Tutorial 4: Obtaining Predictions from a 2-class Regression in CORExpress®

Dataset for running CCR Linear Regression (CCR.lm)

This tutorial 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).

A SPSS (.sav) file of the dataset used in this tutorial can be downloaded by clicking here.

An Excel file (.xls) with the predictions and calculated residuals used in this tutorial can be downloaded by clicking here.

Goal of CCR for 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

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 'OJtutorial2lc.sav' and click Open to load the dataset

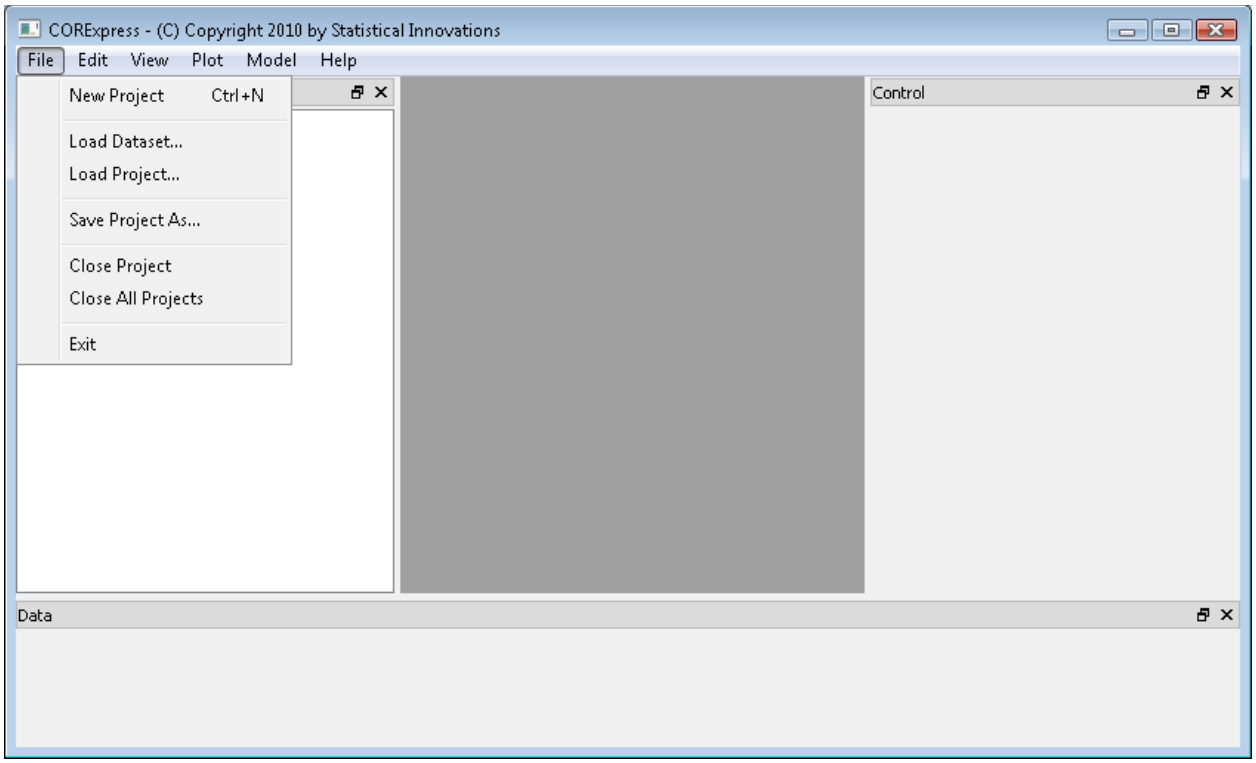


Fig. 1: File Menu

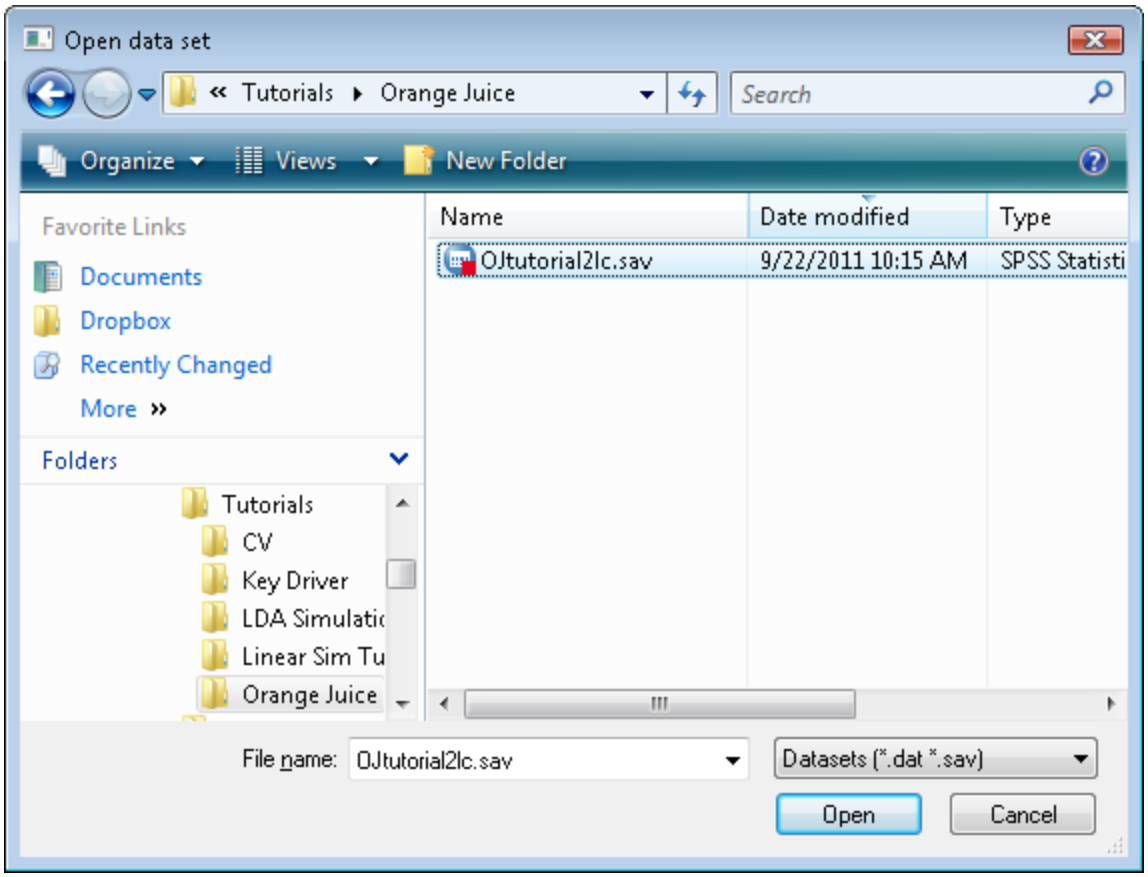


Fig. 2: Loading a Dataset

You will now see the 'OJtutorial2lc' 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.

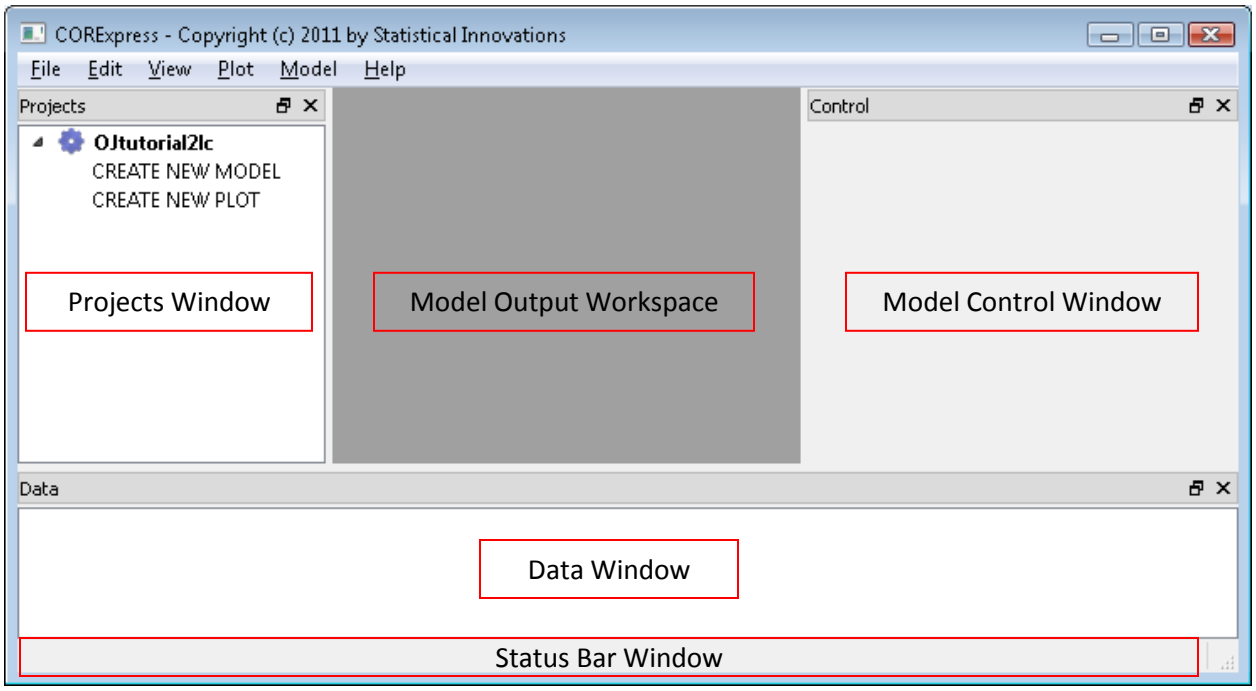
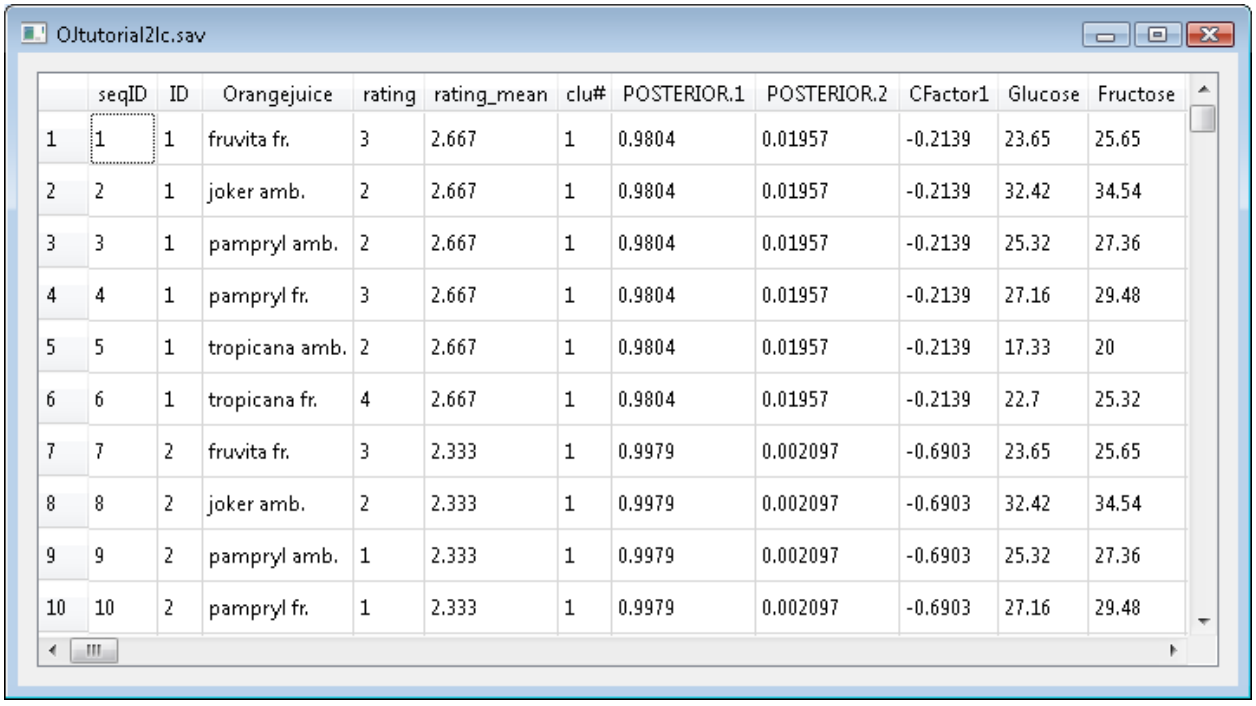


Fig. 3: COREXpress Windows

You can view the complete dataset in a new window by double clicking on 'OJtutorial2lc' 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).


OJtutorial2lc.sav

	seqID	ID	Orangejuice	rating	rating_mean	clu#	POSTERIOR.1	POSTERIOR.2	CFactor1	Glucose	Fructose
1	1	1	fruvita fr.	3	2.667	1	0.9804	0.01957	-0.2139	23.65	25.65
2	2	1	joker amb.	2	2.667	1	0.9804	0.01957	-0.2139	32.42	34.54
3	3	1	pampryl amb.	2	2.667	1	0.9804	0.01957	-0.2139	25.32	27.36
4	4	1	pampryl fr.	3	2.667	1	0.9804	0.01957	-0.2139	27.16	29.48
5	5	1	tropicana amb.	2	2.667	1	0.9804	0.01957	-0.2139	17.33	20
6	6	1	tropicana fr.	4	2.667	1	0.9804	0.01957	-0.2139	22.7	25.32
7	7	2	fruvita fr.	3	2.333	1	0.9979	0.002097	-0.6903	23.65	25.65
8	8	2	joker amb.	2	2.333	1	0.9979	0.002097	-0.6903	32.42	34.54
9	9	2	pampryl amb.	1	2.333	1	0.9979	0.002097	-0.6903	25.32	27.36
10	10	2	pampryl fr.	1	2.333	1	0.9979	0.002097	-0.6903	27.16	29.48

Fig. 4: COREXpress Dataset View

Step 1: Determining the Optimal Number of Components

Selecting the Type of Model:

- Double click on 'CREATE NEW MODEL' in the Projects window under 'OJtutorial2lc'

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 'rating' as the dependent variable.

Ratings are available 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.

Selecting the Predictors:

- In the Control window below 'Predictors', click and hold on 'CFactor1' and move the cursor down to 'Sweetness' to highlight all 17 predictors. Click on the box next to 'Sweetness' to select all 17 predictors.

CFactor1, a random intercept obtained from the LC regression analysis based solely on the OJ ratings, is highly correlated with the variable '*Rating_mean'*, representing each judges' mean rating across all 6 juices. Its inclusion as a predictor serves a function similar to 'centering'.

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 'CFactor1' and move the cursor down to 'Sweetness' to highlight all 17 predictors in the left box.
- Click on the '>>' box in the middle to select all 17 predictors and move them to the right box as candidate predictors.

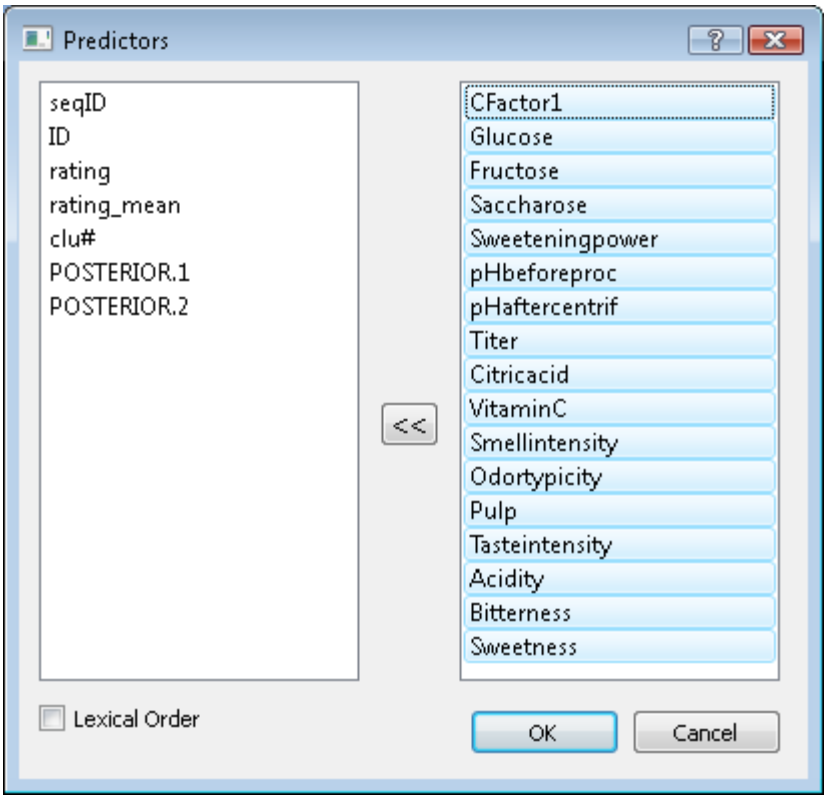


Fig. 5: Predictor Window

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, replace the default number with '16'

The estimation begins with all P=17 predictors in the model, and the CCR step-down procedure is applied, eliminating the weakest predictors until 1 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 1-16. The model output displayed will be for the model with P* predictors, where P* is the one in the range 1-16 that achieves the highest R^2 (CV-R^2).

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.

Your Control window should now look like this:

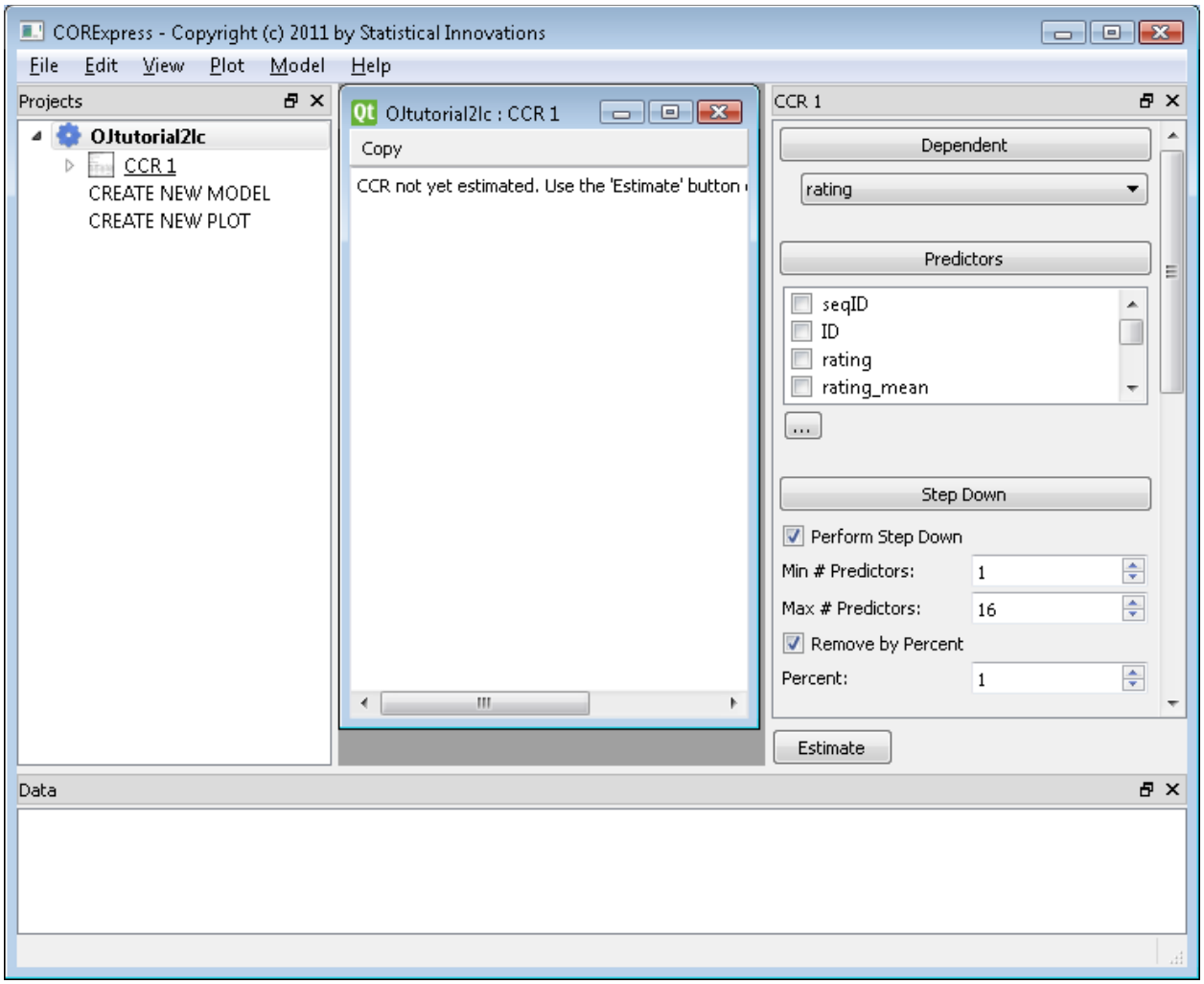


Fig. 6: Model Control Window

Selecting the Number of Components & Activating the Automatic Option:

- Under Options, click in the box to the right of '# Components', delete '4', and type '6'
- Then, check the 'Automatic' box

Selecting the Model Type:

- Click on 'CCR.lm' to select a CCR linear regression model

Selecting the Case Weights:

- Click on the Weights drop down menu and select 'POSTERIOR.1'.

Selecting the Case ID:

- Click on the Case ID drop down menu and select 'ID' to select the judge number as case ids.

The case **ID** variable (ID) is included so that all 6 records for each judge are grouped and assigned to the same fold during cross-validation.

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 '# rounds:' box, delete the default and type '10'
- In the '# Folds:' box, keep the default,'10'
- Uncheck the 'Stratify' option

This divides the analysis sample into 10 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. If a fold variable is not specified, COREXpress assigns cases randomly to each fold.

M-fold cross-validation is a common technique used in datamining. The CV-R^2 statistic is 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 statistic. Thus, the performance of the model is measured using cases not used at all in the development of the model.

Your Control window should now look like this:

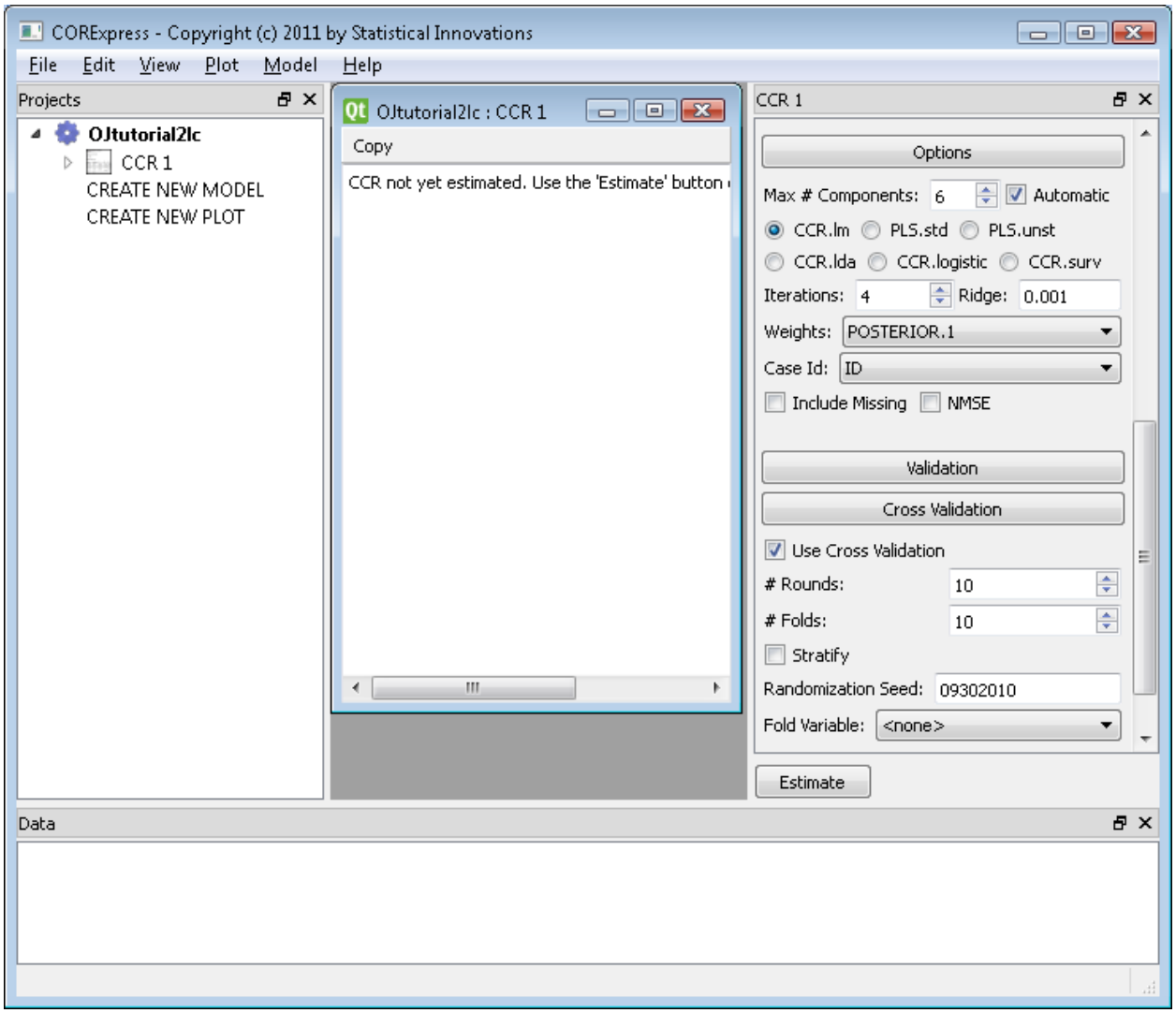


Fig. 7: Control Window

Estimate the Specified Model:

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

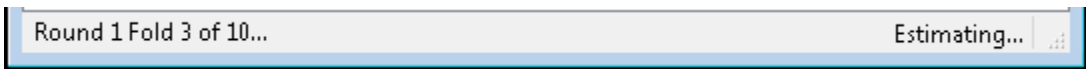
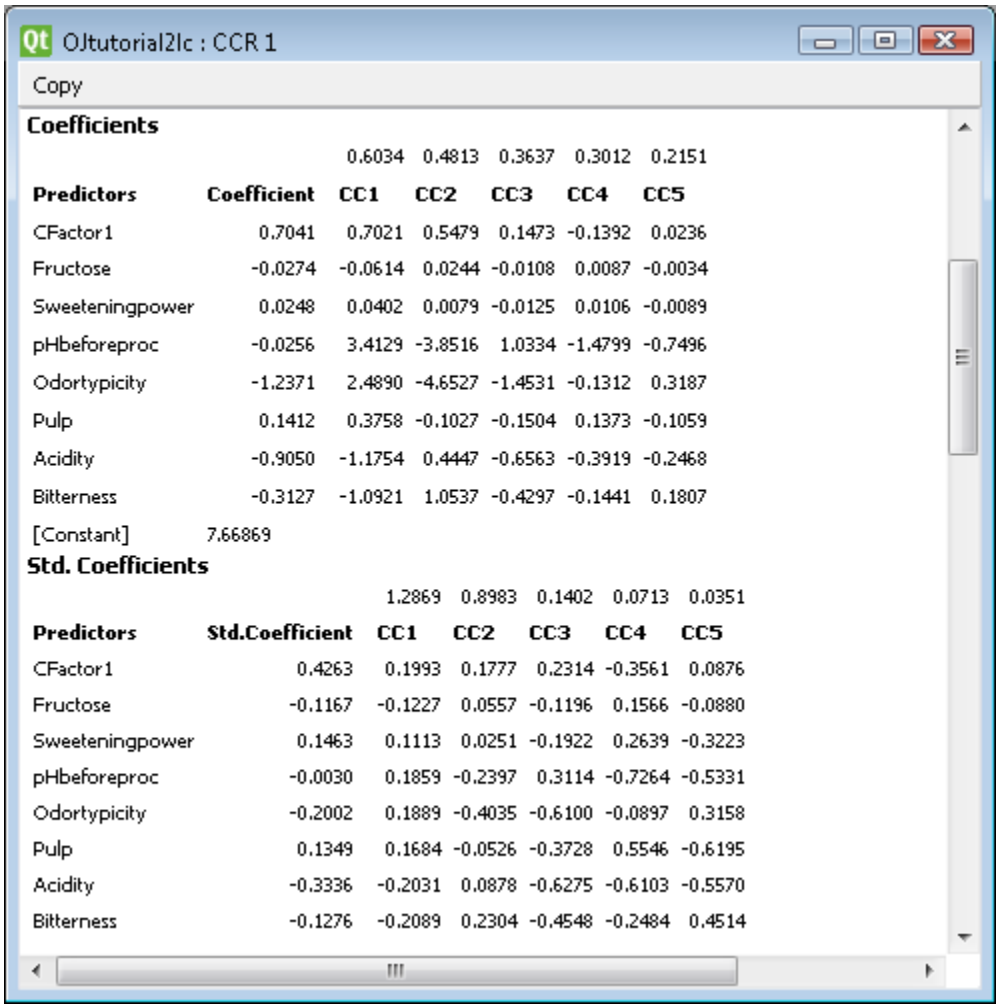


Fig. 8 COREXpress Status Bar

Interpreting CCR results for Segment #1

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=5 components (CC1, CC2, … , CC5) as well as the component weights for the components.


OJtutorial2lc : CCR 1

Copy

Coefficients

		0.6034	0.4813	0.3637	0.3012	0.2151
Predictors	**Coefficient**	**CC1**	**CC2**	**CC3**	**CC4**	**CC5**
CFactor1	0.7041	0.7021	0.5479	0.1473	-0.1392	0.0236
Fructose	-0.0274	-0.0614	0.0244	-0.0108	0.0087	-0.0034
Sweeteningpower	0.0248	0.0402	0.0079	-0.0125	0.0106	-0.0089
pHbeforeproc	-0.0256	3.4129	-3.8516	1.0334	-1.4799	-0.7496
Odortypicity	-1.2371	2.4890	-4.6527	-1.4531	-0.1312	0.3187
Pulp	0.1412	0.3758	-0.1027	-0.1504	0.1373	-0.1059
Acidity	-0.9050	-1.1754	0.4447	-0.6563	-0.3919	-0.2468
Bitterness	-0.3127	-1.0921	1.0537	-0.4297	-0.1441	0.1807
[Constant]	7.66869					

Std. Coefficients

		1.2869	0.8983	0.1402	0.0713	0.0351
Predictors	**Std.Coefficient**	**CC1**	**CC2**	**CC3**	**CC4**	**CC5**
CFactor1	0.4263	0.1993	0.1777	0.2314	-0.3561	0.0876
Fructose	-0.1167	-0.1227	0.0557	-0.1196	0.1566	-0.0880
Sweeteningpower	0.1463	0.1113	0.0251	-0.1922	0.2639	-0.3223
pHbeforeproc	-0.0030	0.1859	-0.2397	0.3114	-0.7264	-0.5331
Odortypicity	-0.2002	0.1889	-0.4035	-0.6100	-0.0897	0.3158
Pulp	0.1349	0.1684	-0.0526	-0.3728	0.5546	-0.6195
Acidity	-0.3336	-0.2031	0.0878	-0.6275	-0.6103	-0.5570
Bitterness	-0.1276	-0.2089	0.2304	-0.4548	-0.2484	0.4514

Fig. 9. Coefficient estimates obtained from the 6-component (saturated) CCR model

For Segment #1 judges, 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).

*This correlation is obtained using the Segment 1 posterior as a case weight variable.

Predictor	**All**	1	2	3	4	5	6	7	8	9	10
CFactor1	100	10	10	10	10	10	10	10	10	10	10
Acidity	99	10	9	10	10	10	10	10	10	10	10
Sweeteningpower	86	10	7	9	8	10	10	10	8	4	10
Pulp	61	10	2	1	2	10	10	10	2	6	8
Fructose	54	10	5	0	0	9	8	8	7	0	7
Odortypicity	54	10	2	0	0	10	10	10	2	0	10
Bitterness	53	10	3	0	0	10	10	10	0	0	10
pHbeforeproc	36	9	0	0	0	7	6	8	1	0	5
VitaminC	20	10	0	0	0	1	3	2	0	0	4
Glucose	13	10	0	0	0	0	1	1	0	0	1
Saccharose	13	10	1	0	0	0	0	0	0	0	2
Tasteintensity	11	8	0	0	0	2	0	0	0	0	1
Smellintensity	10	10	0	0	0	0	0	0	0	0	0
Sweetness	10	6	1	0	0	0	2	0	0	0	1
Titer	8	8	0	0	0	0	0	0	0	0	0
Citricacid	8	8	0	0	0	0	0	0	0	0	0
pHaftercentrif	4	1	0	0	0	1	0	1	0	0	1
Total	640	150	40	30	30	80	80	80	40	30	80
Predictors		**15**	**4**	**3**	**3**	**8**	**8**	**8**	**4**	**3**	**8**

Table 1. Predictor Table for Segment 1 Regression

From the CV components table we see that the maximum CV-R^2 = .396 occurs with K = 5 components.

Cross-Validated Components

# Components	R^2	Std. Dev
1	0.2196	0.0017
2	0.3414	0.0033
3	0.3922	0.0020
4	0.3954	0.0022
5	0.3963	0.0021
6	0.3962	0.0022

Table 2. Cross-validated Components Table

From the CV-step down plot and associated table we see that the maximum CV-R^2 = .3964 occurs with P*=8 predictors.

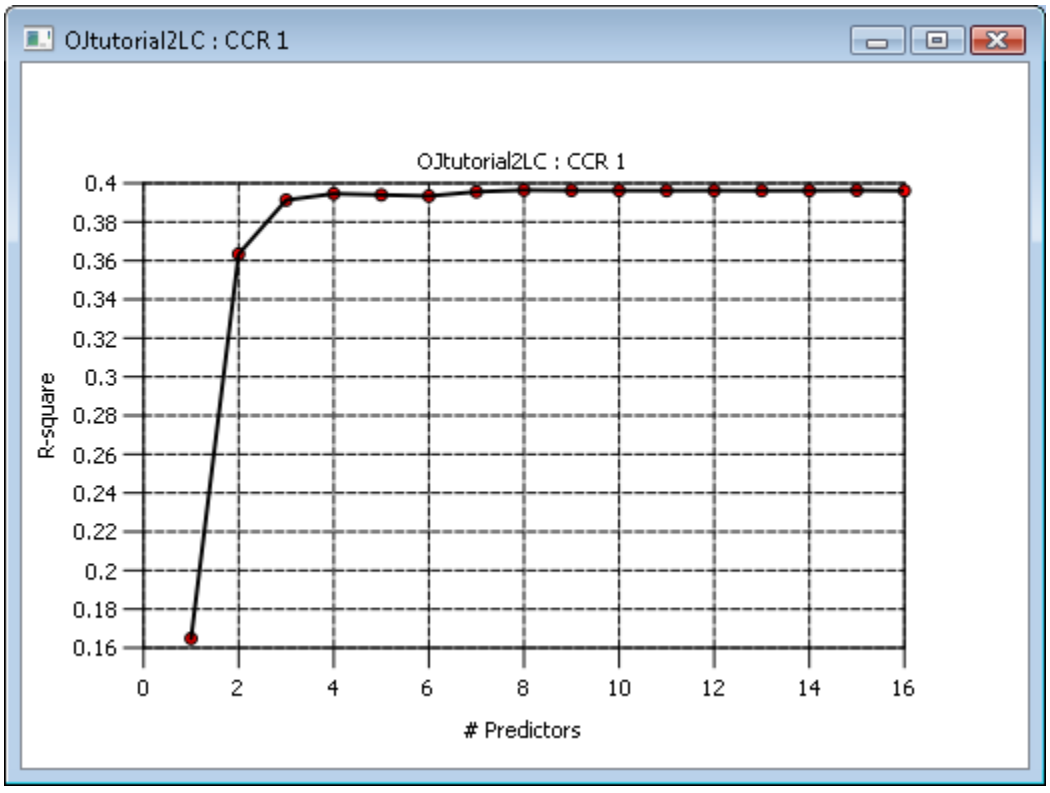


Figure 10. Cross-validation step-down plot (Segment #1)

Cross-Validated Step-Down		
# Predictors	**R^2**	**Std. Dev**
16	0.3961	0.0022
15	0.3963	0.0022
14	0.3962	0.0021
13	0.3961	0.0021
12	0.3962	0.0021
11	0.3962	0.0021
10	0.3962	0.002
9	0.3962	0.0021
8	0.3964	0.0021
7	0.3956	0.0023
6	0.3934	0.0036
5	0.394	0.0033
4	0.3946	0.0037
3	0.3913	0.0083
2	0.3634	0.004
1	0.1648	0.0142

Table 3. Cross-validated Step-Down Table

Table 4 below shows that **Acidity** is an important predictor in the model. The negative standardized coefficient (-.3336) supports the inference that Segment #1 judges tend to dislike OJs with high acid content.

Std. Coefficients

		1.2869	0.8983	0.1402	0.0713	0.0351
Predictors	**Std.Coefficient**	**CC1**	**CC2**	**CC3**	**CC4**	**CC5**
CFactor1	0.4263	0.1993	0.1777	0.2314	-0.3561	0.0876
Fructose	-0.1167	-0.1227	0.0557	-0.1196	0.1566	-0.088
Sweeteningpower	0.1463	0.1113	0.0251	-0.1922	0.2639	-0.3223
pHbeforeproc	-0.003	0.1859	-0.2397	0.3114	-0.7264	-0.5331
Odortypicity	-0.2002	0.1889	-0.4035	-0.61	-0.0897	0.3158
Pulp	0.1349	0.1684	-0.0526	-0.3728	0.5546	-0.6195
Acidity	-0.3336	-0.2031	0.0878	-0.6275	-0.6103	-0.557
Bitterness	-0.1276	-0.2089	0.2304	-0.4548	-0.2484	0.4514

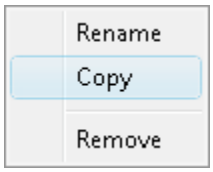
Table 4. Standardized coefficients based on the 5-component model for Segment #1

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

Developing the Corresponding CCR Model for Segment #2

Copying CCR 1 model specifications:

- In the Projects window, right click on 'CCR1' and click 'Copy' to create a new model with the same model specifications as 'CCR 1'.



Next, we will replace the current observation weights by the corresponding values (POSTERIOR.2) associated with Segment #2 (**Posterior2**)

Selecting the Case Weights for Segment #2:

- In the Model Control Window, click on the Weights drop down menu and select 'POSTERIOR.2' .

Estimate the Specified Model:

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

The relevant output for Segment #2 is shown below.

Predictor	**All**	1	2	3	4	5	6	7	8	9	10
CFactor1	100	10	10	10	10	10	10	10	10	10	10
Acidity	98	9	10	10	10	10	10	10	9	10	10
Sweeteningpower	95	9	10	9	9	10	10	9	9	10	10
Smellintensity	94	9	10	9	9	9	10	9	10	10	9
Pulp	31	1	5	1	1	1	10	1	1	10	0
VitaminC	25	1	2	1	1	0	8	1	1	10	0
Tasteintensity	17	0	2	0	0	0	5	0	0	10	0
Odortypicity	12	0	0	0	0	0	2	0	0	10	0
pHbeforeproc	11	1	0	0	0	0	1	0	0	9	0
Bitterness	7	0	1	0	0	0	2	0	0	3	1
pHaftercentrif	6	0	0	0	0	0	0	0	0	6	0
Fructose	2	0	0	0	0	0	1	0	0	1	0
Sweetness	2	0	0	0	0	0	1	0	0	1	0
Total	500	40	50	40	40	40	70	40	40	100	40
Predictors		**4**	**5**	**4**	**4**	**4**	**7**	**4**	**4**	**10**	**4**

Table 5. Predictor Table for Segment 2 Regression

Cross-Validated Components

# Components	R^2	Std. Dev
1	0.1494	0.0056
2	0.3590	0.0074
3	0.4072	0.0070
4	0.4090	0.0062
5	0.4088	0.0062
6	0.4088	0.0062

Table 6. Cross-validated Components Table

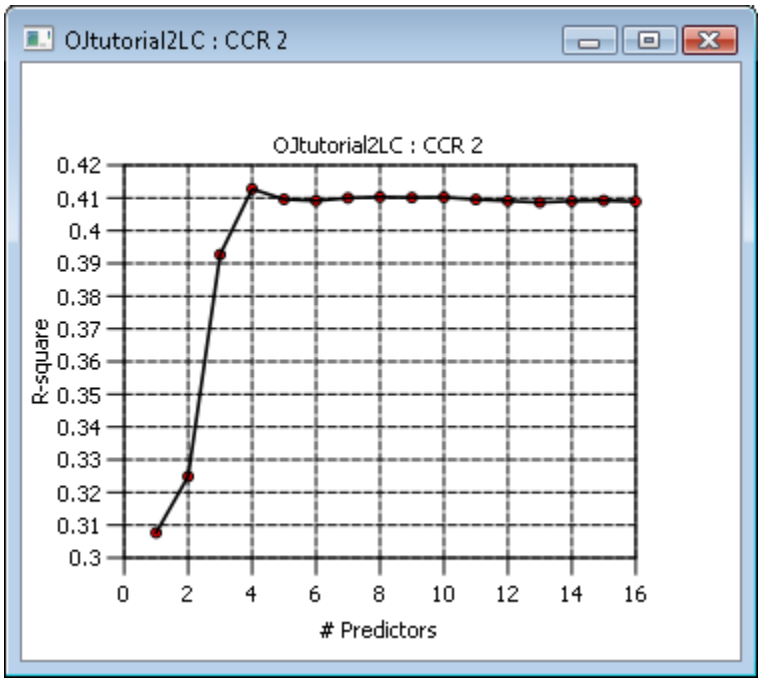


Figure 11. Cross-validation step-down plot (Segment #2). CV-R^2 = .4127

Cross-Validated Step-Down		
# Predictors	**R^2**	**Std. Dev**
16	0.4089	0.0061
15	0.4092	0.0062
14	0.409	0.0063
13	0.4086	0.0063
12	0.4091	0.0061
11	0.4095	0.006
10	0.4102	0.006
9	0.4101	0.0058
8	0.4103	0.0056
7	0.41	0.0055
6	0.4091	0.0061
5	0.4096	0.0055
4	0.4127	0.0054
3	0.3927	0.009
2	0.3248	0.0183
1	0.3077	0.0001

Table 7. Cross-validated Step-Down Table

Table 8 below 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. The table shows that Segment #2 judges *prefer* juices with higher acidity (.214), low sweetening power (-.169), and low smell intensity (-.129).

Std. Coefficients						
			0.6584	0.133	0.0295	0.0056
Predictors	**Std.Coefficient**		**CC1**	**CC2**	**CC3**	**CC4**
CFactor1	0.5548		0.7393	0.6801	-0.9423	0.962
Sweeteningpower	-0.1685		-0.3288	0.386	-0.0922	-0.118
Smellintensity	-0.1294		-0.3066	0.4618	0.3313	0.235
Acidity	0.2136		0.3356	-0.0892	0.1808	-0.1492

Table 8. Standardized Coefficients

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:

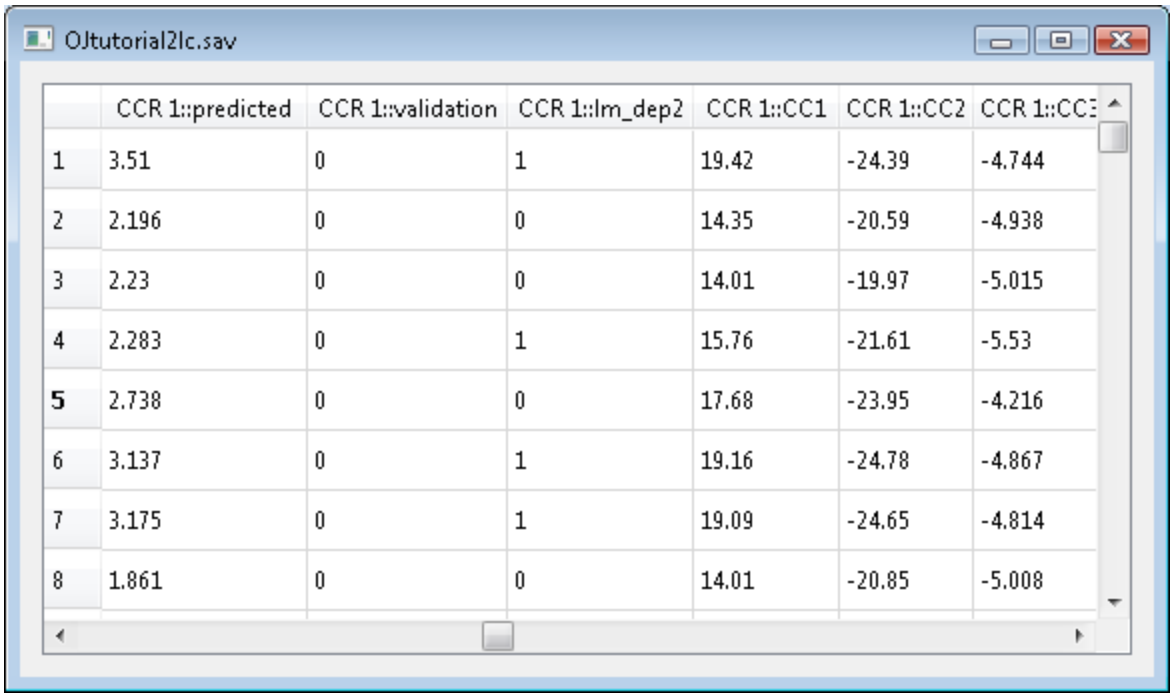
Prediction = .98(3.51) + .02(2.373) = 3.49

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 (3.51) is weighted more heavily for this judge than that from the Segment #2 model (2.373), resulting in a prediction of 3.49 based on the 2-class regression model.

For illustrative purposes, these and other calculations are provided in sheet 'Predictions' (highlighted in yellow) of the Excel file "OJPredictions.xls". This Excel worksheet was created using the predicted scores appended to the dataset.

Viewing the K Components and Predicted Scores on the Dataset:

- Double click on "OJtutorial2lc" in the Projects window
- Click on the window with the dataset and scroll all the way to the right.



	CCR 1::predicted	CCR 1::validation	CCR 1::lm_dep2	CCR 1::CC1	CCR 1::CC2	CCR 1::CC3
1	3.51	0	1	19.42	-24.39	-4.744
2	2.196	0	0	14.35	-20.59	-4.938
3	2.23	0	0	14.01	-19.97	-5.015
4	2.283	0	1	15.76	-21.61	-5.53
5	2.738	0	0	17.68	-23.95	-4.216
6	3.137	0	1	19.16	-24.78	-4.867
7	3.175	0	1	19.09	-24.65	-4.814
8	1.861	0	0	14.01	-20.85	-5.008

Figure 7. Data Viewer

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 " OJtutorial2lc " 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.

For example, cell O3 provides the formula for computing the predicted value 3.49 from the corresponding Segment #1 and Segment #2 output.

			Posterior1			
seqID	**ID**	**rating**	**POSTERIOR.1**	**CCR 1::predicted**	**Residuals**	**ResidSq**
1	1	3	0.9804	3.510	-0.510	0.260
2	1	2	0.9804	2.196	-0.196	0.038

Table 9A. Predictions and residuals output for model with **Posterior1** weights (first 2 rows)

			Posterior2			
seqID	**ID**	**rating**	**POSTERIOR.2**	**CCR 2::predicted**	**Residuals**	**ResidSq**
1	1	3	0.01957	2.373	0.627	0.393
2	1	2	0.01957	3.113	-1.113	1.239

Table 9B. Predictions and residuals output for model with **Posterior2** weights (first 2 rows)

rating	Predictions	Residual	ResidSq
3	3.488	-0.488	0.238
2	2.214	-0.214	0.046

Table 9C. Predictions and residuals computed for 2-class regression model (first 2 rows)

Row 1 in Tables 9A, 9B and 9C 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 (3.51 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)