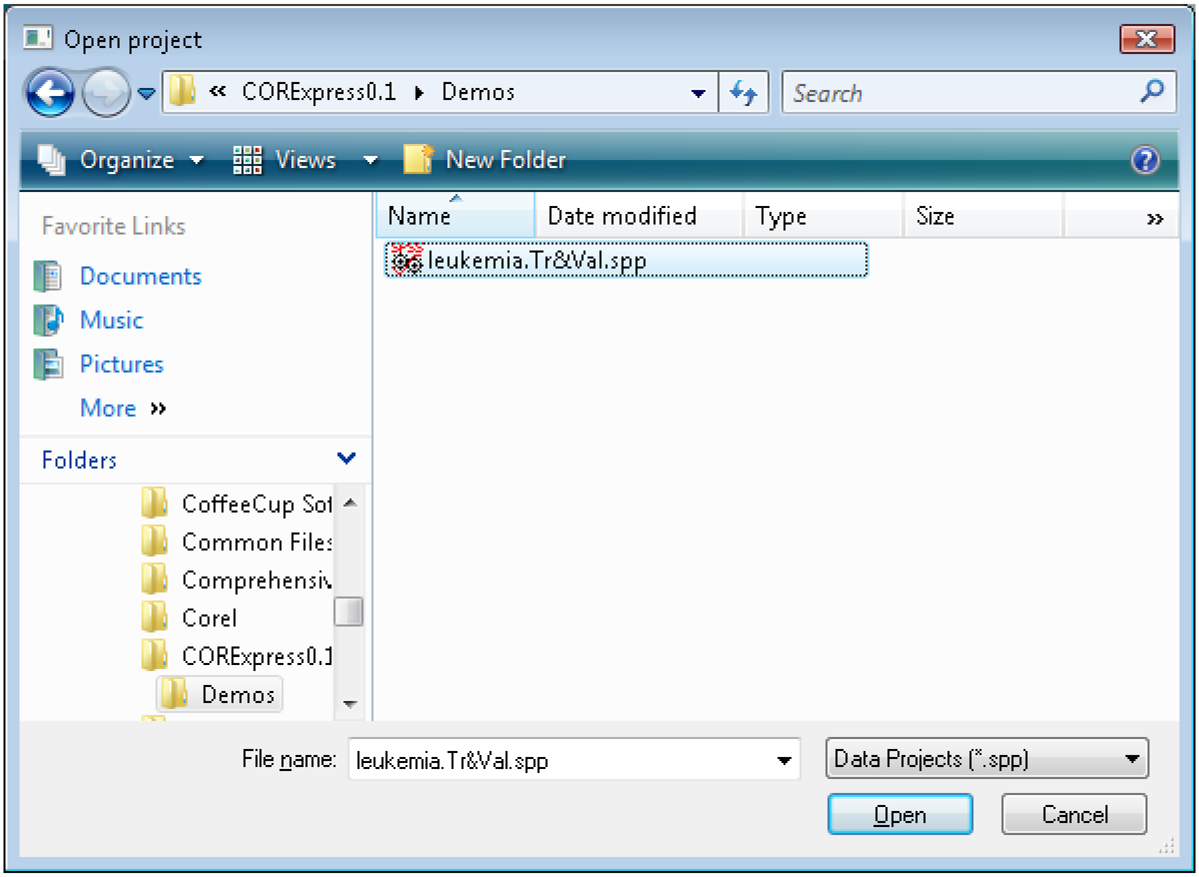
Session 2

Tutorial 2A. Estimating the Naïve Bayes Model

As the starting point for this tutorial, we will retrieve model settings that have been saved in a file. Alternatively, we could open the data file 'leukemia.Tr&Val.sav' and apply the model settings using the GUI.

Opening a Previously Saved Project:

- ➢ File → Load Project…
- ➢ Select 'leukemia.Tr&Val.spp' and click Open to load the project



leukemia.Tr&Val.spp

File name: leukemia.Tr&Val.spp

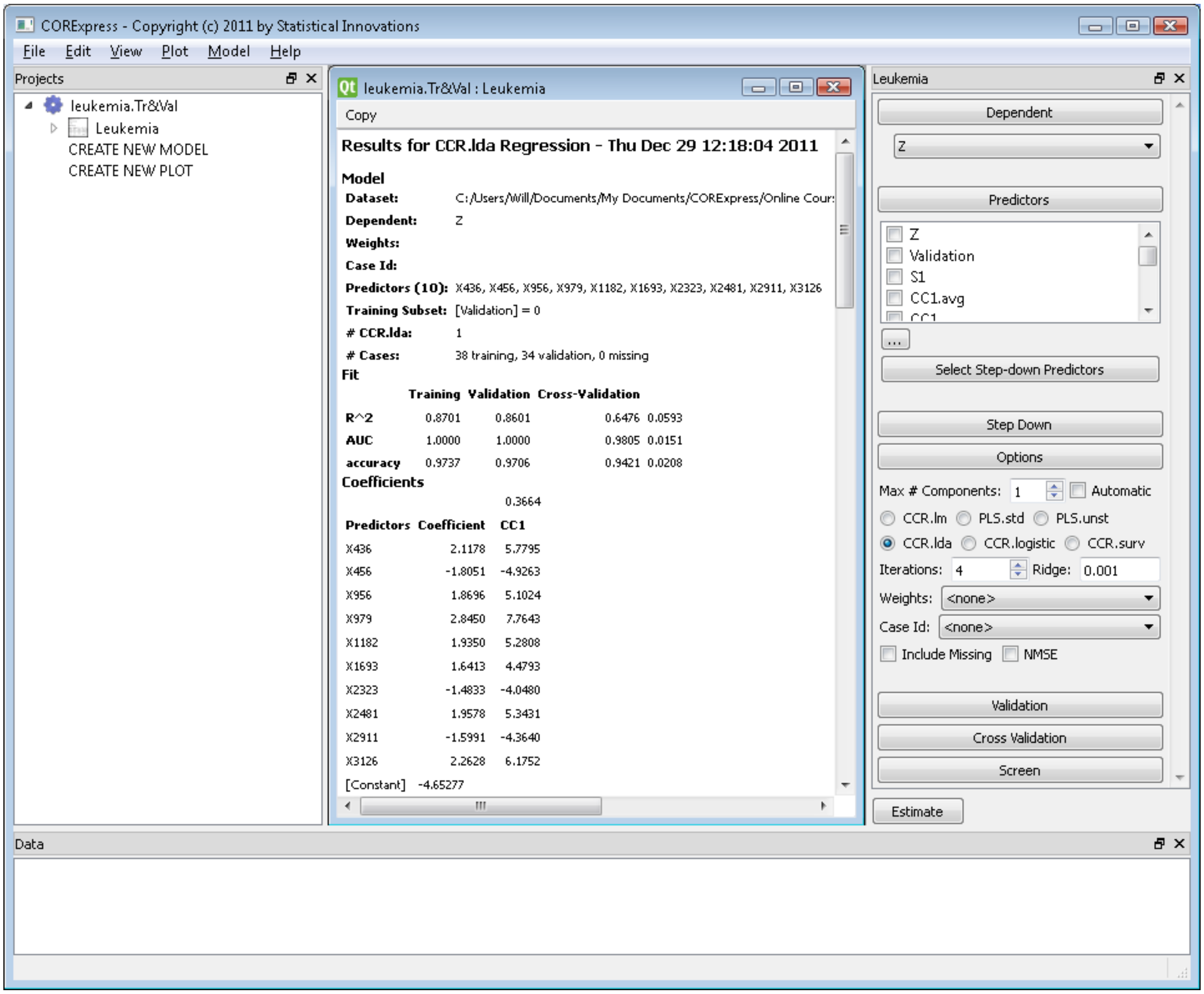
Data Projects (*.spp)

Fig. 1: Loading a previously saved project

Opening the Model Specifications for the Saved Project:

- ➢ Double click on “Leukemia” in the Datasets window

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



```
Results for CCR.lda Regression - Thu Dec 29 12:18:04 2011

Model
 Dataset:          C:/Users/Will/Documents/My Documents/CORExpress/Online Cour:
 Dependent:        Z
 Weights:
 Case Id:
 Predictors (10):  X436, X456, X956, X979, X1182, X1693, X2323, X2481, X2911, X3126
 Training Subset:  [Validation] = 0
 # CCR.lda:        1
 # Cases:          38 training, 34 validation, 0 missing
Fit
           Training  Validation  Cross-Validation
 R^2         0.8701      0.8601            0.6476  0.0593
 AUC         1.0000      1.0000            0.9805  0.0151
 accuracy    0.9737      0.9706            0.9421  0.0208
Coefficients
                         0.3664
 Predictors  Coefficient    CC1
 X436             2.1178   5.7795
 X456            -1.8051  -4.9263
 X956             1.8696   5.1024
 X979             2.8450   7.7643
 X1182            1.9350   5.2808
 X1693            1.6413   4.4793
 X2323           -1.4833  -4.0480
 X2481            1.9578   5.3431
 X2911           -1.5991  -4.3640
 X3126            2.2628   6.1752
 [Constant]  -4.65277
```

Dependent: Z

Max # Components: 1 □ Automatic

○ CCR.lm ○ PLS.std ○ PLS.unst

● CCR.lda ○ CCR.logistic ○ CCR.surv

Iterations: 4 Ridge: 0.001

Weights: <none>

Case Id: <none>

□ Include Missing □ NMSE

Fig. 2: Results from Previously Saved Project

Note the following:

- The dependent variable is *Z*, coded ‘1’ and ‘0’, to distinguish the two types of leukemia.
- We selected all 3,571 predictors as candidate predictors.
- For step down options, we have selected a minimum of 1 predictor and a maximum of 10 predictors.

- We request that 95% of predictors be removed at each step, until 100 predictors remain, at which time it resumes with the default step-down of 1 predictor at a time. For the 1-component model, choosing the '95%' option speeds up the process substantially, without altering the solution at all.
- CCR.lda has been selected with 1 component (this specifies the Naïve Bayes model, under the assumptions of Linear Discriminant Analysis).
- The cases to be used to develop the model (analysis sample) is the training data set, defined by 'Validation=0' (N=38 cases). (The remaining 34 cases, 'Validation'=1, are retained for use as a 'test' set).
- We selected the "Use Cross Validation" option, requesting the use of 10 rounds of 6-folds on the training data.
- We selected the "Stratify" option, which attempts to equalize the folds with respect to the distribution of the dependent variable Z.

Estimate the Model:

➢ Click Estimate

The program quickly reduces the number of predictors in the model from 3,571 to 10.

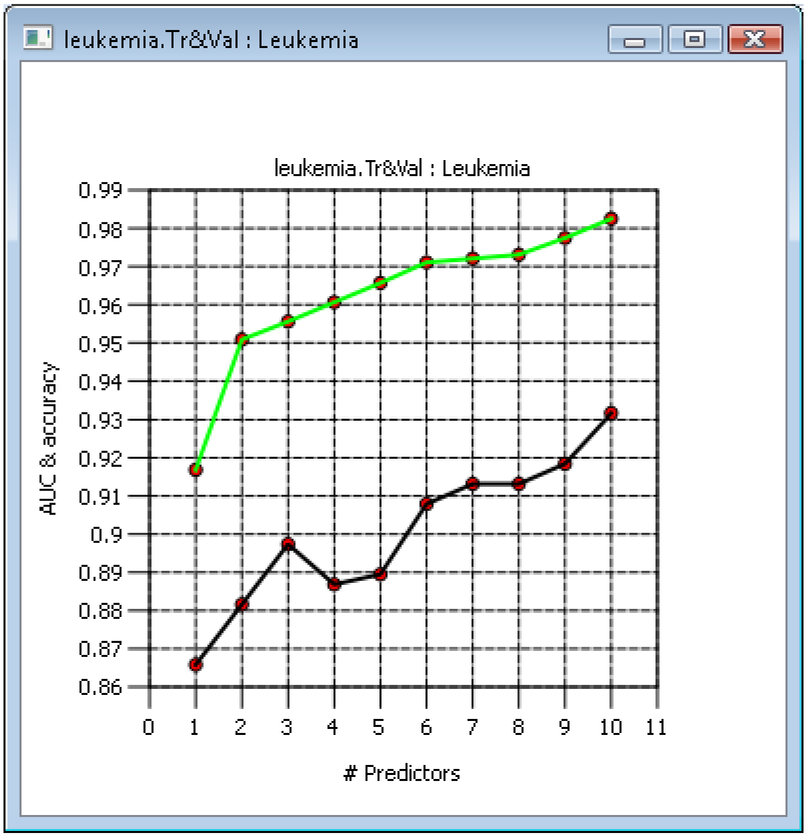
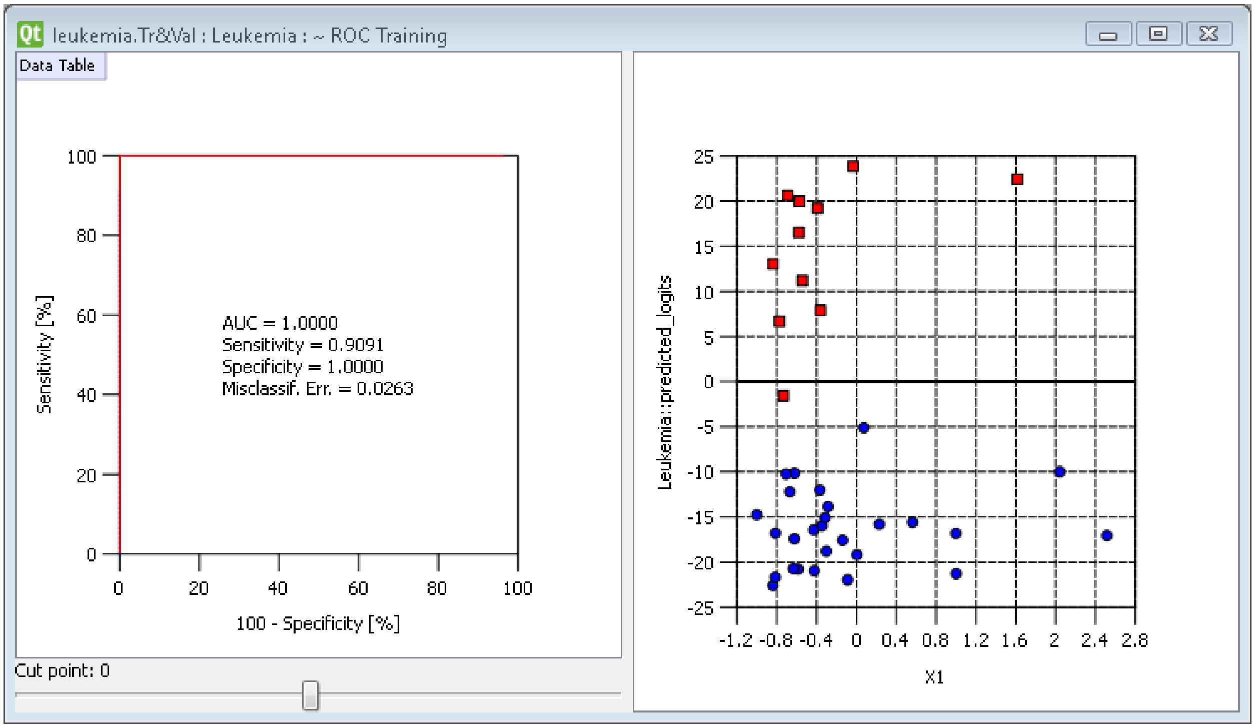


Fig. 3: CV-AUC and CV-ACC Plot

Viewing the Training Interactive Plot:

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

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



AUC = 1.0000
Sensitivity = 0.9091
Specificity = 1.0000
Misclassif. Err. = 0.0263

Fig. 4: Training Dataset ROC & Scatter Plot

Notice that AUC = 1, meaning that all cases coded Z=1 have higher predicted scores than any case coded Z=0.

The scatterplot shows that 1 of the red square symbols (representing cases in the Z=1 group) is misclassified based on the default logit score cut-point of 0 (logit = 0 corresponds to predicted probability of .5). Reducing the cut-point, lowers the classification line to produce 100% correct classification.

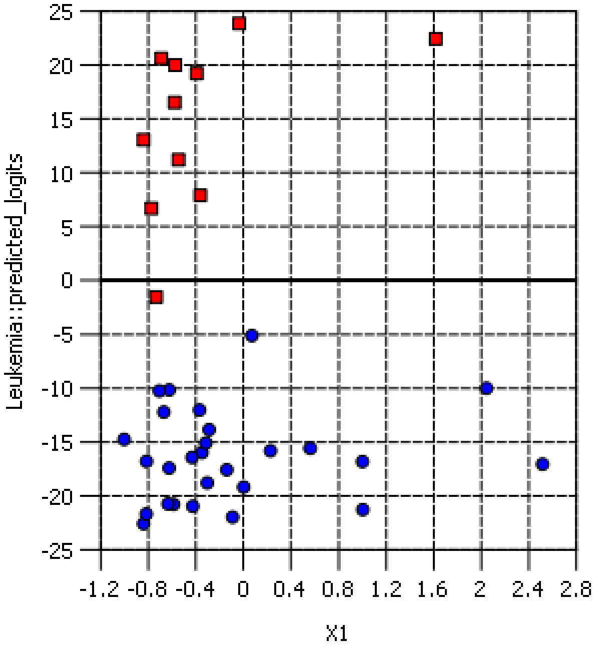


Fig. 5: Training Data

Using the cutpoint slider below the ROC curve, you can adjust the cutpoint so that all the red subjects are above the cutpoint and all blue subjects are below the line.

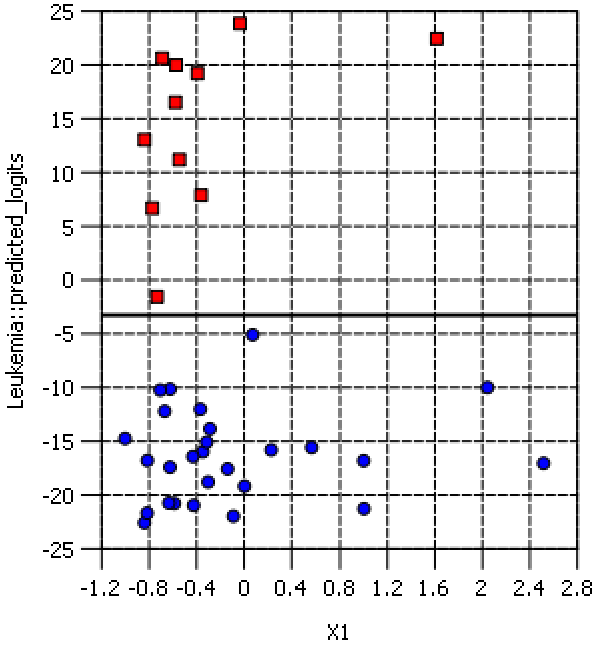


Fig. 6: Training Data with adjusted cutpoint

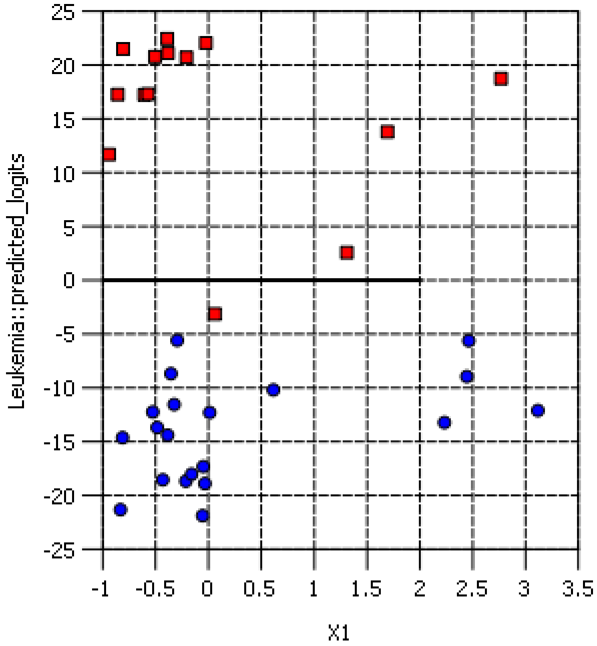


Fig. 7: Validation Data

For the validation data, we also obtain AUC = 1, which confirms that the results obtained using the training data were not due to chance.

As an exercise, under "Step Down", change the percent from '95' to '50', re-estimate the model and confirm that you get the same results.

Change Remove by Percent:

- Double click on "Leukemia" in the Datasets window to bring up the model specifications in the Control Window.
- Under Stepdown, delete '95' and type '50'
- Click Estimate

The reason that the results are the same is that the loadings from the 1-component model are not affected by the number of predictors in the model, eliminating predictors that have the smallest magnitude of the standardized coefficient is equivalent to eliminating predictors that have the smallest magnitude of the loadings. Thus, it does not matter how many predictors are eliminated at the same – the 10 predictors remaining in the model will be same.

To compare the coefficients with those obtained from LDA, we can estimate the LDA model with COREXpress. LDA is equivalent to a saturated CCR model, which with 10 predictors occurs by changing the number of components from 1 to 10.

To select these 10 predictors do the following:

Select Step-down Predictors:

- In the Control Window, click the “Select Step-down Predictors” button below the Predictors window

This will check off only the 10 predictors from the previously estimated model. Note that the “Step Down” feature will automatically be unchecked.

Change Number of Components:

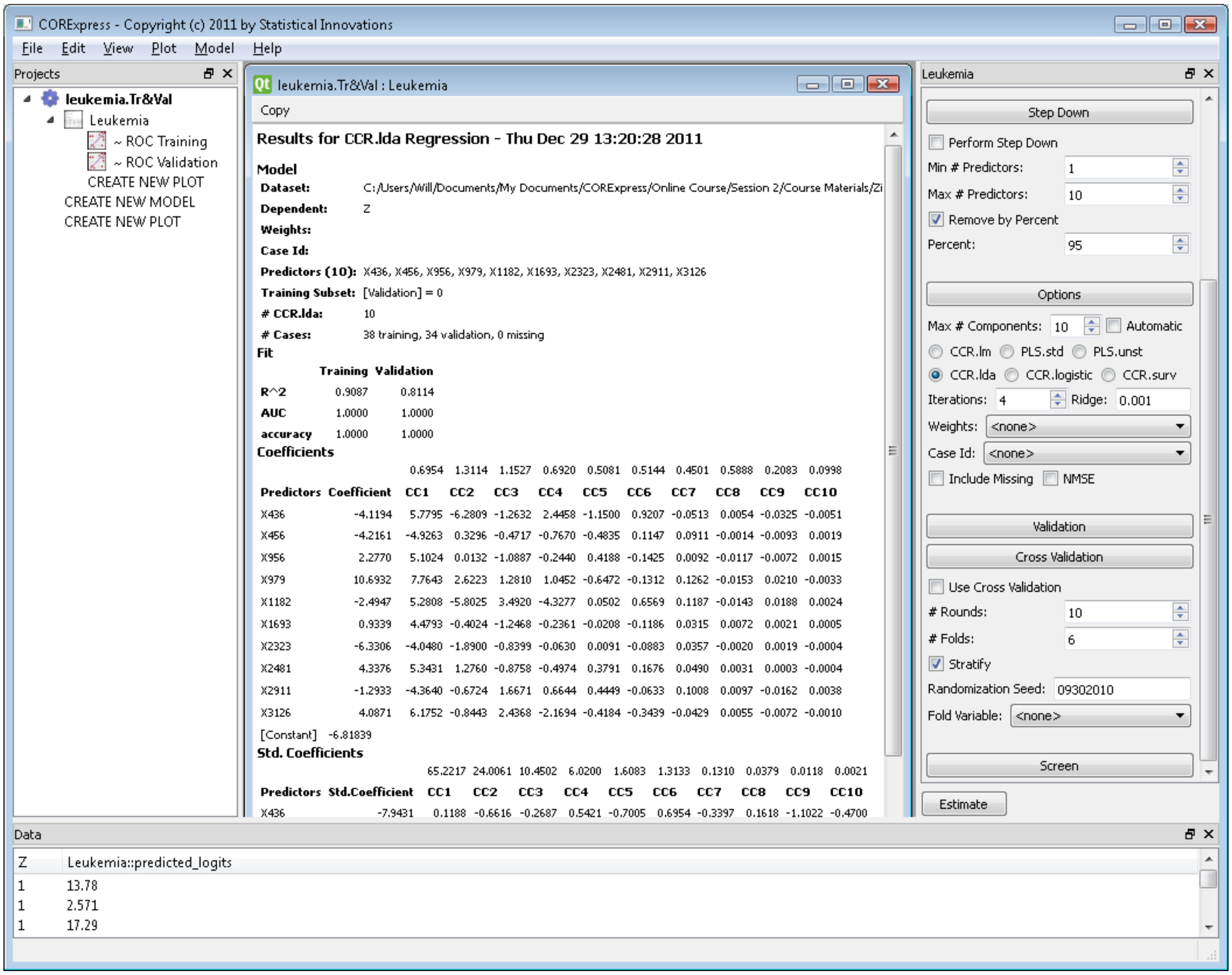
- Change the # components from 1 to 10 to estimate the saturated model

Turn off the Cross Validation Option:

- Scroll down and uncheck the “Use Cross Validation” option under “Cross Validation”

Estimate the Model:

- Click Estimate



Results for CCR.lda Regression - Thu Dec 29 13:20:28 2011

Model

Dataset:	C:/Users/Will/Documents/My Documents/CORExpress/Online Course/Session 2/Course Materials/Zi
Dependent:	Z
Weights:	
Case Id:	
Predictors (10):	X436, X456, X956, X979, X1182, X1693, X2323, X2481, X2911, X3126
Training Subset:	[Validation] = 0
# CCR.lda:	10
# Cases:	38 training, 34 validation, 0 missing

Fit

	Training	Validation
R^2	0.9087	0.8114
AUC	1.0000	1.0000
accuracy	1.0000	1.0000

Coefficients

		0.6954	1.3114	1.1527	0.6920	0.5081	0.5144	0.4501	0.5888	0.2083	0.0998
Predictors	**Coefficient**	**CC1**	**CC2**	**CC3**	**CC4**	**CC5**	**CC6**	**CC7**	**CC8**	**CC9**	**CC10**
X436	-4.1194	5.7795	-6.2809	-1.2632	2.4458	-1.1500	0.9207	-0.0513	0.0054	-0.0325	-0.0051
X456	-4.2161	-4.9263	0.3296	-0.4717	-0.7670	-0.4835	0.1147	0.0911	-0.0014	-0.0093	0.0019
X956	2.2770	5.1024	0.0132	-1.0887	-0.2440	0.4188	-0.1425	0.0092	-0.0117	-0.0072	0.0015
X979	10.6932	7.7643	2.6223	1.2810	1.0452	-0.6472	-0.1312	0.1262	-0.0153	0.0210	-0.0033
X1182	-2.4947	5.2808	-5.8025	3.4920	-4.3277	0.0502	0.6569	0.1187	-0.0143	0.0188	0.0024
X1693	0.9339	4.4793	-0.4024	-1.2468	-0.2361	-0.0208	-0.1186	0.0315	0.0072	0.0021	0.0005
X2323	-6.3306	-4.0480	-1.8900	-0.8399	-0.0630	0.0091	-0.0883	0.0357	-0.0020	0.0019	-0.0004
X2481	4.3376	5.3431	1.2760	-0.8758	-0.4974	0.3791	0.1676	0.0490	0.0031	0.0003	-0.0004
X2911	-1.2933	-4.3640	-0.6724	1.6671	0.6644	0.4449	-0.0633	0.1008	0.0097	-0.0162	0.0038
X3126	4.0871	6.1752	-0.8443	2.4368	-2.1694	-0.4184	-0.3439	-0.0429	0.0055	-0.0072	-0.0010
[Constant]	-6.81839										

Std. Coefficients

		65.2217	24.0061	10.4502	6.0200	1.6083	1.3133	0.1310	0.0379	0.0118	0.0021
Predictors	**Std.Coefficient**	**CC1**	**CC2**	**CC3**	**CC4**	**CC5**	**CC6**	**CC7**	**CC8**	**CC9**	**CC10**
X436	-7.9431	0.1188	-0.6616	-0.2687	0.5421	-0.7005	0.6954	-0.3397	0.1618	-1.1022	-0.4700

□ Perform Step Down
Min # Predictors: 1
Max # Predictors: 10
☑ Remove by Percent
Percent: 95

Max # Components: 10 □ Automatic
○ CCR.lm ○ PLS.std ○ PLS.unst
● CCR.lda ○ CCR.logistic ○ CCR.surv
Iterations: 4 Ridge: 0.001
Weights: <none>
Case Id: <none>
□ Include Missing □ NMSE

□ Use Cross Validation
Rounds: 10
Folds: 6
☑ Stratify
Randomization Seed: 09302010
Fold Variable: <none>

Z	Leukemia::predicted_logits
1	13.78
1	2.571
1	17.29

Fig. 8:

Notice that the coefficients reported under the first component CC1 are unchanged from the Naïve Bayes (NB) model. NB can be viewed as an extreme form of regularization, where all components except for the first component are set to 0.