Tutorial 3.

Obtaining Predictions from a 2-class Regression

This tutorial[1] illustrates a reanalysis of data analyzed by Tenenhaus, et al. (2005): Tenenhaus, M., Pagès, J., Ambroisine L. and & Guinot, C. (2005); PLS methodology for studying relationships between hedonic judgments and product characteristics; Food Quality and Preference. 16, 4, pp 315-325.

The data consists of liking ratings on each of 6 different orange juice (OJ) products by 96 judges. Each of the 6 juices is also described by 16 physico-chemical attributes. In addition, the data contains classification information for weighting the judges according to their (posterior membership) probability of being in two different segments which have distinctly different OJ preferences. (click here for details of the random intercept latent class (LC) regression analysis used to obtain these posterior membership probabilities).

An Excel sheet containing both the data and the results for use in this tutorial can be downloaded by clicking here.

L21 | fx | 22.92

	A	B	C	D	E	F	G	H	I	J	K
1	seqID	ID	Orangejuice	rating	rating_mean	clu#	POSTERIOR.1	POSTERIOR.2	CFactor1	Glucose	Fructose
2	1	1	fruvita fr.	3	2.666667	1	0.980429	0.019571	-0.21392	23.65	25.65
3	2	1	joker amb.	2	2.666667	1	0.980429	0.019571	-0.21392	32.42	34.54
4	3	1	pampryl amb.	2	2.666667	1	0.980429	0.019571	-0.21392	25.32	27.36
5	4	1	pampryl fr.	3	2.666667	1	0.980429	0.019571	-0.21392	27.16	29.48
6	5	1	tropicana amb.	2	2.666667	1	0.980429	0.019571	-0.21392	17.33	20
7	6	1	tropicana fr.	4	2.666667	1	0.980429	0.019571	-0.21392	22.7	25.32
8	7	2	fruvita fr.	3	2.333333	1	0.997903	0.002097	-0.690251	23.65	25.65
9	8	2	joker amb.	2	2.333333	1	0.997903	0.002097	-0.690251	32.42	34.54
10	9	2	pampryl amb.	1	2.333333	1	0.997903	0.002097	-0.690251	25.32	27.36
11	10	2	pampryl fr.	1	2.333333	1	0.997903	0.002097	-0.690251	27.16	29.48
12	11	2	tropicana amb.	3	2.333333	1	0.997903	0.002097	-0.690251	17.33	20
13	12	2	tropicana fr.	4	2.333333	1	0.997903	0.002097	-0.690251	22.7	25.32
14	13	3	fruvita fr.	4	2.5	1	0.995635	0.004365	-0.450104	23.65	25.65

[1] To reproduce the results shown in this tutorial exactly, you will need to fix the seed to '123456789'. To fix the seed in XLSTAT, go to Options, then click on the Advanced tab. Check the box to activate the option 'Fix the seed to:', and change the seed to 123456789.

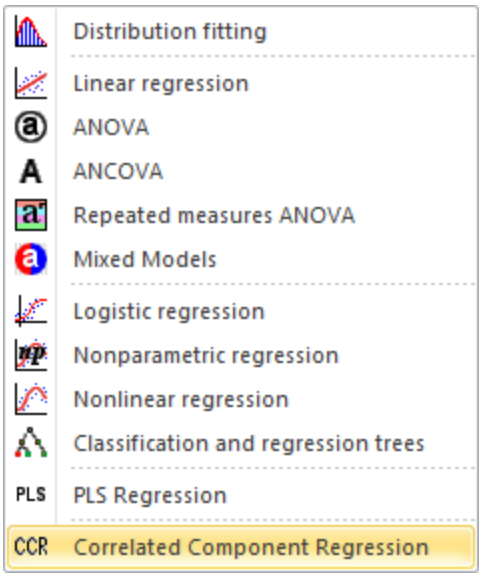
Goal of the Correlated Component Regression in this example

When data consists of multiple records per case, traditional (1-class) regression methods suffer from violation of the independent observations assumption which yields suboptimal prediction, since residuals from records associated with the same case will typically be correlated. In this tutorial we show how CCR can improve prediction of the liking ratings from the OJ attributes by allowing differing attribute effects for each of the 2 LC segments which show different OJ preferences.

In particular, this tutorial illustrates the second step in a 2-step process. In step 1, a 2-class regression model is developed based solely on dummy variables associated with the OJ products. In step 2, CCR is used to predict ratings based on the 16 product descriptors (rather than the dummy variables) to determine those that are the most important in predicting OJ liking. We develop separate models for each LC segment, obtained in step 1, and then combine models for both segments to obtain a single best set of predicted ratings. Use of this 2-step, 2-class regression analysis provides substantial improvement over the traditional regression (cross-validated R-square increases from .28 to .48).

Setting up a Correlated Component Regression (CCR) model

To activate the Correlated Component Regression dialog box, start XLSTAT by clicking on the  button in the Excel toolbar, and 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 selected CCR, the **Correlated Component Regression** dialog box is displayed.

To setup the CCR runs, in the **Y / Dependent variable(s)** field, select (see the tutorial on Selecting data for more information on this topic) with the mouse Column D (*rating*), containing the ratings for each of the 6 juices given by the judges (6 rows for each of the 96 judges). The ratings are the "Ys" of the model as we want to explain these ratings given by the judges as a function of the juice attributes.

In the **X / Predictors** field, select columns I through Y corresponding to the variable *CFactor1* (column I) plus the 16 juice attributes. *CFactor1*, a random intercept obtained from the LC regression analysis based solely on the OJ ratings, is highly correlated with the variable *Rating_mean* (column E), representing each judges' mean rating across all 6 juices. Its inclusion as a predictor serves a function similar to 'centering'.

The case **ID** variable (column B) is entered in the **Observation labels** field, so that all 6 records for each judge are grouped and assigned to the same fold during cross-validation.

Separate models will be developed for each segment. For Segment #1, we select the probability of being in that segment (**Posterior1**) as the weight variable (column G). (For theory on the use of posterior membership probabilities as weights see Magidson and Vermunt, 2005.)

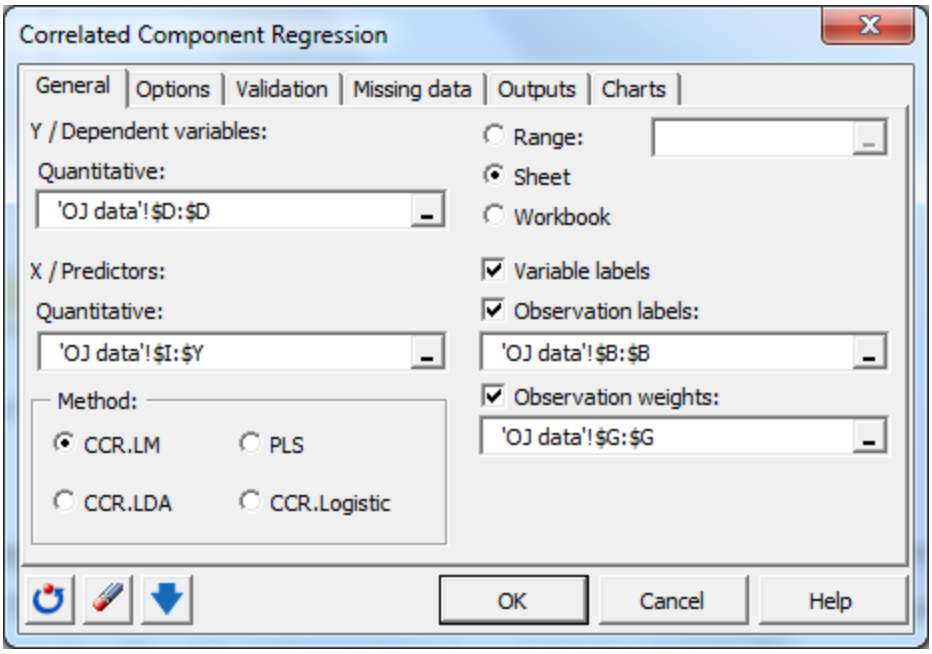


Figure 1. General Tab

To determine the number of components, in the **Options** tab of the dialog box, activate the 'Automatic' option and enter '17' in the 'Max components' field. To determine the number of predictors, activate the step-down procedure. Note that the Cross-validation option in the **Validation** tab is automatically activated with the default parameters (1 round of 10-folds).

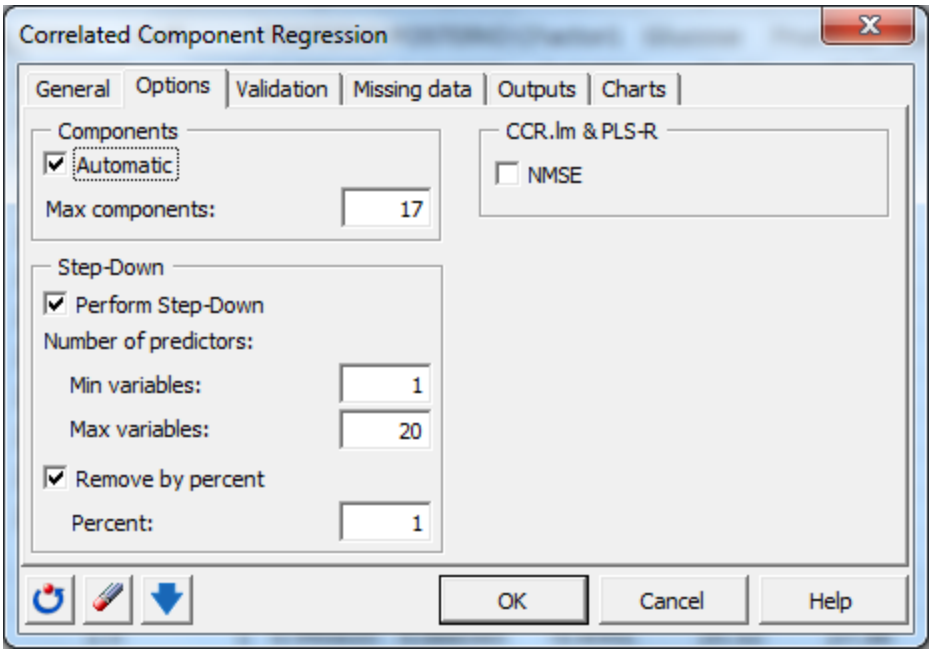


Figure 2. Options Tab

The fast computations start when you click on **OK**.

Interpreting CCR results for Segment #1

From the ***correlation matrix*** output it can be seen that the correlation between **rating** and **Acidity** equals -.433, suggesting that Segment #1 judges tend to dislike OJs with a high acidic nature. We will see later that in contrast to Segment #1 judges, Segment #2 judges tend to *prefer* OJs that have a high acid content (correlation = .252).

From the CV components table and associated plot we see that the maximum CV-R^2 = .398 occurs with K = 5 components (note also that the CV-R^2 deteriorates rapidly after K=9 components indicating substantial amount of collinearity for K>9).

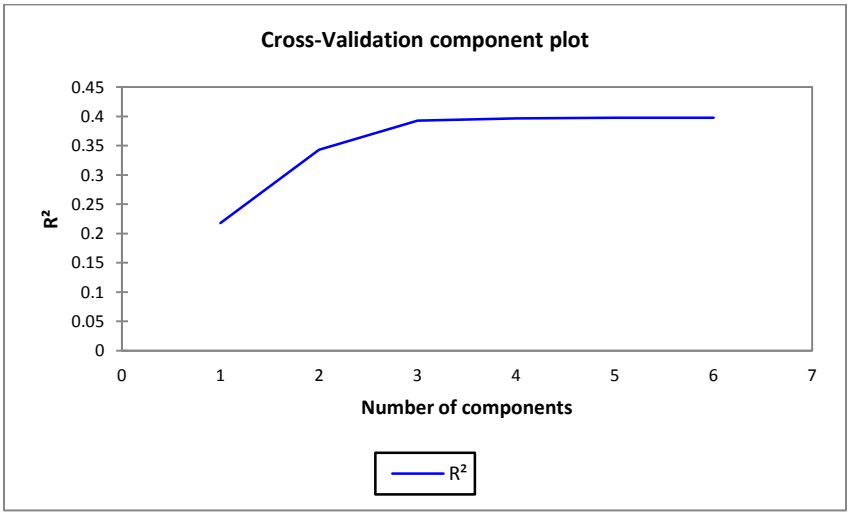


Figure 3. Cross-validation component plot (Segment #1)

From the CV-step down plot we see that the maximum CV-R^2 = .402 occurs with P*=4 predictors (and since P*<5, K is reduced automatically from 5 to K*=4 components to achieve an identifiable model).

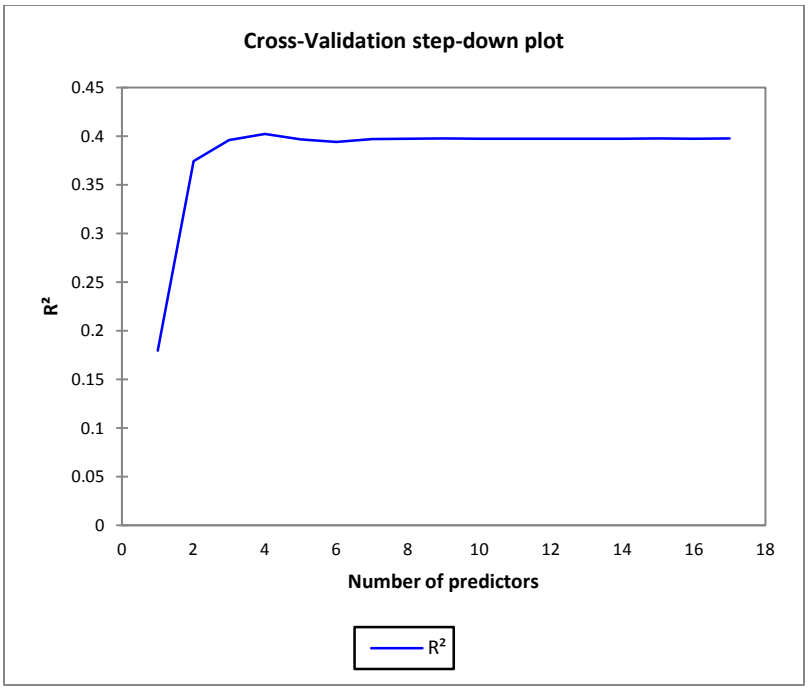


Figure 4. Cross-validation step-down plot (Segment #1)

Table 1 shows that **Acidity** is an important predictor in the model. The negative standardized coefficient (-.325) supports the inference that Segment #1 judges tend to dislike OJs with high acid content.

Standardized coefficients:

Variable	Coefficient
CFactor1	0.425
Fructose	-0.128
Sweeteningpower	0.238
Acidity	-0.325

Table 1. Standardized coefficients based on the 4-component model for Segment #1

For comparison, we will obtain results for Segment #2 next.

Developing the Corresponding CCR Model for Segment #2

Re-open the CCR dialog box by click on **Modeling data / Correlated Component Regression** command in the Excel menu or click the corresponding button on the **Modeling data** toolbar.

The previous model specifications are currently displayed. In the **General** tab, replace the current observation weights by the corresponding values (column H) associated with Segment #2 (**Posterior2**). To produce the Segment #2 model output on the same sheet as the Segment #1 model output, change the output option from 'Sheet' to 'Range' and select cell V1 in the 'CCR.LM' tab (the tab which contains the output from our previous model estimation).

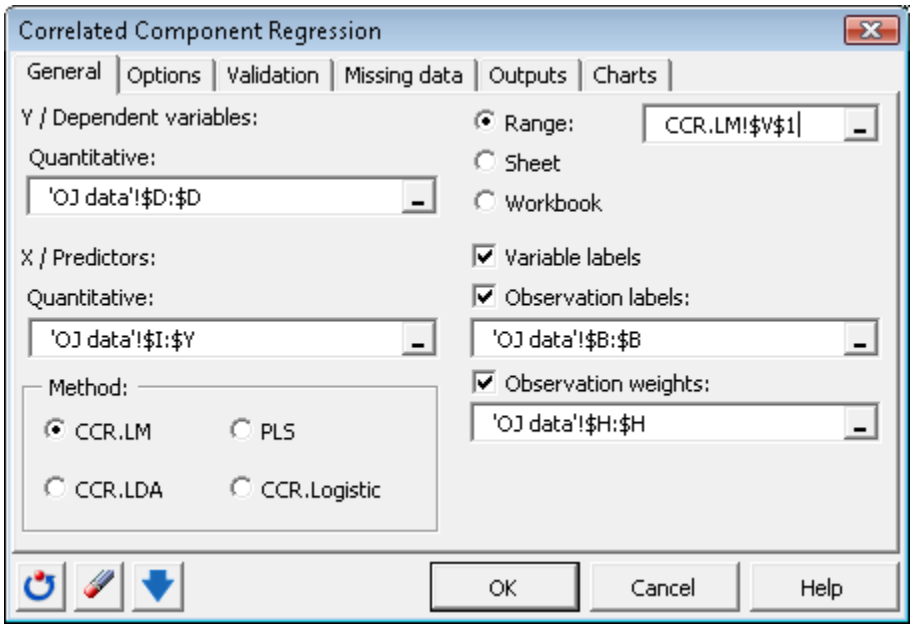


Figure 5. General Tab

Click on **OK** to estimate.

The relevant output for Segment #2 is shown below.

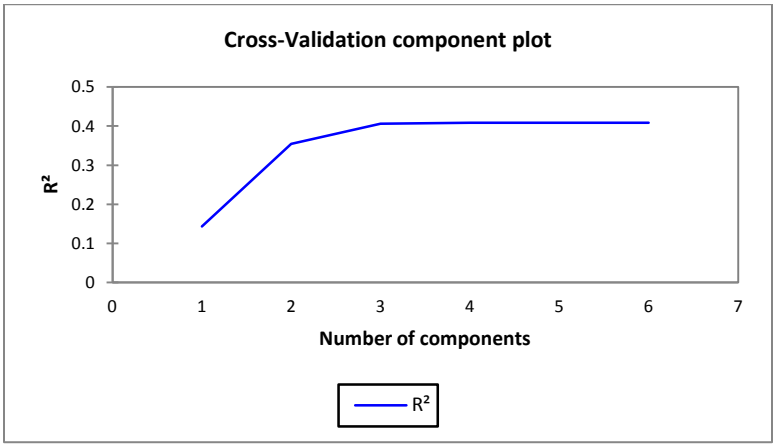


Figure 6. Cross-validation component plot (Segment #2). CV-R^2 = .409

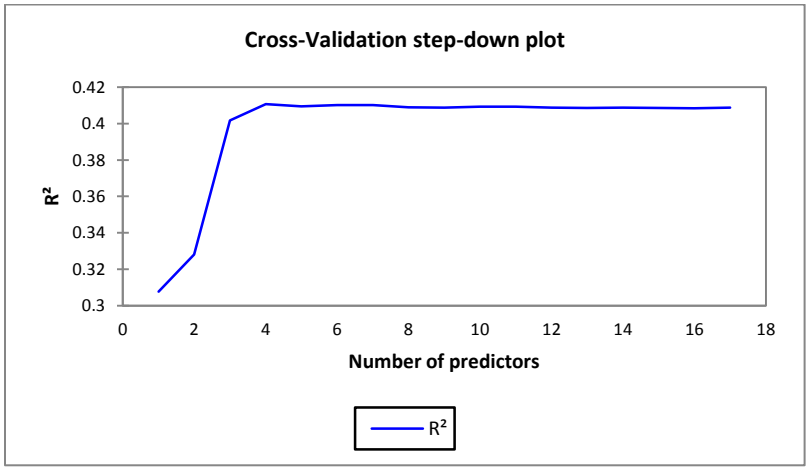


Figure 7. Cross-validation step-down plot (Segment #2). CV-R^2 = .411

Table 2 shows that Acidity is an important predictor for Segment #2 as well as Segment #1. However, in contrast to the model result for Segment #1, the standardized coefficient for Acidity is now positive. Table 2 show that Segment #2 judges *prefer* juices with higher acidity (.214), low sweetening power (-.169), and low smell intensity (-.129).

Standardized coefficients:

Variable	Coefficient
CFactor1	0.555
Sweeteningpower	-0.169
Smellintensity	-0.129
Acidity	0.214

Table 2. Standardized coefficients for Segment #2

Obtaining Predictions from the 2-class Model

Improved prediction over the 1-class model is due to the value of the additional information provided by the LC segmentation results. If we knew that a judge was from Segment #1 (i.e., preferred OJs that had lower acidity), we would use the Segment #1 model for prediction. Similarly, if we knew that a judge was from Segment #2 (i.e., preferred OJs that had higher acidity), we would use the Segment #2 model for prediction. While we do not know with certainty to which segment each judge belongs, we have the posterior membership probabilities to use as weights.

Our prediction from the 2-class CCR model is a weighted average of the 2 sets of predictions obtained from the 2 models. For example, our prediction for the rating for OJ#1 (fruvita fr.) given by judge #1 is obtained as a weighted average of the corresponding predictions from the 2 models, where the weights are the posterior membership probabilities:

Prediction = .98(3.441) + .02(2.373) = 3.42

For judge #1, the probability of being in Segment #1 is about .98, and thus the probability of being in Segment #2 is about .02. The predicted rating from the Segment #1 model (3.441) is weighted more heavily for this judge than that from the Segment #2 model (2.373), resulting in a prediction of 3.42 based on the 2-class regression model.

For illustrative purposes, these and other calculations are provided in sheet 'CCR.LM' (highlighted in yellow). These yellow highlighted cells were added manually to the output provided by XLSTAT-CCR. For example, cell L237 provides the formula for computing the predicted value 3.42 from the corresponding Segment #1 and Segment #2 output.

Predictions and residuals:

Observation	Weight	rating	Pred(rating)	Residual	Std. residual	ResidSq
1	0.980429	3.000	3.441	-0.441	-0.552	0.194
1	0.980429	2.000	2.154	-0.154	-0.193	0.024

Table 3A. Predictions and residuals output for model with **Posterior1** weights (first 2 rows)

Predictions and residuals:

Observation	Weight	rating	Pred(rating)	Residual	Std. residual	ResidSq
1	0.019571	3.000	2.373	0.627	0.804	0.393136
1	0.019571	2.000	3.113	-1.113	-1.428	1.239502

Table 3B. Predictions and residuals output for model with **Posterior2** weights (first 2 rows)

rating	Predictions	Residual	ResidSq
3	3.420	-0.420	0.176
2	2.173	-0.173	0.030

Table 3C. Predictions and residuals computed for 2-class regression model (first 2 rows)

Row 1 in Tables 3A, 3B and 3C corresponds to OJ#1 (fruvita fr.). Since this juice has a lower acidity level, Segment #1 judges are predicted to rate it higher than Segment #2 judges (3.441 vs. 2.373).

Note that judge #1 (corresponding to Observation = 1), tends to rate the 6 juices somewhat lower than the average judge (e.g., for Observation = 1, CFactor1 = -.214 and rating mean = 2.67). As mentioned above, the predictions provided by this 2-class model are substantially better than those provided by a 1-class model which ignores the segments. A food product manager might use these results to customize separate OJ products for each segment, based on the attributes used in each model.

References

Popper, R., J. Kroll, Jeff and J. Magidson (2004). Applications of latent class models to food product development: a case study. *Sawtooth Software Proceedings*, 2004.

Magidson, J., and Vermunt, J.K. (2006). Use of latent class regression models with a random intercept to remove overall response level effects in ratings data. In: A. Rizzi and M Vichi (eds.), Proceedings in Computational Statistics, 351-360, Heidelberg: Springer. (pdf)

Magidson, J., and Vermunt, J.K. (2005). An Extension of the CHAID Tree-based Segmentation Algorithm to Multiple Dependent Variables. C. Weihs & W. Gaul, Classification: The Ubiquitous Challenge, 176-183. Heidelberg: Springer. (pdf)

For CCR **Tutorial 1** click here: **Getting Started with Correlated Component Regression (CCR) in XLSTAT**

For CCR **Tutorial 2** click here: **Using Correlated Component Regression with a Dichotomous Y and Many Correlated Predictors**

For other XLSTAT Tutorials click here