

# Correlated Component Regression (CCR)

Use this module to model and predict a designated dependent variable  $Y$  (continuous or dichotomous) from a set of  $P$  correlated explanatory variables (predictors)  $X = (X_1, X_2, \dots, X_P)$ .

In this section:

Description

Dialog box

Results

Example

References

**Copyright ©2011 Statistical Innovations Inc. All rights reserved.**

## Description

The four regression methods available in the Correlated Component Regression (CCR) module use fast cross-validation to determine the amount of regularization to 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 methods are based on Generalized Linear Models (GLM). As an option, the CCR step-down algorithm may be activated to exclude irrelevant  $X$ s.

The linear part of the model is a weighted average of  $K$  components  $S = (S_1, S_2, \dots, S_K)$  where each component itself is a linear combination of the predictors. For  $Y$  *dichotomous*, these methods provide an alternative to *Logistic* regression (CCR-Logistic) and *linear discriminant analysis* (CCR-LDA). For a *continuous*  $Y$ , these procedures provide an alternative to traditional *linear* regression methods, where components may be correlated (CCR-LM procedure), or restricted to be uncorrelated with components obtained by PLS regression techniques (CCR-PLS). Typically  $K < P$ , resulting in model regularization that reduces prediction error.

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 methods, 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 used to determine the amount of regularization  $K^*$  to apply, and the number of predictors  $P^*$  to include in the model when the step-down algorithm is utilized.

When the CCR step-down option is activated with  $M$ -fold cross-validation, output includes a table of predictor counts, reporting the number of times each predictor is included in a model estimated with one omitted fold. The counts 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 regression methods 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:

$$\hat{Y} = S(S'DS)^{-1}S'DY \quad (1.1)$$

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

For example, with K=2 components we have:

$$\hat{Y} = \alpha + b_{1,2}S_1 + b_{2,1}S_2 \quad (1.2)$$

where  $b_{1,2}$  and  $b_{2,1}$  are the component weights, the components defined as:

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

where  $\lambda_{g,1}$  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:

$$\hat{Y} = \alpha + \sum_{g=1}^P (b_{1,2}\lambda_{g,1} + b_{2,1}\lambda_{g,2}) X_g \quad (1.3)$$

where

$$\beta_g = b_{1,2}\lambda_{g,1} + b_{2,1}\lambda_{g,2} \quad (1.4)$$

is the (regularized) regression coefficient associated with predictor  $X_g$ .

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 or 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.

Methods 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.Ida)

Logistic regression is the standard regression (classification) approach for predicting 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.

$$\text{Logit}(Y | S) = \alpha + b_{1,2}S_1 + b_{2,1}S_2$$

which in reduced form yields:

$$\text{Logit}(Y | X) = \alpha + \sum_{g=1}^P (b_{1,2}\lambda_{g,1} + b_{2,1}\lambda_{g,2})X_g$$

Logit(Y), defined as the natural logarithm of the probability of being in dependent variable group 1 (say Y=1) divided by the probability of being in group 2 (say Y=0), 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:

$$\begin{aligned} \text{Prob}(Y = 1 | X) &= \exp(\text{Logit}(Y | X)) / (1 + \exp(\text{Logit}(Y | X))) & (1.5) \\ &= 1 / (1 + \exp(-\text{Logit}(Y | X))) \end{aligned}$$

and  $\text{Prob}(Y = 0 | X) = 1 / (1 + \exp(\text{Logit}(Y | X)))$  (1.6)

Thus, the logistic regression model is a model for predicting the probability of being in a particular group. Predictions are reported for group 1, which is defined as the category of Y associated with the higher of the 2 numeric values taken on by Y.

Linear Discriminant Analysis (LDA) is another model used commonly to obtain predicted probabilities for a dichotomous Y:

- CCR.Ida 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, PLS, CCR.Ida, or CCR.logistic, in the case where  $P < N$ , setting  $K = P$  yields the corresponding (saturated) regression models:

Method CCR.lm (or PLS) is equivalent to OLS regression (for  $K = P$ )

Method CCR.logistic yields traditional Logistic regression (for  $K = P$ )

Method CCR.lda yields traditional Linear Discriminant Analysis (for  $K = P$ )

where prior probabilities are computed from group sizes

## **CCR.LM**

The CCR.LM method applies CCR techniques to obtain a regularized linear regression based on the Correlated Component Regression (CCR) model for a continuous  $Y$  (Magidson, 2010; Magidson and Wassmann, 2010). It is recommended especially in cases where several explanatory variables have moderate to high correlation. Tutorial 1 (see [Example](#)) illustrates an analysis involving a relatively small number of correlated  $X$ s, and Tutorial 3 (see [Example](#)) shows how separate models can be estimated for different (latent class) segments using ‘Observation weights’.

Method CCR.lm differs from Method PLS in that the components are allowed to be correlated, there is no need to deflate (and then restore) predictors, and similar to traditional OLS regression, predictions are invariant to linear transformations applied to the predictors. Thus, the explanatory variables do not need to be standardized prior to estimation.

## **PLS**

The PLS method applies CCR techniques to obtain a regularized linear regression based on the PLS regression (PLS) model for a continuous  $Y$ . For an introduction to PLS regression see Tenenhaus (1998). For a comparison of the CCR.lm and PLS methods see Tenenhaus (2011).

Unlike CCR.lm which is invariant with respect to the scale of the predictors, when  $K < P$ , PLS regression can yield substantially different predictions depending upon whether the predictors are standardized or not. For a detailed comparison of CCR.lm, PLS with unstandardized  $X$ s and PLS with standardized  $X$ s, see Magidson (2011).

## **CCR.lda and CCR.logistic**

The CCR.lda and CCR.logistic methods apply CCR techniques to obtain regularized regressions based on the Correlated Component Regression (CCR) model for a dichotomous  $Y$ .

Tutorial 2 (see [Example](#)) illustrates such an analysis involving many  $X$ s, and where  $P > N$  (i.e., the number of predictors exceeds the number of cases).

## Notes:

### **M-fold Cross-Validation**

R rounds of M-fold Cross-validation (CV) may be used to determine the number of components  $K^*$  and number of predictors  $P^*$  to include in a model. For  $R > 1$  rounds, the standard error of the relevant CV statistic is also reported. When multiple records (rows) are associated with the same case ID (in XLSTAT, case IDs are specified using 'Observation labels'), for each round, the CV procedure assigns all records corresponding to the same case to the same fold.

### **The Automatic Option in M-fold Cross-Validation**

When the CV option is performed in Automatic mode (see 'Automatic' option in Options tab) a maximum number  $K$  is specified for the number of components, all  $K$  models containing between 1 and  $K$  components are estimated, and the  $K^*$  model selected as the one with the best CV statistic. When the step-down option is also activated, the  $K$  models are estimated with all predictors prior to beginning the step-down algorithm.

The CV statistic used to determine  $K^*$  depends upon the model type as follows:

For CCR.lm or PLS: The  $CV-R^2$  is the default statistic. Alternatively, the Normed Mean Squared Error (NMSE) can be used instead of  $CV-R^2$ .

For CCR.lda or CCR.logistic: The CV-Accuracy, based on the probability cut-point of .5, is used by default. In the case of two or more values of  $K$  yielding identical values for CV-Accuracy, the one with the higher value for the Area Under the ROC Curve (AUC) is selected.

### **Predictor Selection Using the CCR/Step-Down Algorithm**

In step 1 of the step-down option, a model containing all predictors is estimated with  $K^*$  components (where  $K^*$  is specified by the user or determined by the program if the Automatic option is activated), and the relevant CV statistics are computed. In step 2, the model is then re-estimated after excluding the predictor whose standardized coefficient is smallest in absolute value, and CV statistics are computed again. Note that both steps 1 and 2 are performed within each subsample formed by eliminating one of the folds. This process continues until the user-specified minimum number of predictors remains in the model (by default,  $P_{\min} = 1$ ). The number of predictors included in the reported model,  $P^*$ , is the one with the best CV statistic.

In any step of the algorithm, if the number of predictors remaining in the model falls below  $K^*$ , the number of components is automatically reduced by 1, so that the model remains saturated. For example, suppose that  $K^*=5$ , but after a certain number of predictors are eliminated  $P=4$  predictors remain. Then, the  $K^*$  is reduced to 4 and the step-down algorithm continues.


If a maximum number of predictors to be included in a model,  $P_{\max}$ , is specified, the step-down algorithm still begins with all predictors included in the model, but results are reported only for  $P$  less than or equal to  $P_{\max}$ , and the CV statistics are only examined for  $P$  in the range  $[P_{\min}, P_{\max}]$ .

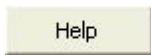
**Copyright ©2011 Statistical Innovations Inc. All rights reserved.**


## Dialog box


The dialog box is divided into several tabs that correspond to options ranging from the selection of data to the display of results. The description of the various elements of the dialog box appears below.



 : Click this button to start the computations.

 : Click this button to close the dialog box without doing any computation.

 : Click this button to display the help.

 : Click this button to reload the default options.

 : Click this button to delete the data selections.

  : Click these buttons to change the way XLSTAT handles the data. If the arrow points down, XLSTAT considers that rows correspond to observations and columns to variables. If the arrow points to the right, XLSTAT considers that rows correspond to variables and columns to observations.

**General** tab:

### **Y / Dependent variables:**

**Quantitative:** Select the dependent variable(s). The data must be numeric. If the ‘Variable labels’ option is activated make sure that the headers of the variable(s) have also been selected. If more than one Y variable is selected, models will be developed for each Y separately.

### **X / Explanatory variables:**

**Quantitative:** Select one or more quantitative explanatory variables. The data must be numeric. If the ‘Variable labels’ option is activated make sure that the headers of the variables have also been selected.

**Method:** Choose the regression method you want to use:

- **CCR.Im:** Activate this option to compute a Correlated Component Linear Regression model with a continuous dependent variable. Predictors assumed to be numeric (continuous, dichotomous, or discrete).
- **PLS:** Activate this option to compute a Partial Least Squares Regression with a continuous dependent variable. Predictors assumed to be numeric (continuous, dichotomous, or discrete).
- **CCR.Ida:** Activate this option to compute a Correlated Component Regression with a dichotomous (binary) dependent variable Y. According to assumptions of Linear Discriminant Analysis (LDA), predictors assumed to be multivariate normal with differing means but constant variances and correlations within each dependent variable group).
- **CCR.logistic:** Activate this option to compute a Correlated Component Logistic Regression model with a dichotomous (binary) dependent variable. Predictors assumed to be numeric (continuous, dichotomous, or discrete).

**Range:** Activate this option if you want to display the results starting from a cell in an existing worksheet. Then select the corresponding cell.

**Sheet:** Activate this option to display the results in a new worksheet of the active workbook.

**Workbook:** Activate this option to display the options in a new workbook.

**Variable labels:** Activate this option if the first row of the data selections (dependent and explanatory variables, weights, observations labels) includes a header.

**Observation labels:** Activate this option if labels are available for the N observations. Then select the corresponding data. If the ‘Variable labels’ option is activated you need to include a header in the selection.

With repeated measures data (multiple records per case) the Observation labels variable serves as a case ID variable, which groups the records from a given case together so that they are assigned to the same fold during cross-validation. If this option is *not* activated, the observations labels are automatically generated by XLSTAT (Obs1, Obs2 ...), so that each case contains a single record.

**Observation weights:** Activate this option if you want to weight the observations. If you do not activate this option, all weights are set to 1. The weights must be non-negative values. Setting a case weight to 2 is equivalent to repeating the same observation twice. If the ‘Variable labels’ option is activated, make sure that the header (first row) has also been selected.

**Options** tab:

**Component options:**

**Automatic:** When the ‘Automatic’ option is activated, XLSTAT-CCR estimates K-component models for all values of K less than or equal to the number specified in the ‘Number of components’ text box, and produces the ‘Cross-validation Component Plot’ (see Charts tab). This chart plots the CV- $R^2$  (or NMSE) if the CCR.lm or PLS method is activated, or for CV-ACC (accuracy) and CV-AUC (Area Under ROC Curve) if the CCR.logistic or CCR.lda method is activated. Coefficients are provided for the model with the best CV result.

Note: Activating the ‘Automatic’ option will have no effect if ‘Cross-validation’ option is not also activated.

**Number of components / Max Components:** When Automatic is activated, separate K-component models are estimated for each value  $K=1, 2, \dots, K_{MAX}$  where the number  $K_{MAX}$  is specified in the ‘Max Components’ field. If Automatic is not activated, enter the desired number of components K (positive integer) in the ‘Number of Components’ field. If the number entered exceeds the number of selected predictors P or N-1, K will automatically be reduced to the minimum of P and N-1.

**Step-Down options:**

**Perform Step Down:** Activate this option to estimate a  $K^*$ -component model containing the subset of candidate predictors selected according to the chosen option settings:

**Min variables:** Enter the minimum number of predictors to be included in the model. The default value is 1.

**Max variables:** Enter the maximum number of predictors to be included in the model. The default value is 20.

**Remove by Percent:** Activate this option to specify the percentage of predictors to be removed at each step. If not activated, the step-down algorithm removes 1 predictor at a time, which might take a considerable amount of time to run when the number of predictors is large.

**Percent:** Enter 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.

Note: If the ‘Automatic’ option is also activated,  $K^*$  is the value for  $K$  having the best cross-validation (CV) statistic. Otherwise,  $K^* =$  the number entered in the ‘Number of Components’ field.

#### **Additional Options for CCR.logistic method**

The following additional options apply to the Iteratively Re-weighted Least Squares (IRLS) algorithm that is used repeatedly to estimate parameters for the CCR.logistic model.

**Iterations:** Enter the number of iterations for IRLS. The default (recommended) number is 4.

**Ridge:** Enter the Ridge penalty number for CCR.logistic models. The default number is 0.001. 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.

#### **Additional Options for CCR.lm and PLS methods:**

**NMSE:** Activate this option to use Normed Mean Squared Error (NMSE) as an alternative to the default criterion,  $CV-R^2$ , for determining the tuning parameters  $K^*$  (if the ‘Automatic’ option is activated) and/or the number of predictors to be included in the model,  $P^*$ , if the ‘Perform Step-down’ option is activated.

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). If the NMSE option is not activated, the default criterion CV-  $R^2$  will be used. 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.

#### **Additional Options for PLS method:**

**Standardize:** Activated by default, this option standardizes the explanatory variables to have variance 1. Unlike the other methods which are invariant with respect to linear transformations on the variables, the PLS regression method produces different results depending upon whether or not the explanatory variables are standardized. Deactivate this option to use the PLS method with unstandardized predictors.

#### **Validation** tab:

**Validation:** Activate this option to define a validation subsample that will be used only to provide independent validation statistics for a model. Cases selected for a validation subsample are used neither for model estimation nor for cross-validation. The model developed by XLSTAT-CCR from the estimation sample is used to score the cases in the validation subsample, and validation statistics are produced for this subsample. If not activated, the model is developed based on the entire estimation (training) sample.

**Validation set:** Choose one of the following options to define how to obtain the subsample of observations used to define the validation subsample:

- **Random:** The validation observations are randomly selected. The ‘Number of observations’ to be selected must then be specified.
- **N last rows:** The N last observations are selected for the validation subsample. The ‘Number of observations’ N must then be specified.
- **N first rows:** The N first observations are selected for the validation. The ‘Number of observations’ N must then be specified.
- **Group variable:** If you choose this option, you need to select a binary variable with only 0s and 1s. The 1s identify the observations to use for the validation.

#### **Cross-Validation Options:**

**Cross-Validation:** Activate this option to use cross-validation.

**Number of Rounds:** The default number is 1. Enter the number of rounds (positive integer) of cross-validation to be performed. When a value greater than 1 is entered, the standard error for the relevant CV statistic is calculated. This option does not apply when a Fold variable is specified.

**Number of Folds:** The default number is 10. Enter the number of cross-validation folds (positive integer greater than 1). Typically, a value between 5-10 is specified that divides evenly (when possible) into the number of observations in the estimation sample. This option does not apply when a Fold variable is specified.

**Stratify:** Activate this option to use the 2 categories of dependent variable Y as a stratifier for fold assignment (applies only to CCR.Ida and CCR.logistic).

**Fold Variable:** Activate this option to use a variable to specify to which fold each case is assigned. If no grouping variable is specified, each case is assigned to 1 of the folds M folds randomly. A fold variable contains positive integer values 1, 2, ..., M where  $M = \# \text{ folds}$ .

NOTE: When Observation labels are specified with the same label for multiple records, all records with the same observation label are grouped together and assigned to the same fold. This assures that in the case of repeated measures data (multiple records per case) the records associated with a given case are all allocated to the same fold during cross-validation.

**Missing data** tab:

**Remove the observations:** Activate this option to remove the observations with missing data (list-wise deletion).

**Include Missing:** Activate this option to impute the mean for missing values. Imputation is done for the estimation and the validation samples separately.

**Outputs** tab:

Following model estimation, standard model output is generated. Output tabs 1, 2, and 3 allow for the following additional output to be generated.

Tab 1

**Descriptive Statistics:** Activate this option to display descriptive statistics for the variables selected.

**Correlations:** Activate this option to display the correlation matrix for the dependent and the explanatory variables.

**Predictors retained in the model:** Activate this option to display the predictors retained in the model.

**Coefficients:**

**Unstandardized:** Activate this option to display the unstandardized regression coefficients.

**Standardized Coefficients:** Activate this option to display the standardized regression coefficients.

**Predictions and residuals:** Activate this option to display the predictions and residuals associated with the dependent variable. For methods CCR.lm and PLS, predictions

For model types CCR.lm and CCR.PLS, predictions are the probability of being in the dependent variable group coded with the higher value.

**Equation of the model:** Activate this option to display the equation for the model.

For model types CCR.lm and PLS, the equation predicts the mean of the dependent variable for given values of the predictors. For model types CCR.lm and CCR.logistic, the equation predicts the probability of being in dependent variable group 1 (group 1 is the group that is coded with the higher value).

Tab 2

The following parameters can be included in the output by activating the associated output options.

**Component Weights:**

**Unstandardized:**

**Standardized:**

**Loadings:**

**Unstandardized:**

**Standardized:**

**Cross-validation predictor count table:** Activate this option to display the predictor count table. This option can only be activated if ‘Step-down’ option is activated in the Options tab *and* the ‘Cross-Validation’ option is activated in the Validation tab.

**Cross-Validated Step-Down table:** Activate this option to display the table corresponding to cross-validation step-down. This option can only be activated if the ‘Step-down’ option is activated in the Options tab *and* the ‘Cross-Validation’ option is activated in the Validation tab.

Tab 3 (available only for model types CCR.Ida and CCR.logistic)

**Classification table:** Activate this option to display the posterior observation classification table (confusion table) using a specified probability cutpoint (default probability cutpoint = 0.5).

**Charts** tab:

**Cross-Validation Component Plot:** Activate this option to display the chart produced when both the **Automatic** option and **Cross-validation** are activated. This chart plots the relevant CV statistic as a function of the number of components  $K=1, 2, \dots, K_{MAX}$ .

For model types CCR.Ida and CCR.logistic:

The Cross-Validation Component Plot corresponds to the cross-validation AUC and model accuracy (ACC) based on the number of components  $K$  ranging from 1 to the specified **Number of components**.

For model types CCR.lm and PLS:

The  $R^2$  plot corresponds to the cross-validation  $R^2$  (or NMSE if this option is activated in the Options tab) based on the number of components  $K$  ranging from 1 to the specified **Number of components**.

**Cross-Validation Step-down Plot:** Activate this option to display the chart associated with the **Step-down** option and **Cross-validation**.

For CCR.lda and CCR.logistic options:

The Cross-Validation Step-down Plot corresponds to the cross-validation AUC and model accuracy based on the specified K-component model for numbers of predictors P ranging from the specified 'Max variables' down to the specified 'Min variables'.

For CCR.lm and PLS:

The  $R^2$  graph corresponds to the cross-validation  $R^2$  (or NMSE if this option is activated in the Options tab) based on the specified K-component model for numbers of predictors P ranging from the specified 'Max variables' down to the specified 'Min variables'.

**Copyright ©2011 Statistical Innovations Inc. All rights reserved.**

# Results

**Summary (descriptive) statistics:** the tables of descriptive statistics display for all the selected variables a set of basic statistics. For the dependent variables (colored in blue), and the quantitative explanatory variables, XLSTAT displays the number of observations, the number of observations with missing data, the number of observations with no missing data, the mean, and the unbiased standard deviation.

**Correlation Matrix:** this table is displayed to allow your visualizing the correlations among the explanatory variables, among the dependent variables and between both groups.

## Goodness of Fit Statistics:

For model types CCR.lm and PLS:

The table displays the model quality indices.

- The  $R^2$  is shown for the estimation sample. If a validation is specified, the Validation- $R^2$  will be included in the table. If the cross-validation option is activated, the CV- $R^2$  will be included in the table. The CV-  $R^2$  reported in the table 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-  $R^2(P^*.r)$ .
- If the NMSE option is activated, the normed mean squared error (NMSE) is reported in addition to  $R^2$ . 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 estimation sample.

For model types CCR.lda and CCR.logistic:

The table displays the model quality indices.

- The Area Under the Curve (AUC) is shown for the estimation sample. If a validation is specified, the Validation-AUC will be included in the table. If the cross-validation option is activated, the CV-AUC will be included in the table.
- The accuracy (ACC) is shown for the estimation sample. If a validation is specified, the Validation-ACC will be included in the table. If the cross-validation option is activated, the CV-ACC will be included in the table.

**Predictors retained in the model:** A list of the names of the predictors retained in the model.

**Number of components:** The number of components in the model.

**Unstandardized component Weights Table:** The unstandardized component weights for each component.

**Standardized component Weights Table:** The standardized component weights for each component.

**Unstandardized loadings Table:** The unstandardized predictor loadings for each component.

**Standardized loadings Table:** The standardized predictor loadings for each component.

**Cross-Validation Component Table (and associated plot):** This output appears only if the 'Automatic' option is activated in the Options tab *and* the 'Cross-Validation' option is activated in the Validation tab. If more than 1 round of M-folds are used, the relevant CV statistics are computed as the average over all Rounds, and the associated standard error is also reported. Coefficients and other output are provided for the model containing  $K^*$  components where  $K^*$  is the value of K shown in this table associated with the best CV statistic.

Results for model types CCR.lm and PLS:

The relevant CV statistic is the CV- $R^2$ . The NMSE statistic is also reported if requested in the Options tab.

Results for model types CCR.lda and CCR.logistic:

The relevant CV statistics are the Cross-Validated Accuracy (CV-ACC) and the CV-AUC.

**Cross-Validated Step-Down Table (and associated plot):** The Cross-Validation Step-Down Table appears only if the **Step-Down** option and the **Cross-Validation** option are activated. If more than 1 round of M-folds are used, the relevant CV statistics are computed as the average over all Rounds, and the associated standard error is also reported. Coefficients and other output are provided for the model containing  $P^*$  predictors where  $P^*$  is the value of P shown in this table associated with the best CV statistic.

Results for model types CCR.lm and PLS:

For each number of predictors in the model, the table reports the CV- $R^2$ . If more than 1 round of M-folds are used, the reported CV- $R^2$  is the average over all Rounds, and the associated standard error is also reported.

Results for model types CCR.lda and CCR.logistic:

For each number of predictors in the model, the table reports the CV-AUC (and associated standard error) and the CV-ACC.

Note: The value for the CV statistic provided in this table for  $P^*$  predictors, along with the associated standard error, may differ from the CV statistic provided in the Goodness of Fit Table. For example, suppose that  $P^* = 4$  predictors and  $R = 10$  rounds of M-folds are used. Then the value of the CV statistic reported in this table is computed as the average over all 10 rounds of the corresponding CV statistic within each of the 10 rounds, where all CV statistics are based on  $P^*$  predictors. On the other hand, as mentioned above, the CV statistic (and associated standard error) reported in the Goodness of Fit Table is computed as the average across all 10 rounds where in each round  $r$  the CV statistic is used based on  $P^*_r$  predictors.

### **Cross-Validation Predictor Count Table:**

The Cross-Validation Step-Down Table is available only if the **Step-Down** option and the **Cross-Validation** option are activated.

In the table, the first column lists the number of times each candidate predictor showed up in the final model for each round. The last column (Total) reports the sum of counts for each round. The last row (Total) reports the sum of the totals for a given round ( $= M * P_r$ ).

**Optimal number of predictors for each round table:** Reports the optimal # of predictors selected in each round ( $P_r$ ).

### **Unstandardized Coefficients Table:**

Unstandardized regression coefficients are used to predict the dependent variable Y.

For CCR.lm and CCR.logistic, Y is dichotomous and predictions are for the probability of being in the dependent variable group associated with the higher of the 2 numeric values taken on by Y.

For PLS with the **Standardize** option activated in the Options tab, predictors are standardized by dividing by their standard deviation. The unstandardized regression coefficient reported is for the *standardized* predictor.

**Equation of the model:** This table displays the equation for the model.

For model types CCR.lm and PLS, the equations compute the predicted value for the dependent variable, while for model types CCR.lm and CCR.logistic the equation computes the conditional mean of the dependent variable, while for model types CCR.lm and CCR.logistic, the equation computes the predicted probability of the dependent variable group coded with the highest value.

### **Standardized Coefficients Table (and associated column chart):**

Standardized regression coefficients are used to assess the importance of the predictors, predictors with the highest magnitude are the most important. Each standardized regression coefficient equals the corresponding unstandardized coefficient multiplied by the ratio  $\text{std}(X_g)/\text{std}(Y)$ , where 'std' denotes standard deviation.

For PLS with the **Standardize** option activated in the Options tab, predictors are standardized by dividing by their standard deviation, so that  $\text{std}(X_g) = 1$  for each predictor  $g = 1, 2, \dots, P$ . The

standardized regression coefficient in this case equals the corresponding unstandardized coefficient reported divided by  $\text{std}(Y)$ .

**Predictions and Residuals:**

This table reports the predictions for the dependent variable, residuals and standardized residuals.

Additional output for model types CCR.lda and CCR.logistic:

**Classification table for the estimation sample (and associated ROC Curve):** The table reports the correct classification rates for each of the 2 dependent variable groups. This classification table is based on the cutpoint specified in Output Tab 3 (default probability = .5).

**Classification table for the validation sample (and associated ROC Curve):** The table reports the correct classification rates for each of the 2 dependent variable groups. This classification table is based on the cutpoint specified in Output Tab 3 (default probability = .5).

**Copyright ©2011 Statistical Innovations Inc. All rights reserved.**

## Examples

The following tutorials on how to use XLSTAT-CCR are available:

[Tutorial 1](#): Getting Started with Correlated Component Regression (CCR) in XLSTAT-CCR

[Tutorial 2](#): Using Correlated Component Regression with a Dichotomous Y and Many Correlated Predictors

[Tutorial 3](#): Obtaining Predictions from a 2-class Regression

**Copyright ©2011 Statistical Innovations Inc. All rights reserved.**

## References

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

[Magidson, J. \(2011\). 'Correlated Component Regression: A Sparse Alternative to PLS Regression'. 5th ESSEC-SUPELEC Statistical Workshop on 'PLS \(Partial Least Squares\) Developments'.](#)

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

Tenenhaus, M. La regression PLS: Theorie et pratique (French Edition). Editions Technip (1998).

Tenenhaus, M. (2011). 'Conjoint use of Correlated Component Regression (CCR), PLS regression and multiple regression'. 5th ESSEC-SUPELEC Statistical Workshop on 'PLS (Partial Least Squares) Developments'.

**Copyright ©2011 Statistical Innovations Inc. All rights reserved.**