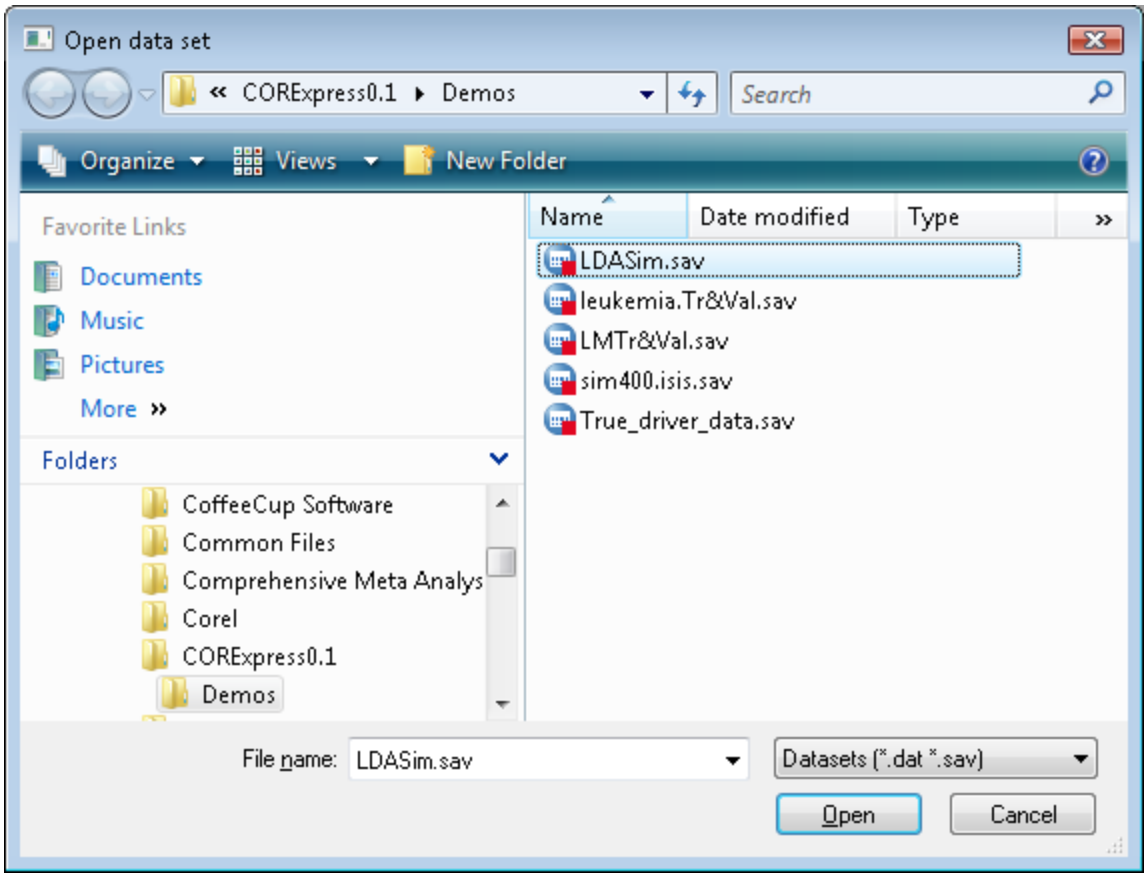


Fig 2. Loading a Dataset

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

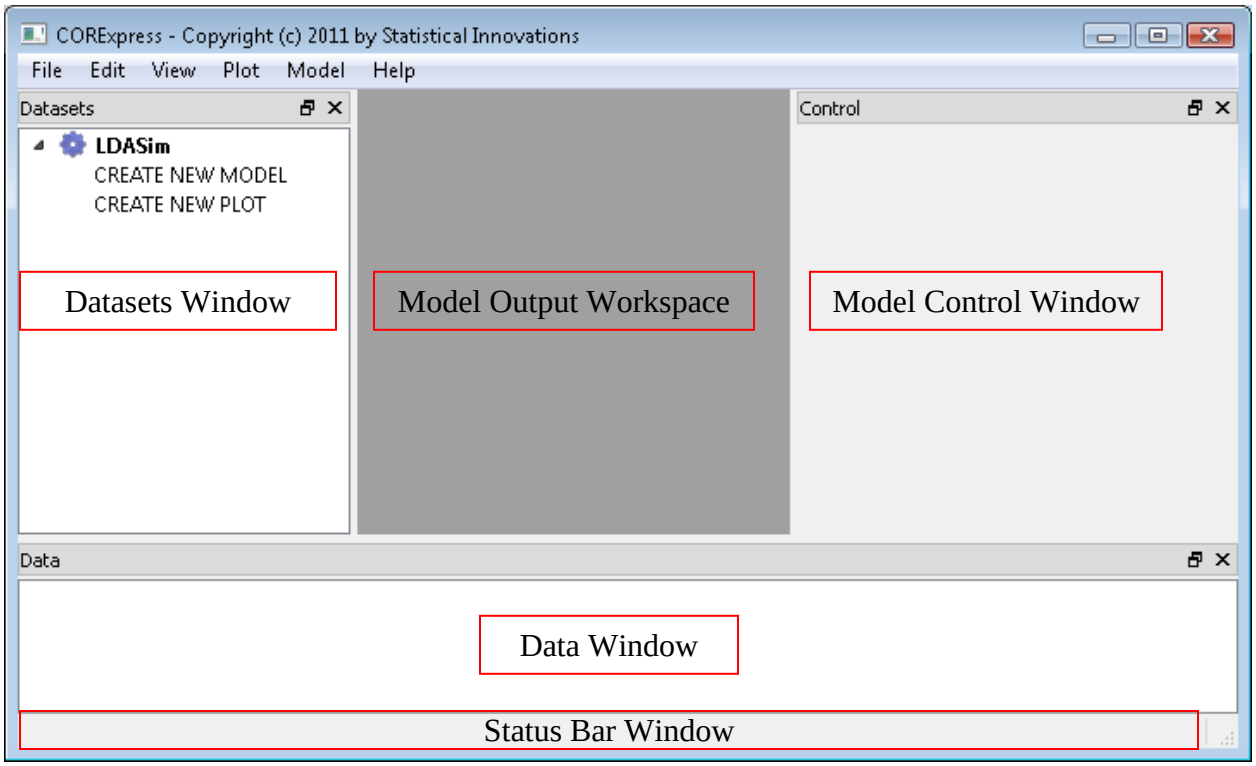
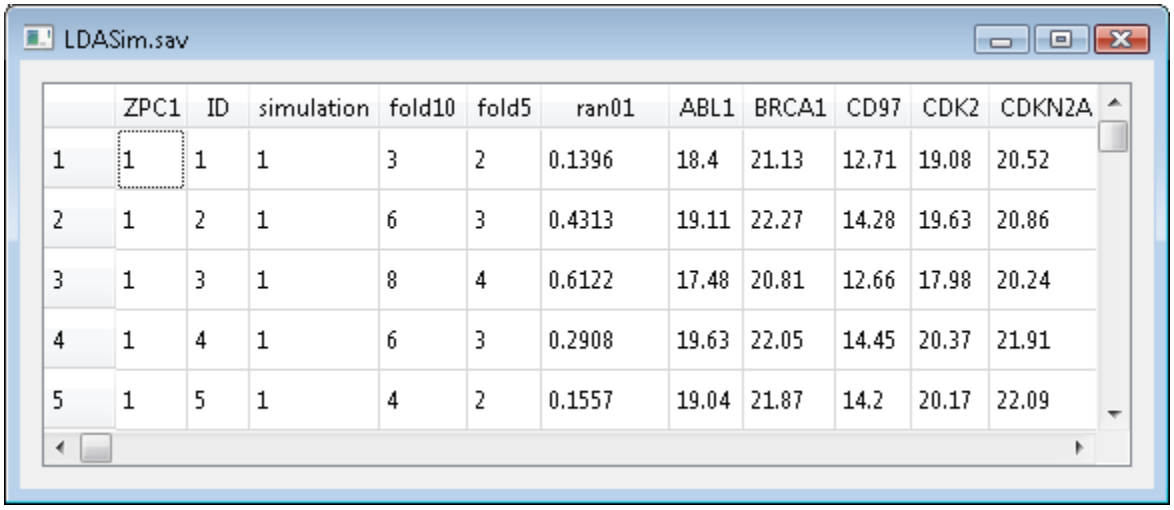


Fig. 3: COExpress Windows

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

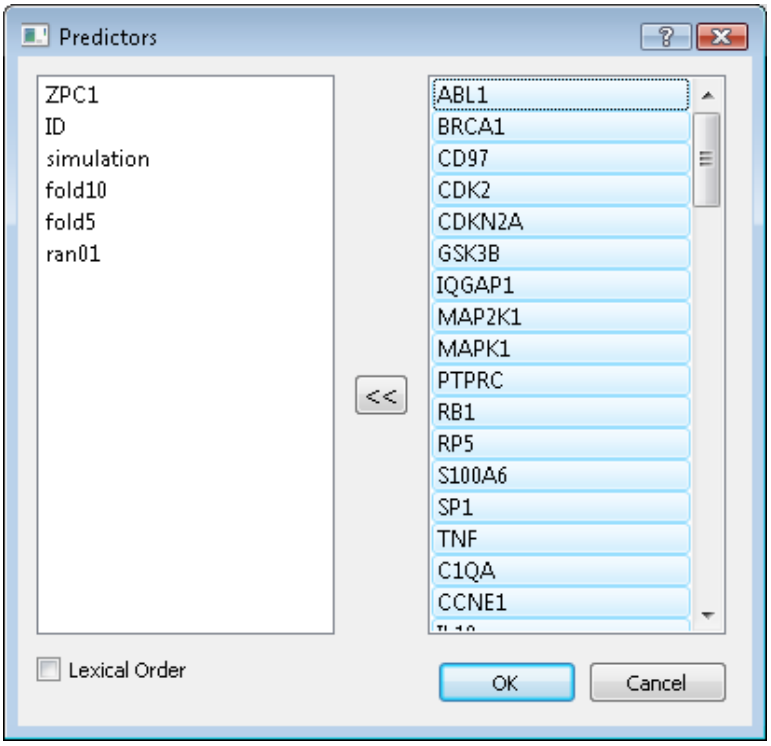


Fig. 5. Predictor Window

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Your Control window should now look like this:

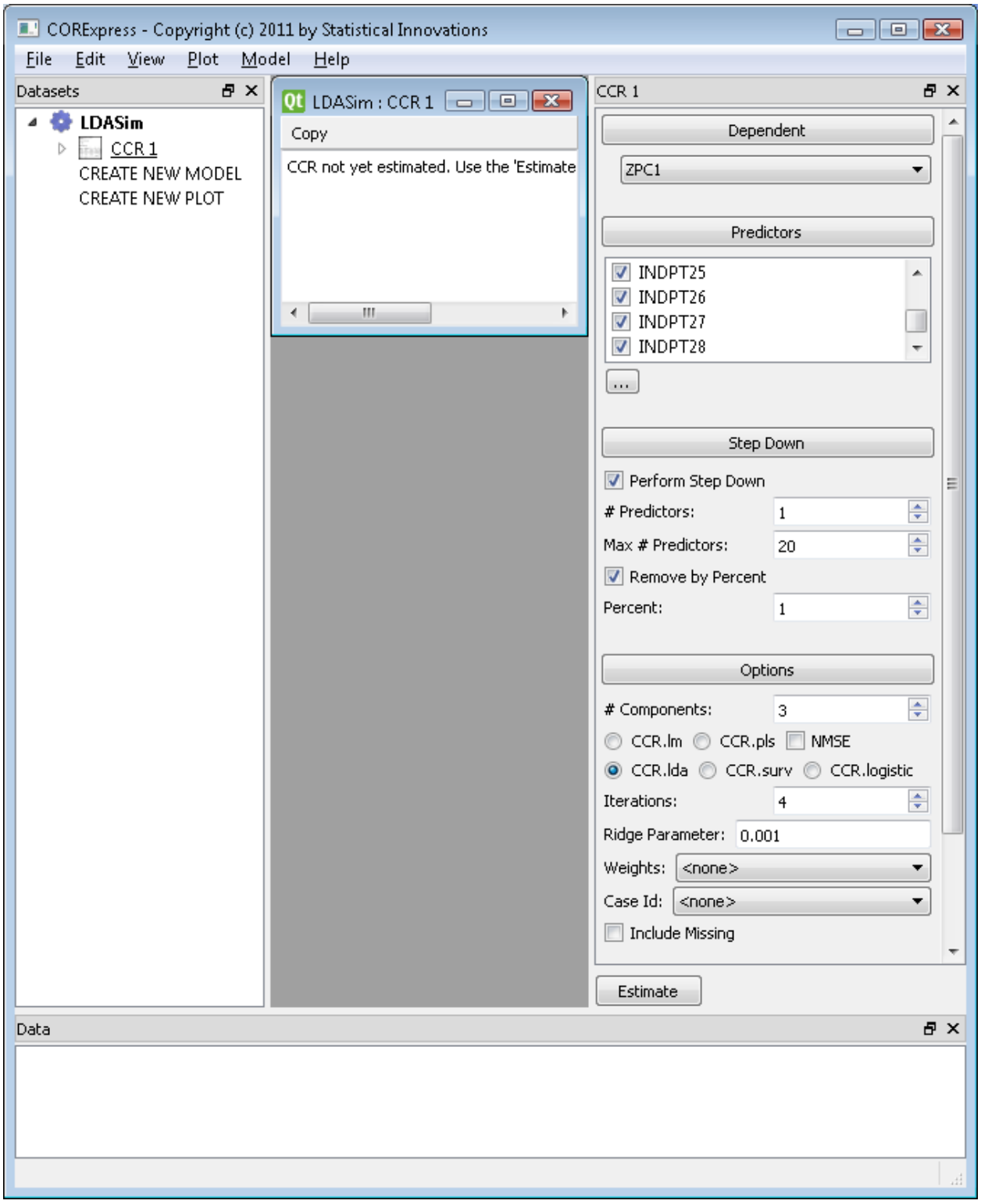


Fig. 6: Control Window

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Your Control window should now look like this:

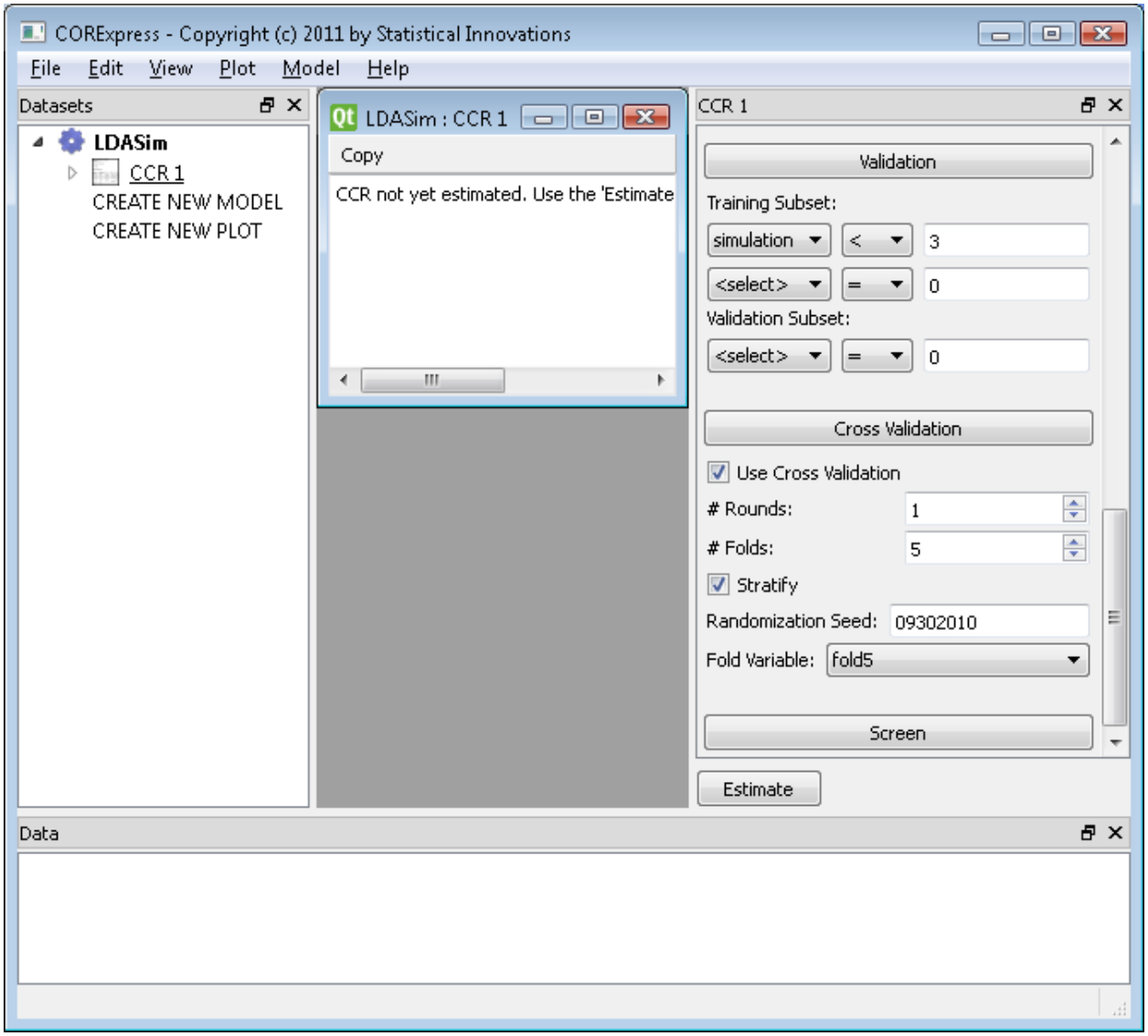


Fig. 7: Control Window

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

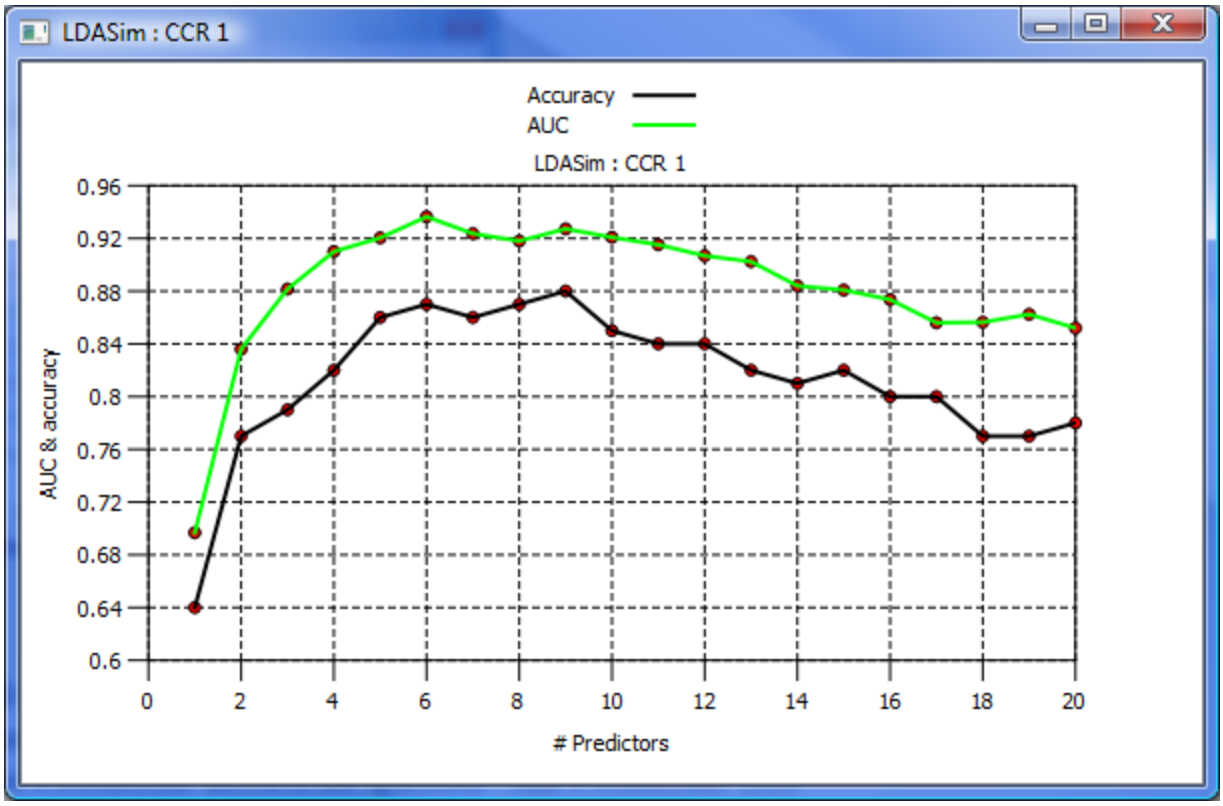


Fig. 8: CV-AUC and CV-ACC Plot

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

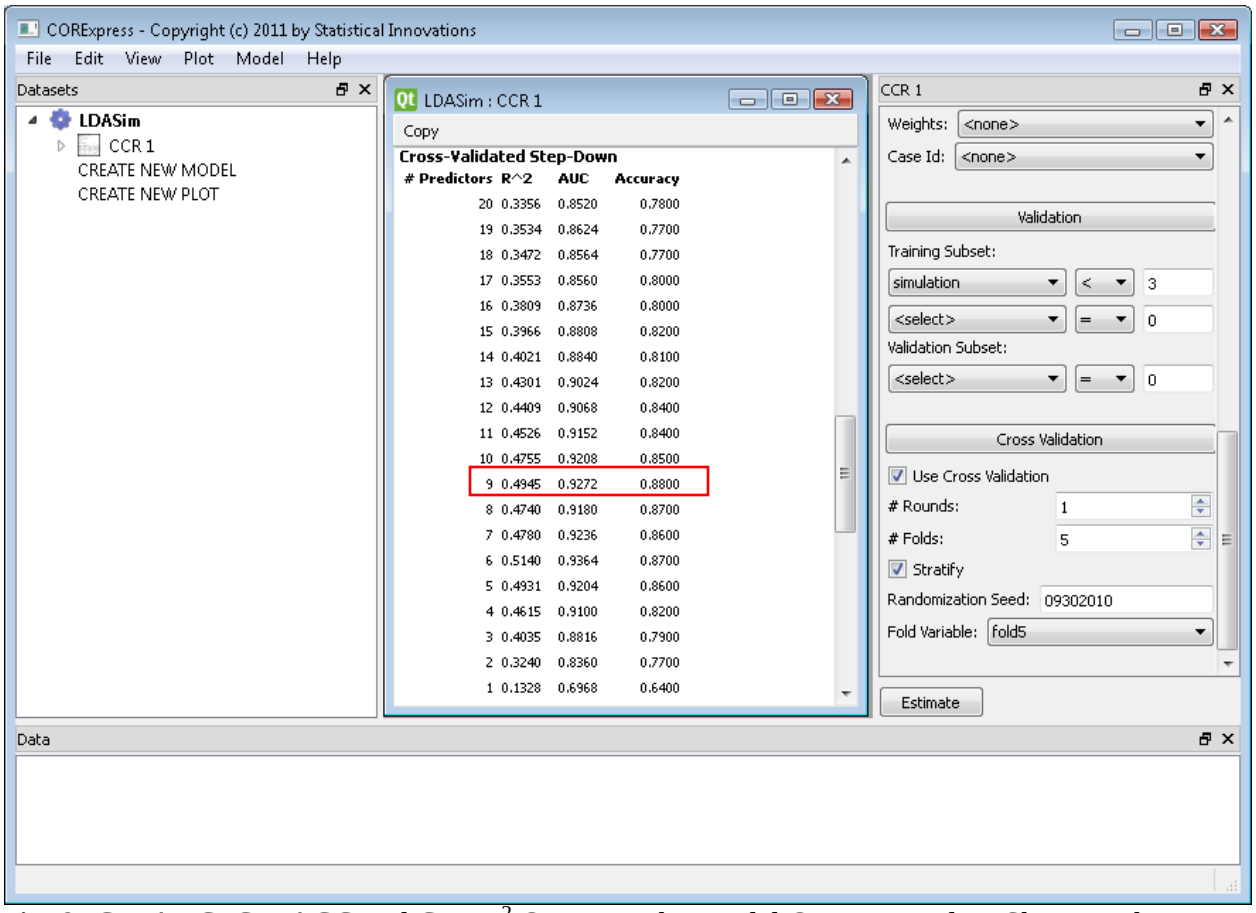


Fig. 9: CV-AUC, CV-ACC and CV- R^2 Output in the Model Output Window Showing the Highest CV-ACC Occurs with P*=9 Predictors

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

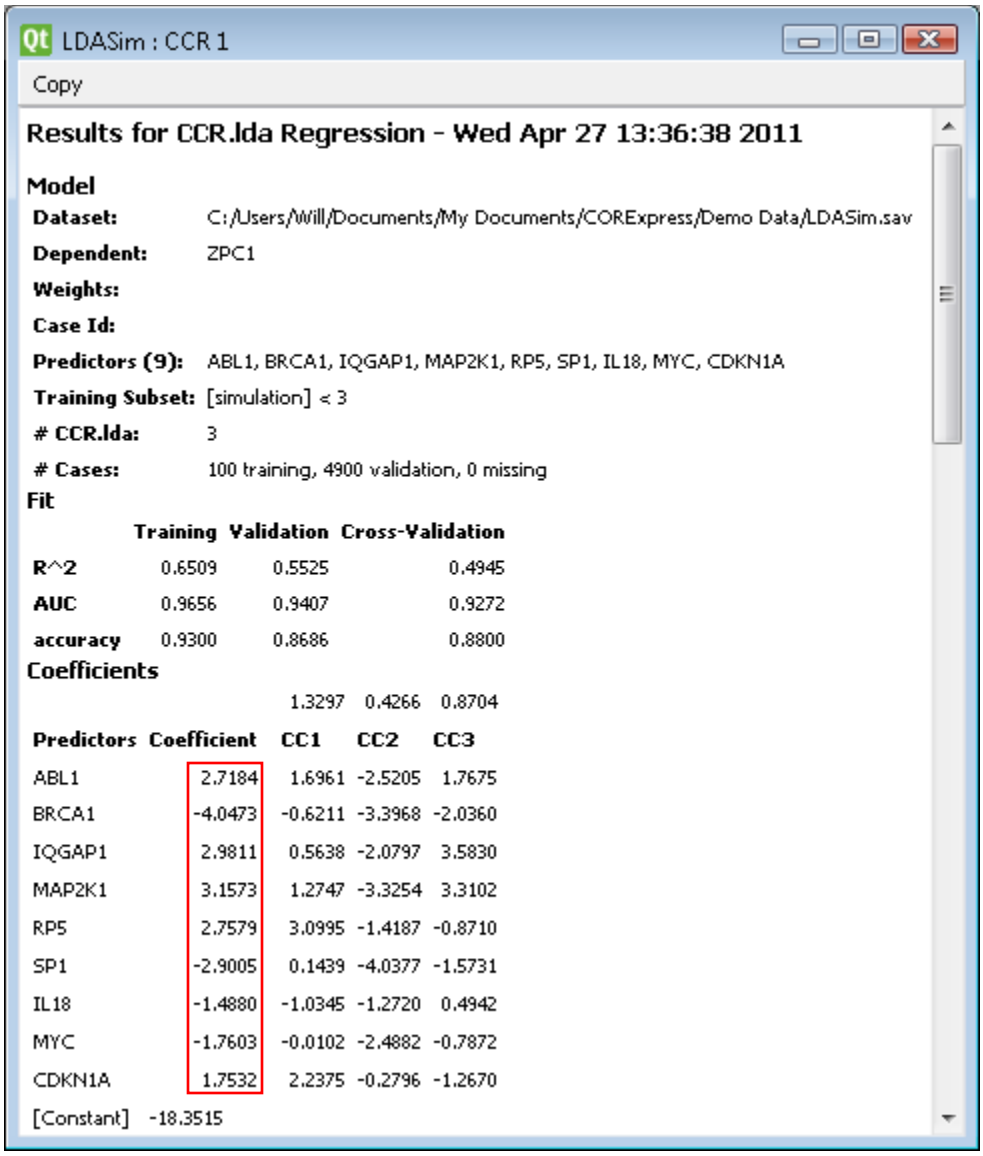


Fig. 10: Unstandardized Coefficients for K=3 in the Model Output Window

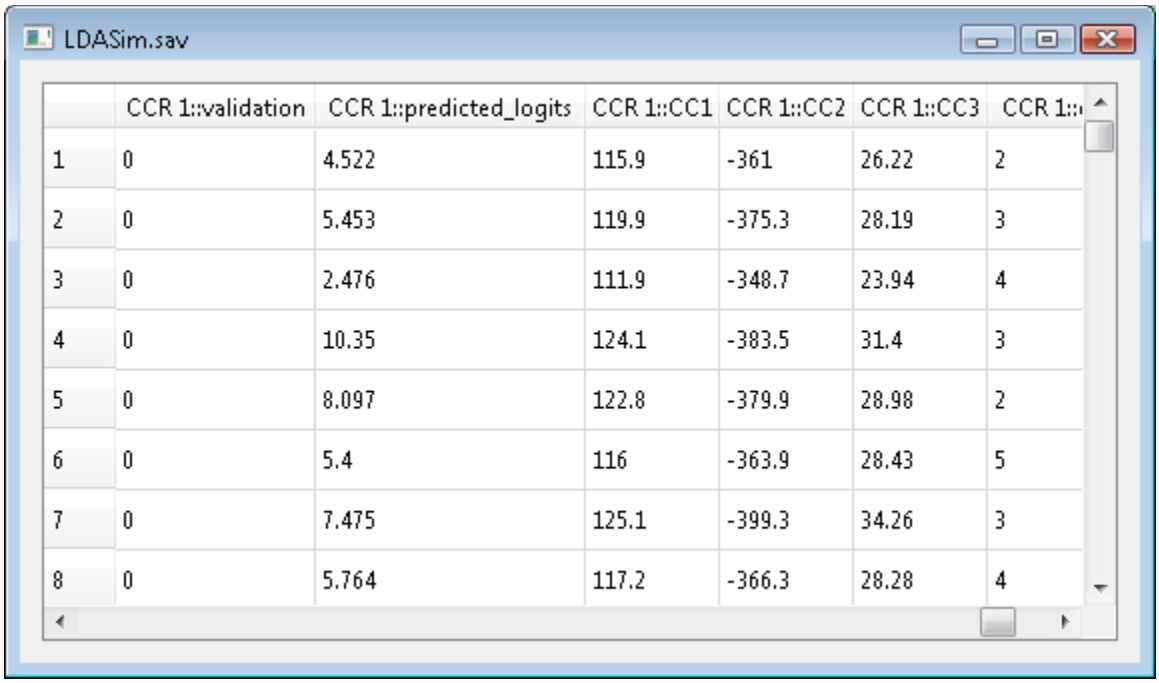
Note that for the Training, the CV-ACC=0.88 and for the Validation the ACC=0.8686.
Note that the validation accuracy and AUC are .8686, and .9407, close to the corresponding cross-validation quantities, CV-ACC = .88, and CV-AUC=0.9272. The results are quite good based on this sample -- the drop-off in 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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- 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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

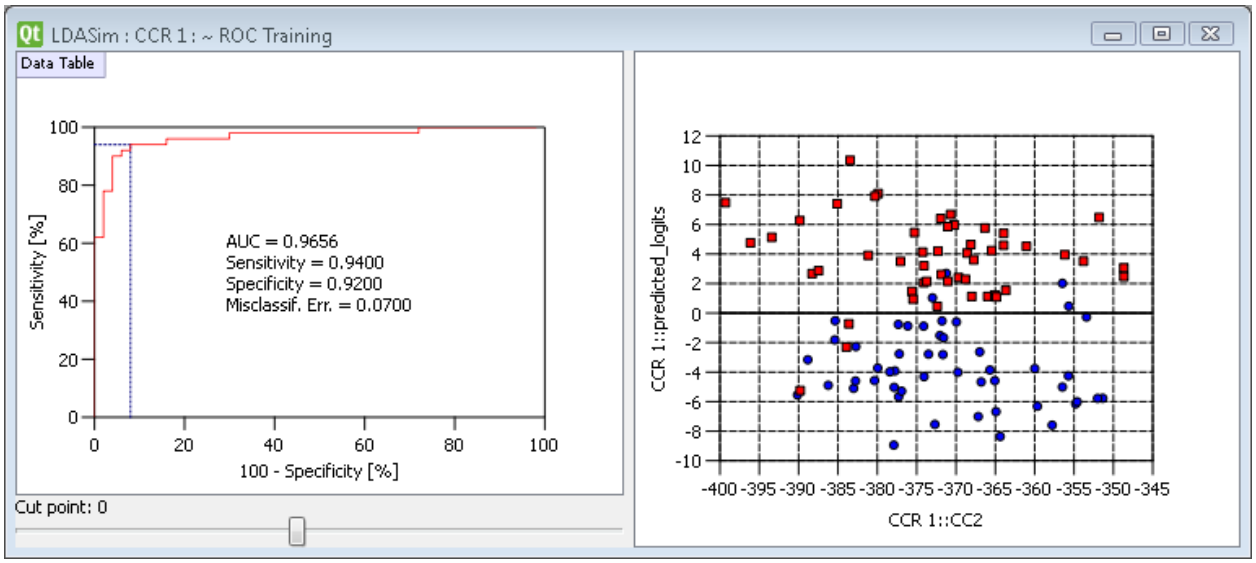


Fig. 12: Training Dataset ROC & Scatter Plot

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

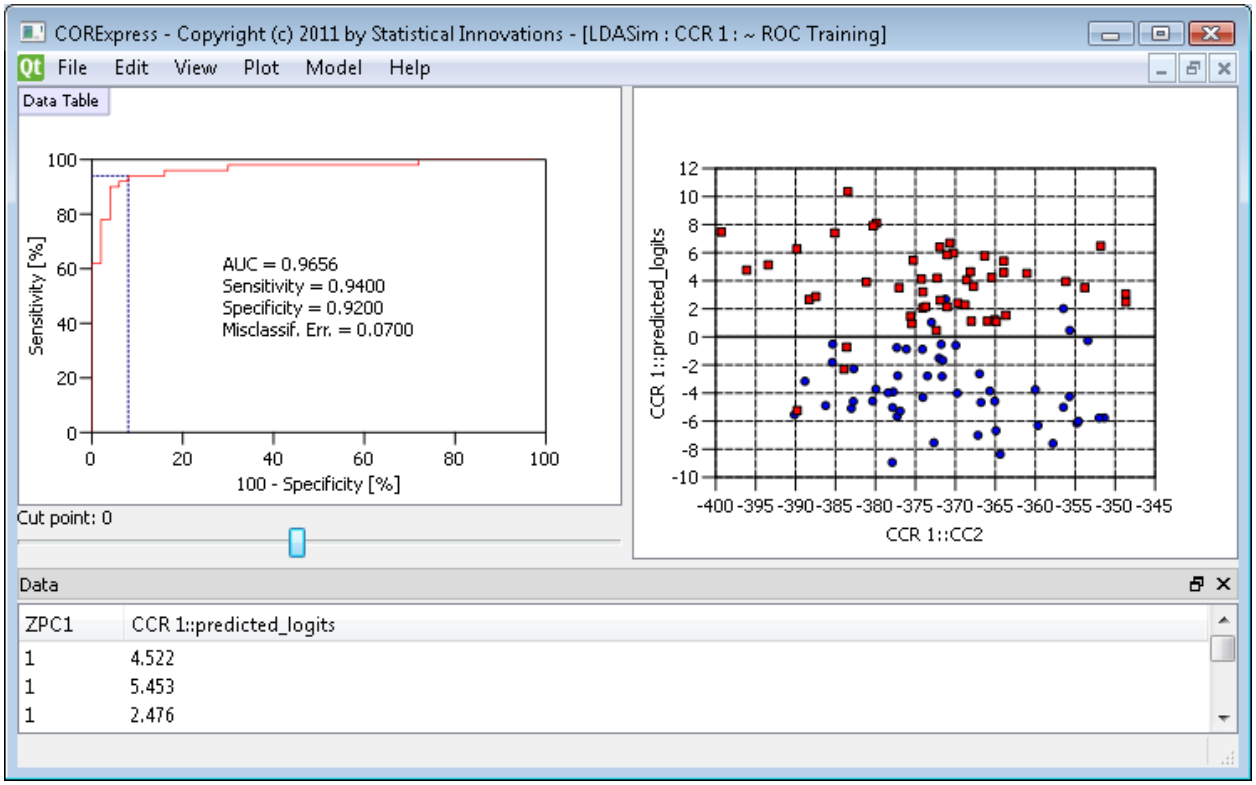


Fig. 13. Training Dataset ROC & Scatter Plot

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

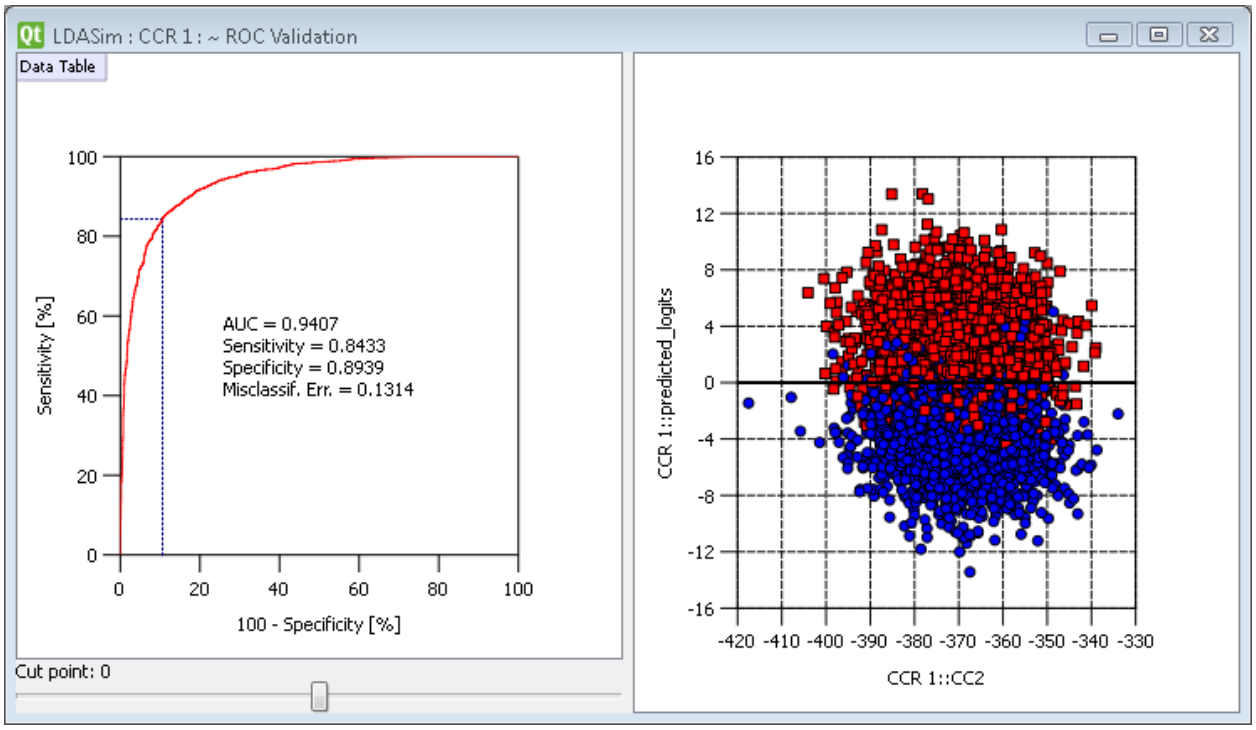


Fig. 14: Validation Dataset ROC & Scatterplot

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

Your screen should now look like this:

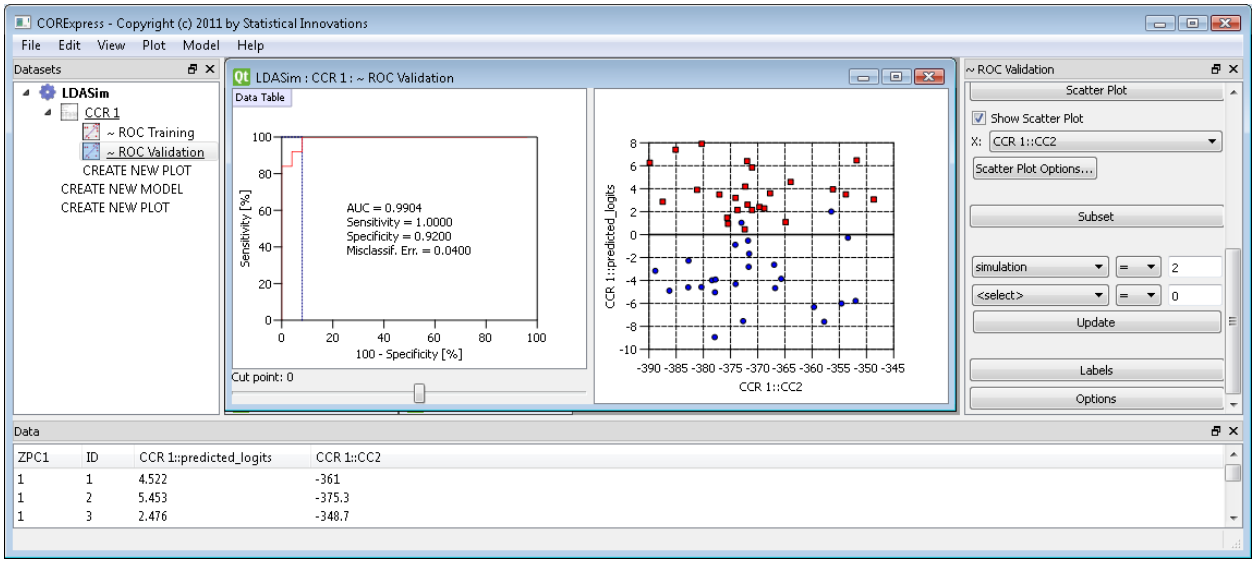


Fig. 15: Validation Dataset ROC & Scatterplot with simulation=2

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

The following window appears:

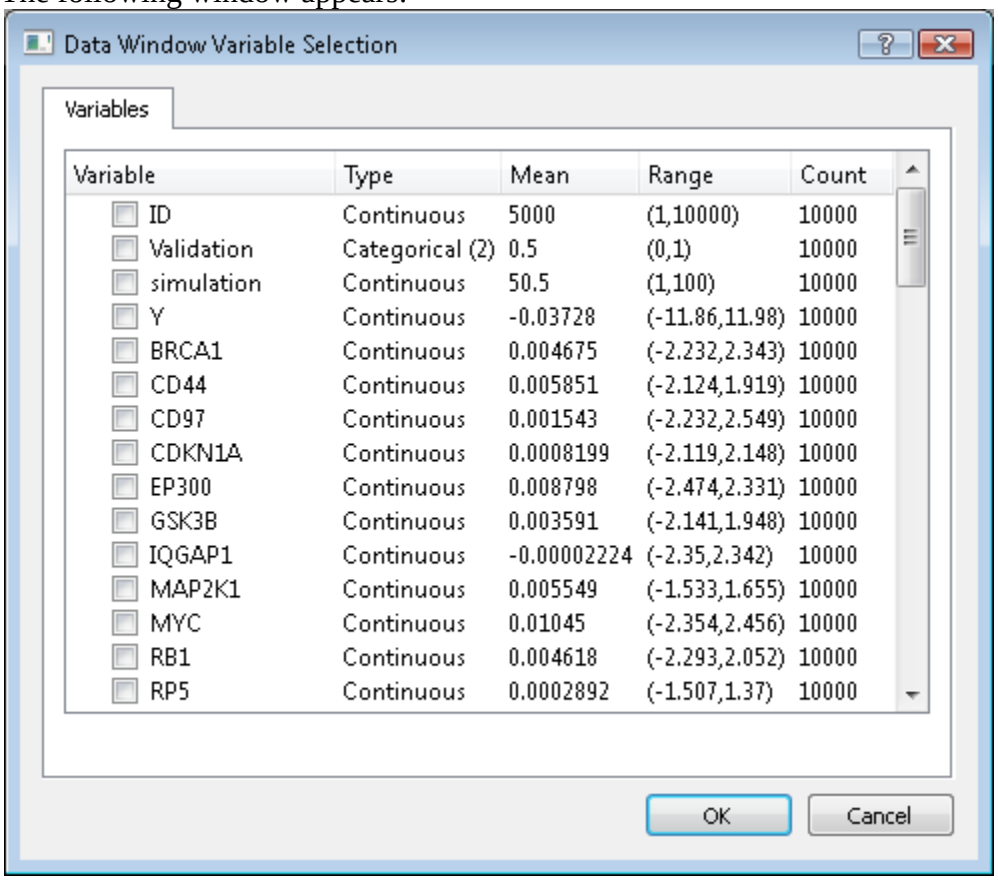


Fig. 16. Project Settings Menu Option: Data Window Variable Selection

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

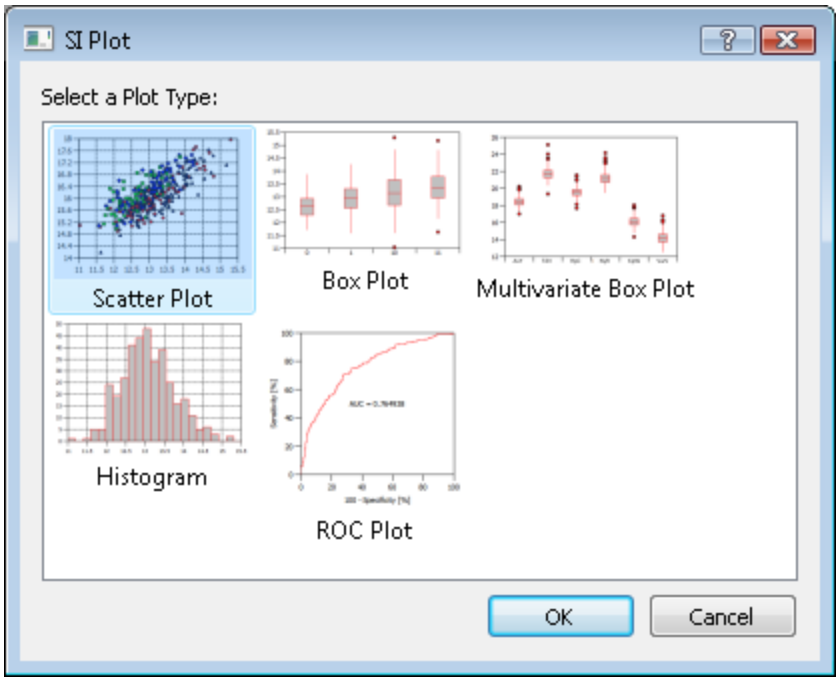


Fig. 17. Create New Plot Options

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

- 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)

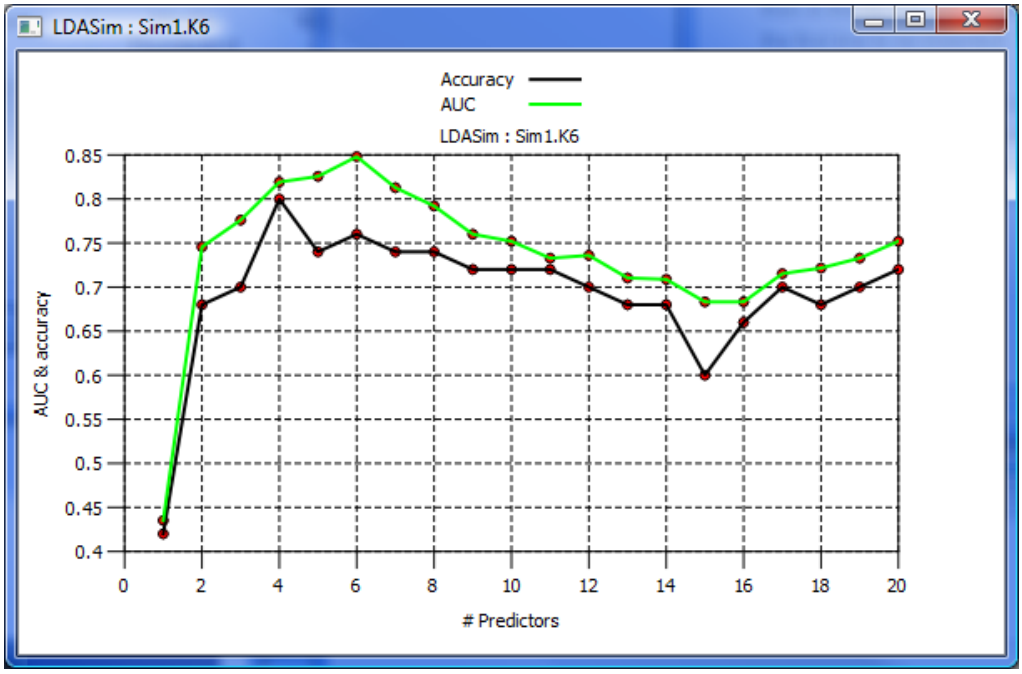


Fig 18. CV-AUC and CV-ACC Plot

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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

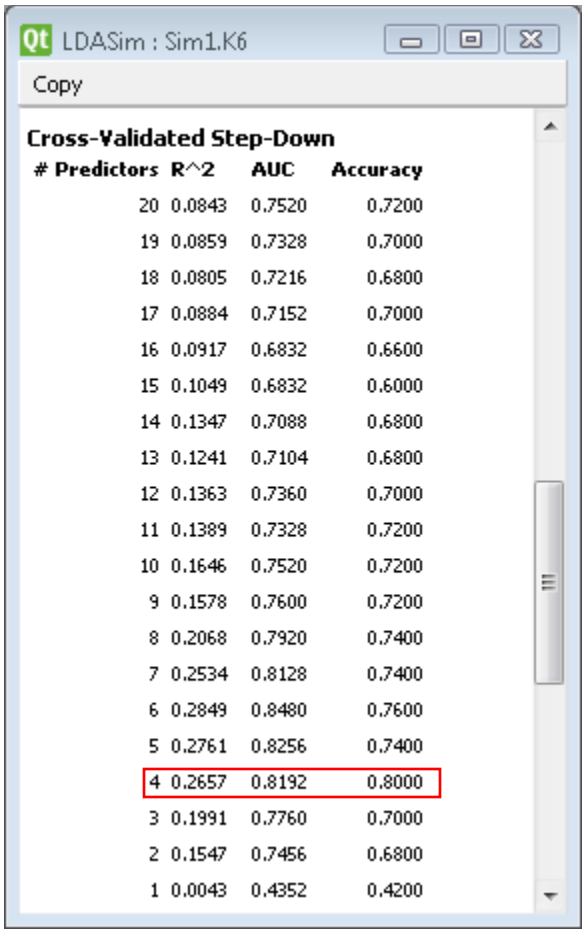


Fig. 19. Cross-Validation Step-down Output

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

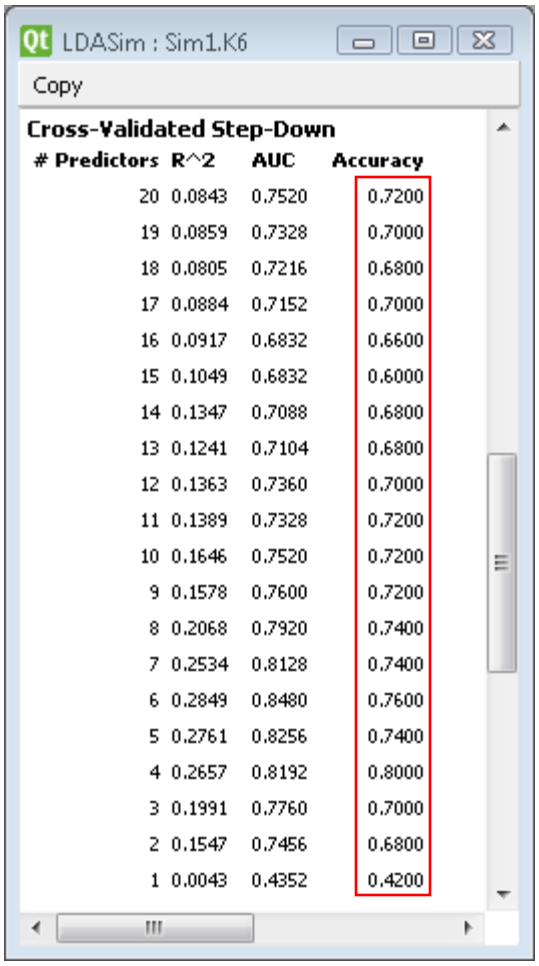


Fig. 20. Cross-Validation Step-down Output

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

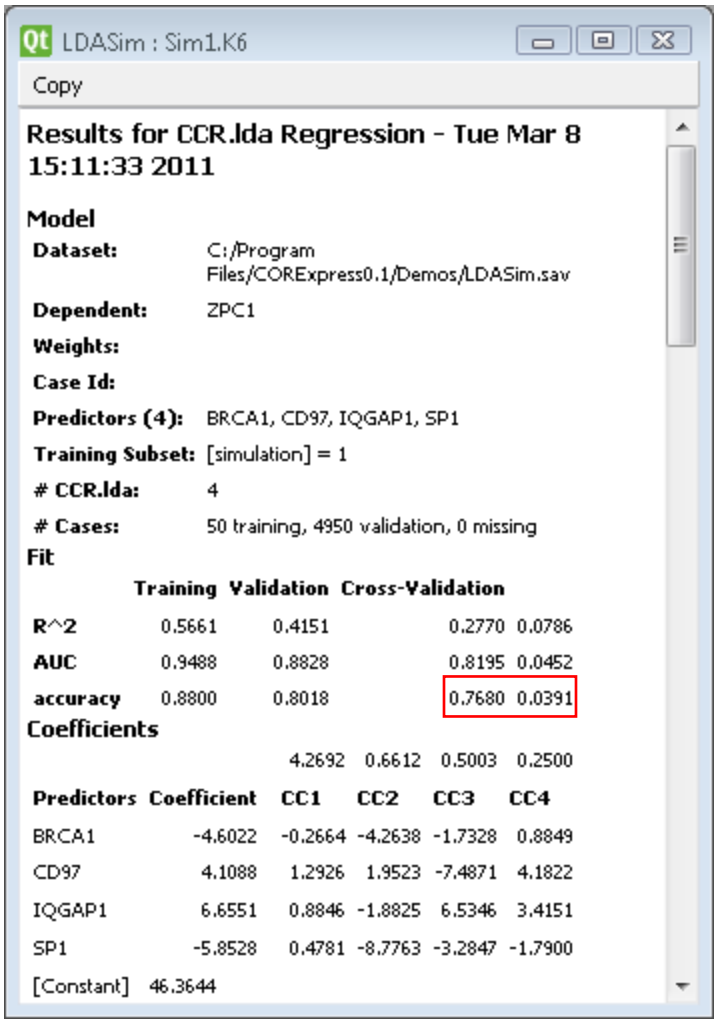


Fig. 21: Results from CCR4, Obtained based on 10 Rounds of 5-Folds
CV-ACC = 0.7680 with Associated Standard Error = 0.0391

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 © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

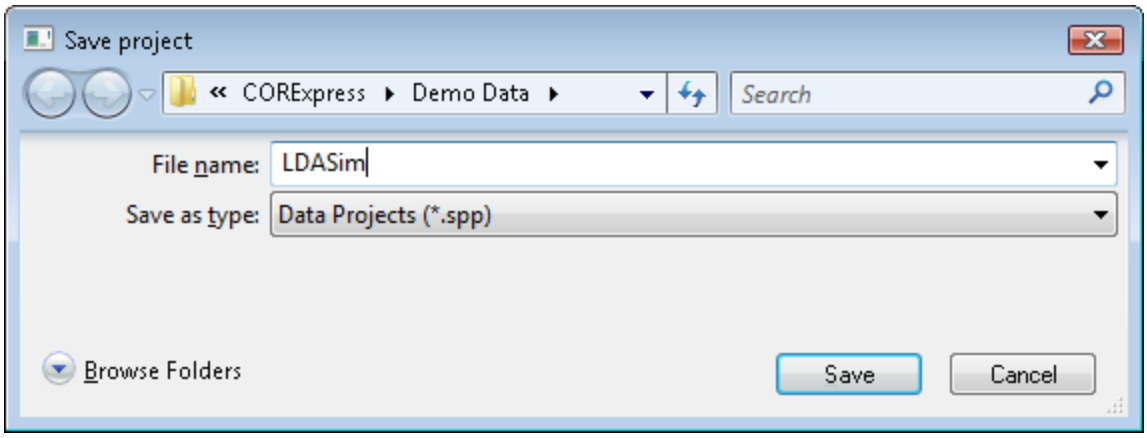


Fig. 22: Saving a Current Project

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

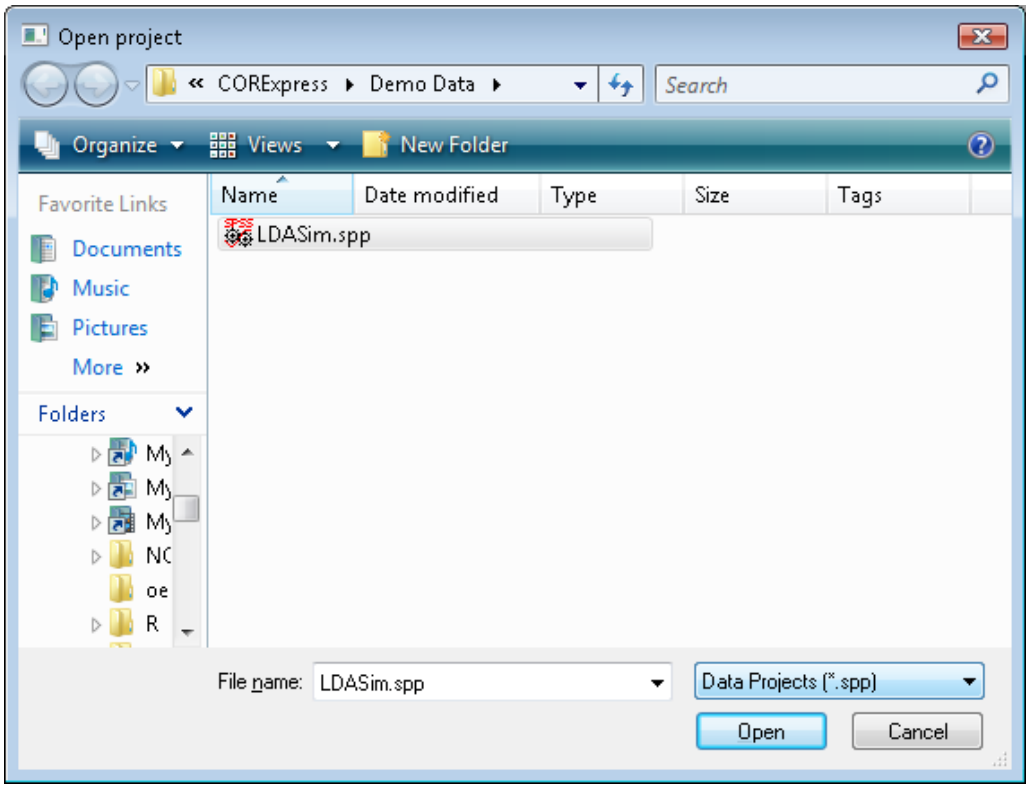


Fig. 23: Loading a previously saved project

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

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.

Copyright © 2014 by Statistical Innovations Inc. All rights reserved. (version: 12/1/2014)

10 Index

Automatic, 24, 27, 32, 43
Case ID, 28, 30, 41
CCR.lda, 28, 36, 38, 40, 42
CCR.lm, 28, 30, 32, 37, 40, 41, 42, 58
CCR.logistic, 28, 29, 36, 38, 42
CCR.pls, 30, 37
CCR.surv, 37
Coefficients, 42
Cross-Validated Component Table, 43
Cross-Validated Step-Down Table, 44
Cross-validation, 25, 27, 30, 31, 36, 41, 45
CV-ACC, 36
CV-AUC, 36, 37
CV-R^2, 26, 27, 30, 36, 37, 41, 43, 45
Fold Variable, 33
Histogram, 7, 17, 48, 53
Iterations, 29
Loadings, 42
Max # Components, 24, 27
Max # Predictors, 25, 36, 37
M-fold, 32, 43, 44, 45, 46, 60
Min # Predictors, 25, 36, 37
Missing, 21, 30
NMSE, 27, 30, 41
Predictor Table, 46
Ridge Parameter, 29
ROC Curve, 36, 38
Rounds, 15, 32
Scatter Plot, 7, 38, 48, 49, 50, 56, 57, 58
Screen, 11, 33
Step-down, 15, 24, 25, 26, 32, 36, 37, 41, 44, 45
Stratify, 32