Session 2

Tutorial 2A. Estimating the Naïve Bayes Model

Datafile and Saved Model Definition

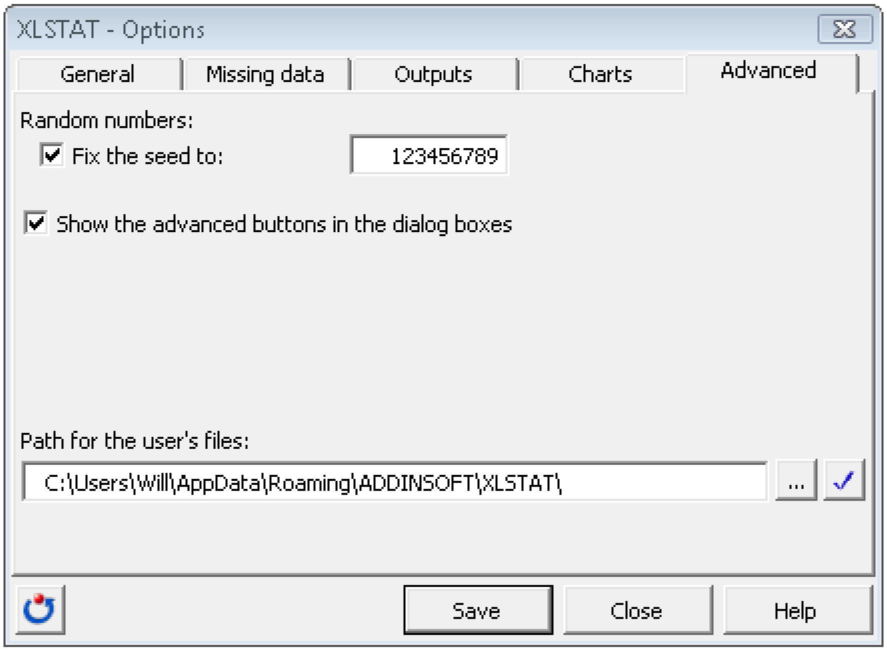
Using saved model results, this tutorial will illustrate how M-fold CV should *not* be performed.

Download this datafile: leukemia.Tr&Val.xlsx

Using this saved model definition: leukemiaTr&Val.txt

Enabling advanced options

Once XLSTAT-Pro is activated, go to the menu **Options**, and in the tab **Advanced** enable the option named **Show the advanced buttons in the dialog boxes**.



Random numbers:

☑ Fix the seed to: 123456789

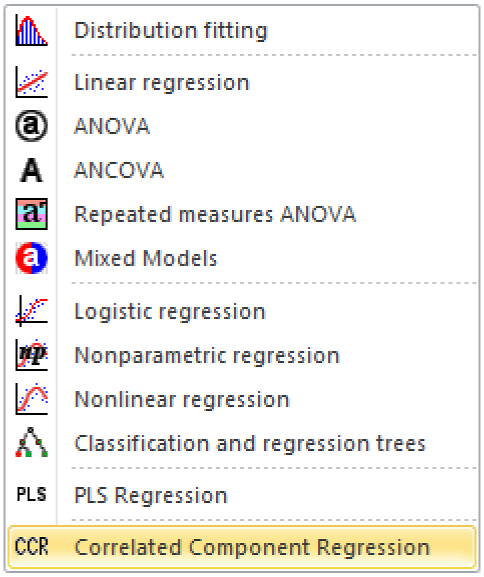
☑ Show the advanced buttons in the dialog boxes

Path for the user's files: C:\Users\Will\AppData\Roaming\ADDINSOFT\XLSTAT\

Figure 1. Advanced tab of XLSTAT Options Dialog Box

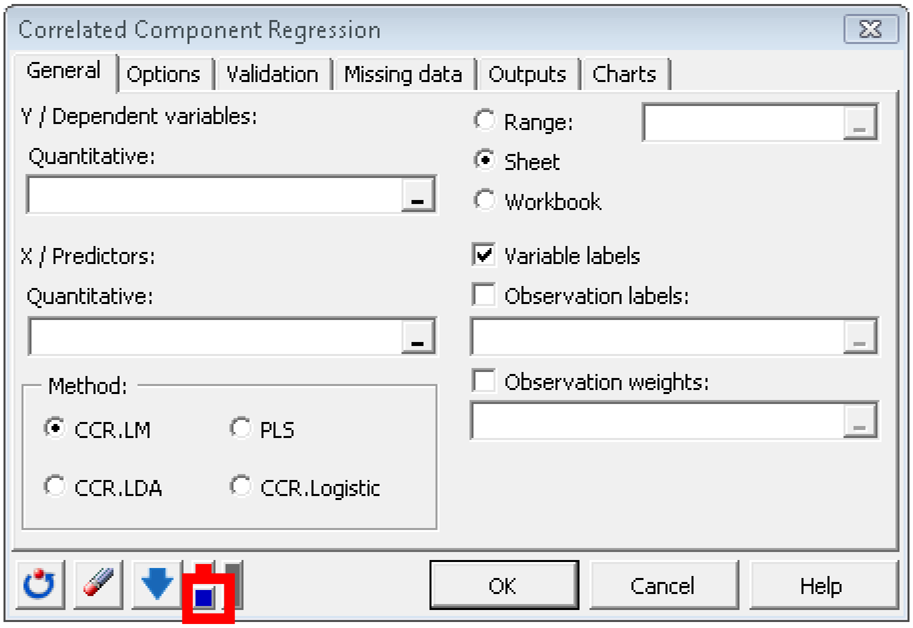
Opening a Previously Setup Correlated Component Regression

To activate the Correlated Component Regression dialog box, first start XLSTAT by clicking on the button in the Excel toolbar, then select the **XLSTAT / Modeling data / Correlated Component Regression** command in the Excel menu or click the corresponding button on the **Modeling data** toolbar.



Once you have clicked the button, the **Correlated Component Regression** dialog box is displayed with the Method=CCR.LM (linear regression model) selected by default.

When the dialog box is open click on the blue button to load the code.



Y / Dependent variables:

Quantitative:

○ Range:

● Sheet

○ Workbook

X / Predictors:

Quantitative:

☑ Variable labels

□ Observation labels:

□ Observation weights:

Method:

● CCR.LM ○ PLS

○ CCR.LDA ○ CCR.Logistic

Click this button to load previously saved settings for this dialog box from a file.

Figure 2. Loading previously saved model settings for CCR dialog box.

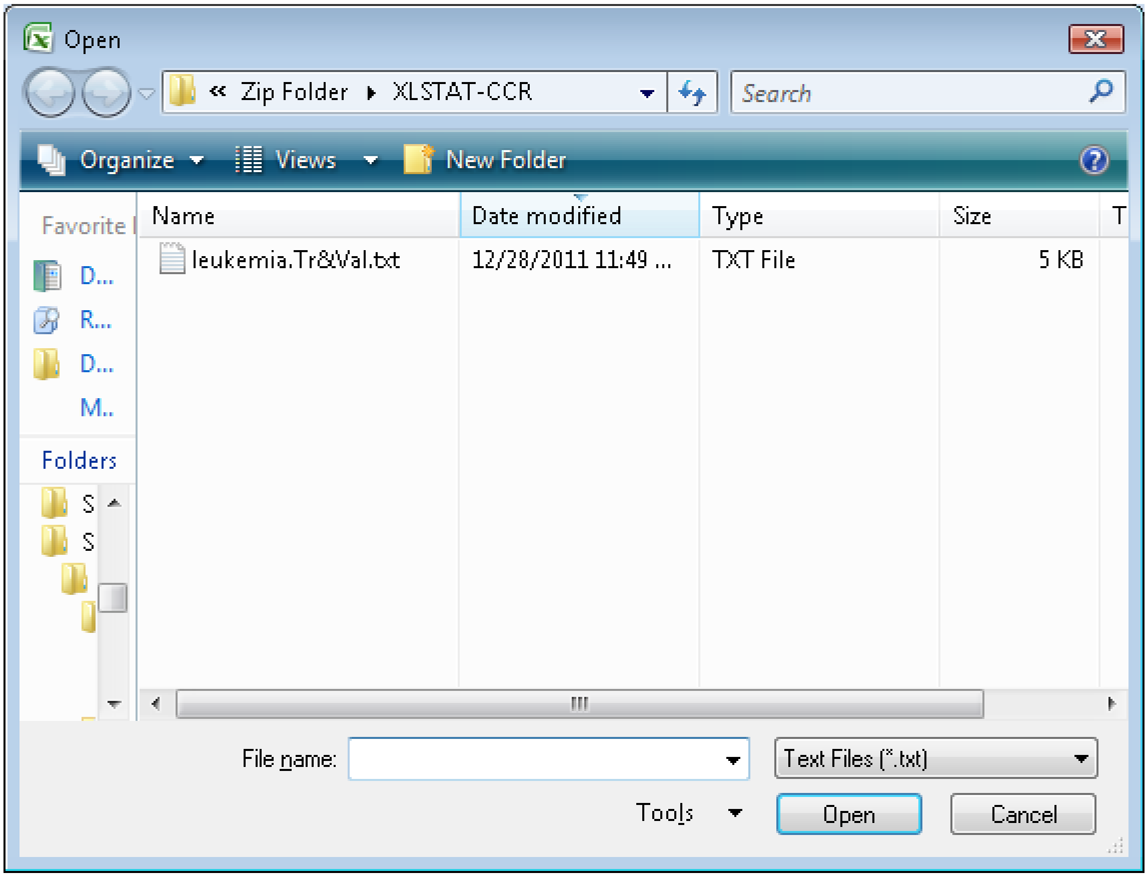


Figure 3. Loading previously saved model settings for CCR dialog box.

Open the text file “leukemiaTr&Val.txt” to load previously saved settings for this dialog box from a file.

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.

Now simply click on **OK.**

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

From the drop down menu, select 'Cross-validation step-down table'. Scroll down to the 'Cross-validation step-down plot' (Row 10,835).

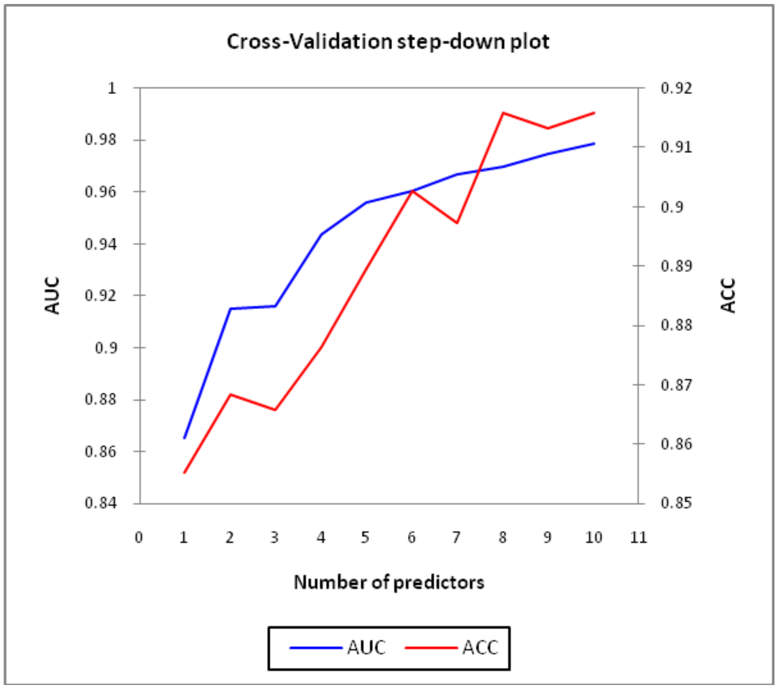


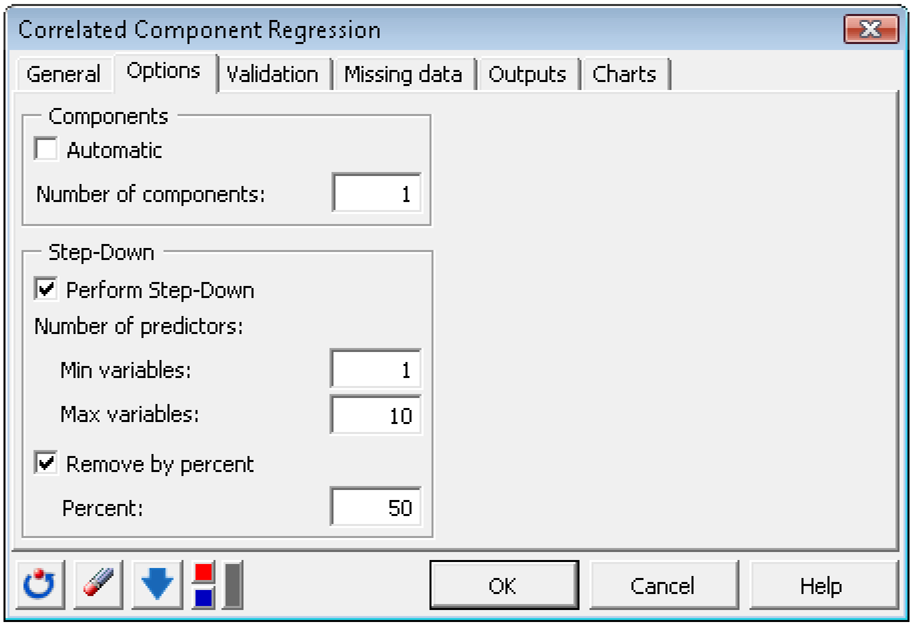
Fig. 3: CV-AUC and CV-ACC Plot

From the drop down menu, select 'Goodness of fit statistics' (Row 10,747).

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

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, open the CCR dialog box. In the Options section, under "Step Down", change the percent from '95' to '50', re-estimate the model and confirm that you get the same results.



Components

☐ Automatic

Number of components: 1

Step-Down

☑ Perform Step-Down

Number of predictors:

Min variables: 1

Max variables: 10

☑ Remove by percent

Percent: 50

Fig. 4: Changing the 'Remove by Percent' Field

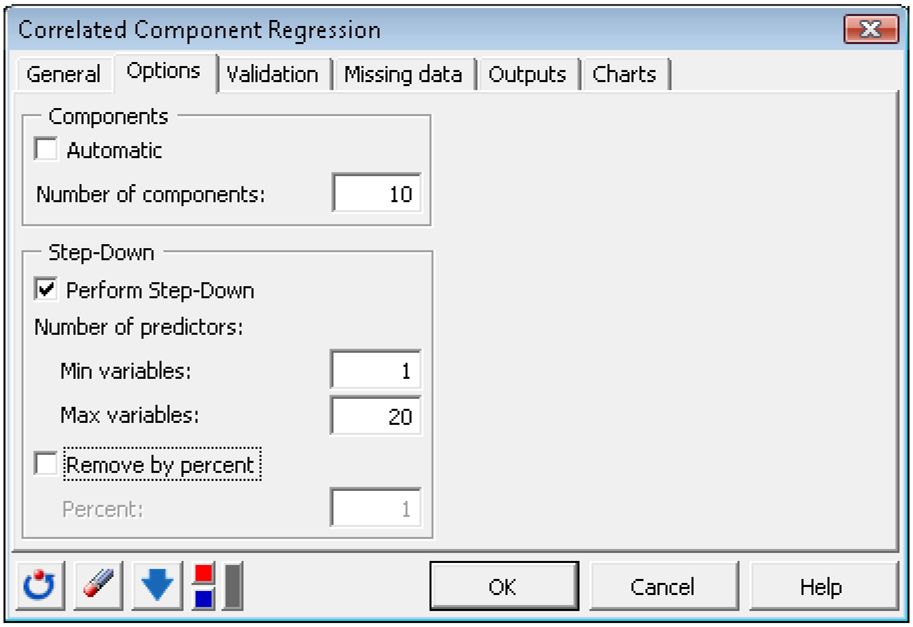
The reason that the results are the same is that the loadings from the 1-component model are not affected by the number of predictors in the model, eliminating predictors that have the smallest magnitude of the standardized coefficient is equivalent to eliminating predictors that have the smallest magnitude 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 XLSTAT-CCR. LDA is equivalent to a saturated CCR model, which with 10 predictors occurs by changing the number of components from 1 to 10.

Select the 10 predictors (for convenience, we have included the 10 predictors in the tab "10 preds" and included the saved model definition "leukemiaTr&Val10pred.txt").

Change Number of Components:

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



Components

□ Automatic

Number of components: 10

Step-Down

☑ Perform Step-Down

Number of predictors:

Min variables: 1

Max variables: 20

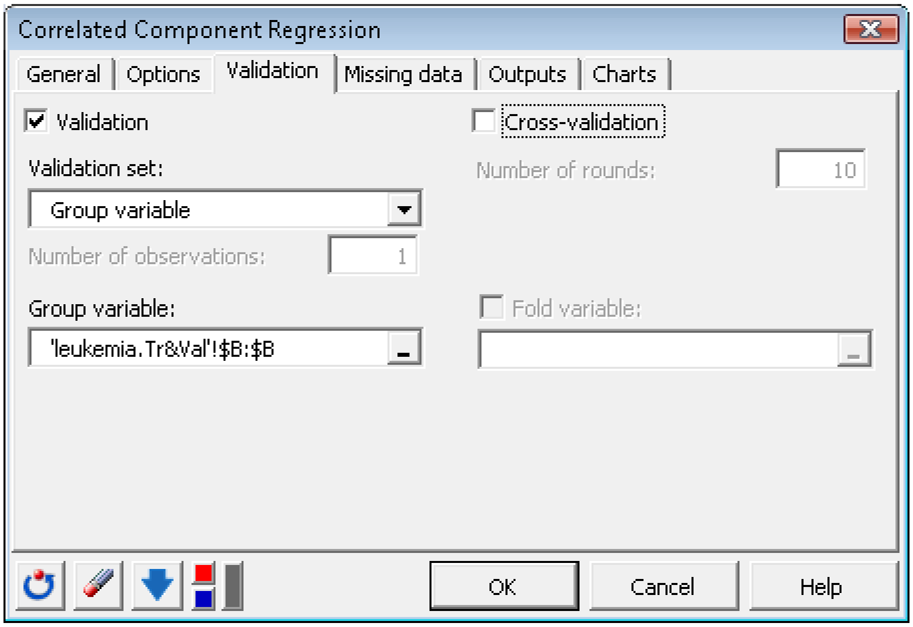
□ Remove by percent

Percent: 1

Fig 5:

Turn off the Cross Validation Option:

- ➢ Uncheck the "Use Cross Validation" option in the "Validation" tab.



☑ Validation

Validation set: Group variable

Number of observations: 1

Group variable: 'leukemia.Tr&Val'!$B:$B

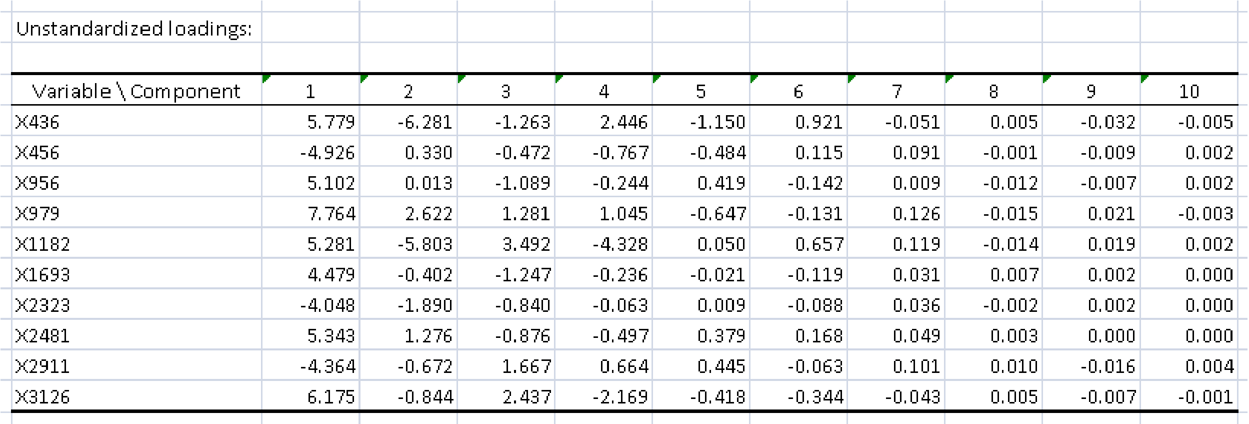
□ Cross-validation

Number of rounds: 10

□ Fold variable:

Fig. 6:

Estimate the Model



Unstandardized loadings:

Variable \ Component	1	2	3	4	5	6	7	8	9	10
X436	5.779	-6.281	-1.263	2.446	-1.150	0.921	-0.051	0.005	-0.032	-0.005
X456	-4.926	0.330	-0.472	-0.767	-0.484	0.115	0.091	-0.001	-0.009	0.002
X956	5.102	0.013	-1.089	-0.244	0.419	-0.142	0.009	-0.012	-0.007	0.002
X979	7.764	2.622	1.281	1.045	-0.647	-0.131	0.126	-0.015	0.021	-0.003
X1182	5.281	-5.803	3.492	-4.328	0.050	0.657	0.119	-0.014	0.019	0.002
X1693	4.479	-0.402	-1.247	-0.236	-0.021	-0.119	0.031	0.007	0.002	0.000
X2323	-4.048	-1.890	-0.840	-0.063	0.009	-0.088	0.036	-0.002	0.002	0.000
X2481	5.343	1.276	-0.876	-0.497	0.379	0.168	0.049	0.003	0.000	0.000
X2911	-4.364	-0.672	1.667	0.664	0.445	-0.063	0.101	0.010	-0.016	0.004
X3126	6.175	-0.844	2.437	-2.169	-0.418	-0.344	-0.043	0.005	-0.007	-0.001

Fig. 7:

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.