

## Tutorial #8: LC Regression with High-dimensional Data: A 2-step Approach

DemoData = 'OJtutorial.sav' ([click](#) to download)

This tutorial is based on data described in Tenenhaus et. al. (2005).

### The Data

The data consists of liking ratings on each of 6 different orange juice (OJ) products by 96 judges. In addition, each of the 6 juices is also described by 16 physico-chemical attributes.

	seqID	ID	Orangejuice	rating	rating_mean	Glucose	Fructose	Saccharose	Sweeteningpower	pt
1	1	1	fruvita fr.	3	2.67	23.65	25.65	52.12	102.22	
2	2	1	joker amb.	2	2.67	32.42	34.54	22.92	90.71	
3	3	1	pampryl amb.	2	2.67	25.32	27.36	36.45	89.95	
4	4	1	pampryl fr.	3	2.67	27.16	29.48	38.94	96.51	
5	5	1	tropicana amb.	2	2.67	17.33	20.00	44.15	82.55	
6	6	1	tropicana fr.	4	2.67	22.70	25.32	45.80	94.87	
7	7	2	fruvita fr.	3	2.33	23.65	25.65	52.12	102.22	
8	8	2	joker amb.	2	2.33	32.42	34.54	22.92	90.71	
9	9	2	pampryl amb.	1	2.33	25.32	27.36	36.45	89.95	
10	10	2	pampryl fr.	1	2.33	27.16	29.48	38.94	96.51	
11	11	2	tropicana amb.	3	2.33	17.33	20.00	44.15	82.55	
12	12	2	tropicana fr.	4	2.33	22.70	25.32	45.80	94.87	
13	13	3	fruvita fr.	4	2.50	23.65	25.65	52.12	102.22	

Figure 1: Orange Juice Data (Tenenhaus, et al., 2005)

The variable ID uniquely identifies each judge. The dependent variable 'rating' contain ratings for each of the 6 juices by each judge. The variable 'rating\_mean' provides the average rating for each judge across the 6 juices. Note that the attributes Glucose, Fructose, etc. describe the juices and are identical for each judge.

## ***The Goal and Methodological Challenges***

The overall goal is to predict the liking ratings as a function of the 16 attributes. There are 2 methodological challenges that need to be addressed in accomplishing this goal:

- Observations are not independent -- Since these data consist of multiple records per case, traditional (1-class) regression methods generally suffer from violation of the independent observations assumption which yields suboptimal prediction, since residuals from records associated with the same judge may be correlated. In this tutorial we show how a latent class (LC) regression can be used to identify 2 LC segments having different OJ preferences, to account for correlated observations.
- High-dimensional data – With only 6 juices being rated, use of the 16 correlated attributes as predictors yields a high-dimensional data situation such that traditional regression is not possible due to multicollinearity. We use the Correlated Component Regression (CCR) methods implemented in CORExpress to address this problem.

Two specific goals are:

- Goal 1 – to determine if the judges can be segmented on the basis of their juice liking ratings.
- Goal 2 – to determine if the juice attributes can predict the liking ratings, and if so which attributes are the most important predictors for each segment.

In this tutorial, we illustrate step 1 of the following 2-step approach:

- Step 1: Estimate LC Regression Models to determine the number of segments
- Step 2: Use results from the LC regression to select the attributes to be used as predictors

Step 2 of the 2-step approach is illustrated in forthcoming CORExpress Tutorial 4 (or the alternative XLSTAT tutorial for [XLSTAT-CCR Tutorial 3](#)). More specifically, in step 2 CCR is used for variable selection with these data in which the predictors/attributes are multicollinear.

## ***Opening the Data File***

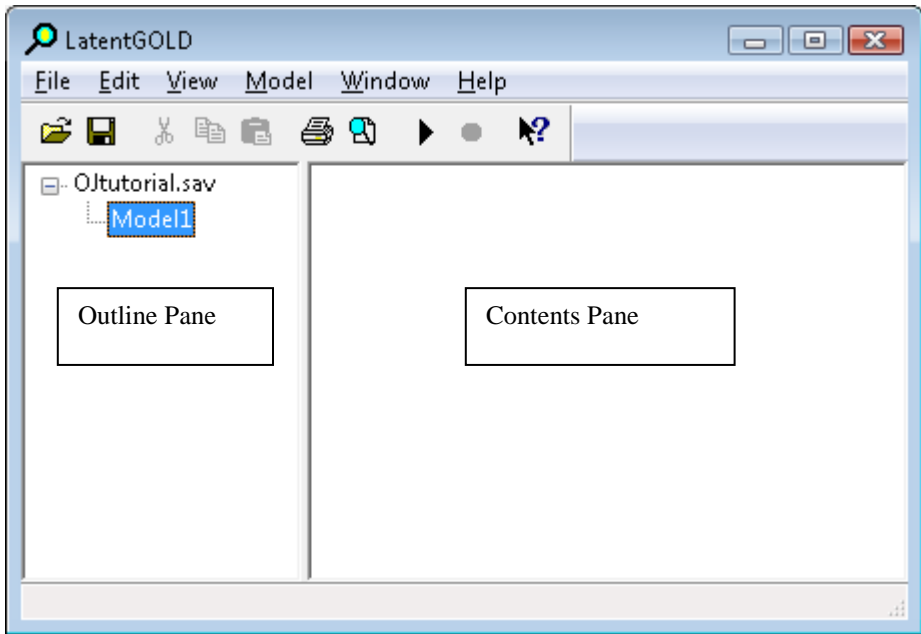
For this example, the data file is in SPSS system file format.

- To open the file, from the menus choose: File→Open
- Open OJtutorial.sav



**Figure 2. Opening OJtutorial.sav**

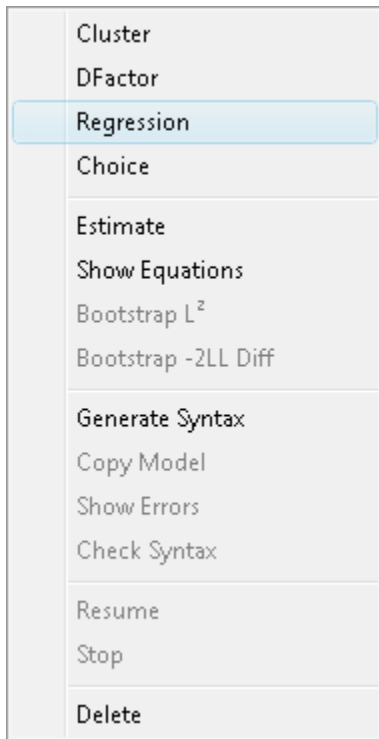
The filename OJtutorial.sav appears in the left-hand pane of Latent GOLD (the Outline pane).



**Figure 3. Data file in the Outline Pane**

## ***Estimating a LC Regression Model***

- Right-click on Model1 and select Regression from the menu:



The Analysis Dialog Box for the Regression Model opens

- Select the variable 'rating', and click on 'Dependent' to move it to the Dependent box
- Select the variable 'ID' and move it to the Case ID box
- Select the nominal variable 'Orangejuice' and move it to the Predictor box

We will begin by estimating 1 to 4 Segments. Thus,

- In the Classes Box, type in '1-4'
- Click 'Scan' to scan the data file

Your model analysis dialog box should now look like this:

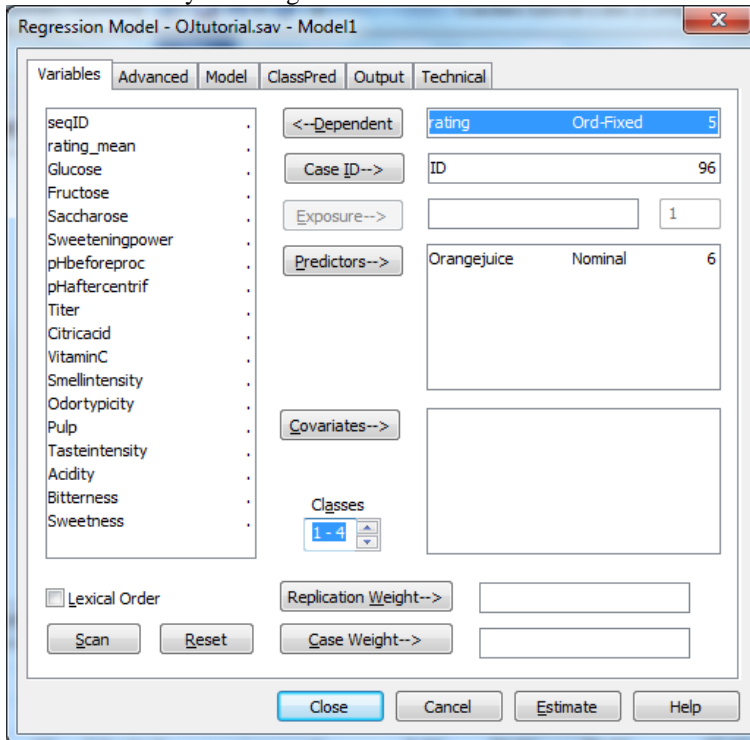


Figure 4. LC Regression Model analysis dialog box prior to estimation

- In the Output tab, select 'Classification – Posterior' to obtain classification output

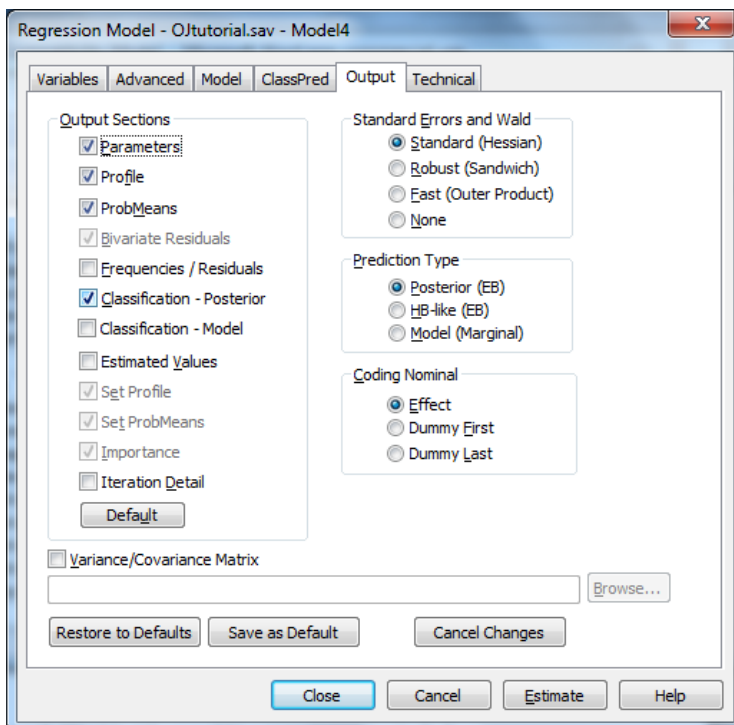


Figure 5. LC Regression Model analysis dialog box prior to estimation

To allow for differential scale usage, as recommended by Magidson and Vermunt (2006) we include a random intercept in the model. To add the random intercept:

- In the Advanced tab, add 1 CFactor

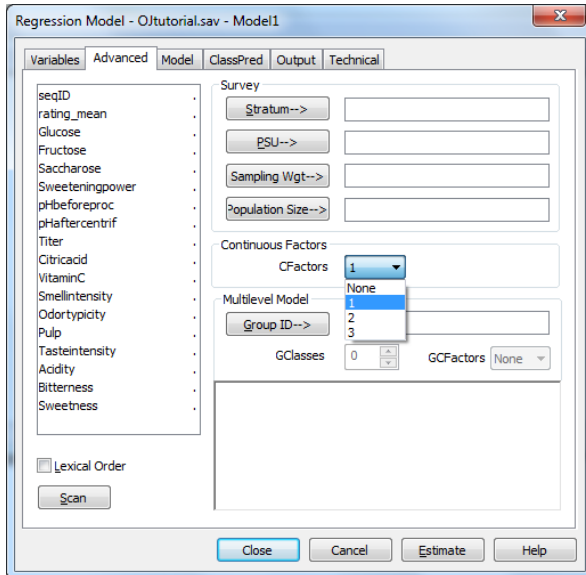


Figure 6. LC Regression Model analysis dialog box prior to estimation

- Click Estimate to estimate the models

After the model has estimated, click on the data file name ‘OJtutorial.sav’ in the Outline (left-hand) Pane. Observing the model summary output, we can see that depending upon the criteria (BIC, AIC, AIC3), the the lowest value for the criterion statistic suggests 1, 3, or 2 classes respectively.

		LL	BIC(LL)	AIC(LL)	AIC3(LL)
Model1	1-Class 1-CFactor Regression	-804.4045	1654.4525	1628.8090	1638.8090
Model2	2-Class 1-CFactor Regression	-786.4798	1664.2465	1612.9595	1632.9595
Model3	3-Class 1-CFactor Regression	-771.9031	1680.7367	1603.8062	1633.8062
Model4	4-Class 1-CFactor Regression	-764.8436	1712.2611	1609.6871	1649.6871

Figure 7. 2-LC Regression Summary Model Output showing the best fit

According to the AIC3, the 2-class solution provides the best model fit. This means that 2 segments are sufficient to explain the correlated observations. From the Profile output it can be seen that segment 1 consists of about 65% of the judges and segment 2 consists of the remaining 35% of the judges. For this tutorial, we will use the 2-class model.

Now, in the Outline Pane,

- Click on the '+' symbol to the left of Model2 to display the various output.
- Click on 'Profile' to display the Profile output:

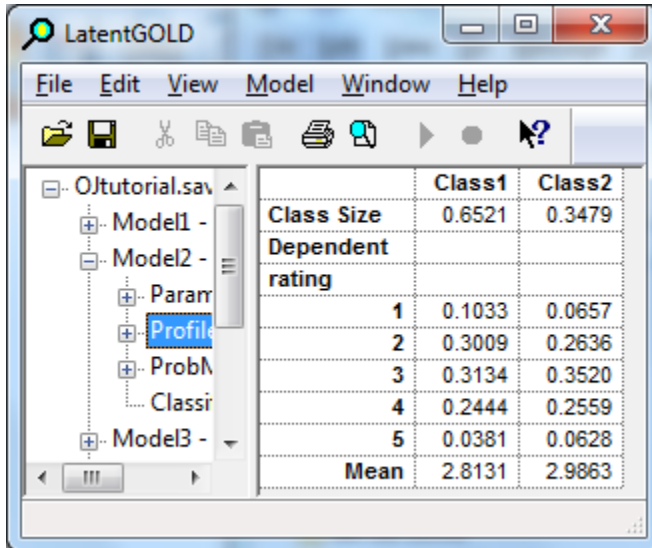


Figure 8. Profile Output showing the sizes of the 2 classes (segments) and their mean ratings

To view the posterior membership probabilities and random intercepts:

- Click on 'Classification' to display the Classification output:

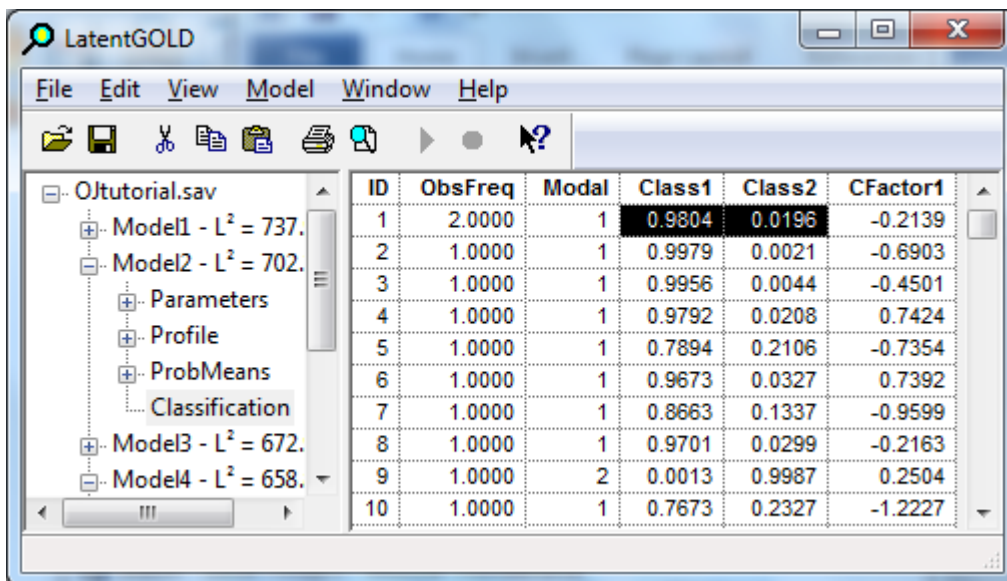


Figure 9. Classification Output for Model 2

In row 1 of the Class1 and Class2 columns, we see that judge #1 has a .9804 probability of being in segment 1, and therefore a .0196 probability of being in segment 2. (The ObsFreq value of '2' means that a second judge provides exactly the same ratings as judge #1 and thus has the same CFactor score and posterior membership probabilities).

The CFactor1 field contains the random intercept (with mean 0), which turns out to be highly correlated (>.99) with the variable 'rating\_mean'. Since the value of CFactor1 for judge#1 is less than 0 (-.2139), this indicates that the ratings given by this judge tends to be somewhat lower than the average judge.

### Step 2: Predicting ratings based on the attributes

For food manufacturers to customize separate orange juice products for these 2 segments, it is important to have a model that predicts liking as a function of the attributes.

From the Prediction Statistics section of the Latent GOLD summary output, it can be seen that  $R^2 = .4147$  for the 2-class regression model based on knowledge of which juice is being rated (capture by dummy variables associated with the Nominal predictor 'Orangejuice'), class membership (Class), and the random intercept (CFactor1).

Prediction Statistics				
rating				
	Error Type	Baseline	Model	R <sup>2</sup>
	Squared Error	1.0690	0.6257	0.4147
	Minus Log-likelihood	1.4304	1.1879	0.1696
	Absolute Error	0.8416	0.6524	0.2248
	Prediction Error	0.6736	0.5313	0.2114

Figure 10. Prediction Statistics Section of the Model Summary Output

Since the nominal variable Orangejuice has 6 categories, it can be perfectly predicted by any 5 attributes. Thus, replacing the predictor 'Orangejuice' by 16 attributes creates a multicollinearity problem. To the extent to which use of these attributes (along with CFactor1) can achieve an  $R^2$  close to .4147, our model might be considered successful since the attributes would be explaining the differences among the juices. However, if this achievement was obtained from traditional regression techniques such an inference would be questionable since good prediction can always be achieved in this high-dimensional setting by overfitting.

To avoid overfitting, we use the Correlated Component Regression (CCR) methods implemented in the CORExpress program, with 10 rounds of 10-fold cross-validation (for details see CORExpress Tutorial 4, forthcoming). The resulting predictor tables shown below (Tables 1 and 2), are used to determine the attributes to include in the model. For example, for the segment 1 regression (Table 1), 'Acidity' was included in cross-validation regression models 99 out of a possible 100 times, and for segment 2 'Acidity' was included 98 of a possible 100 times. For further details on the interpretation of Tables 1 and 2, see CORExpress Tutorial 4 (forthcoming).

The resulting predictor tables for segment 1 and segment 2 are shown below.

**Table 1. Predictor Table for Segment 1 Regression**

<b>Predictor</b>	<b>All</b>	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		<b>15</b>	<b>4</b>	<b>3</b>	<b>3</b>	<b>8</b>	<b>8</b>	<b>8</b>	<b>4</b>	<b>3</b>	<b>8</b>

**Table 2. Predictor Table for Segment 2 Regression**

<b>Predictor</b>	<b>All</b>	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	4	7	4	4	10	4

From the cross-validation results summarized in these tables we note the following:

- For segment 1, the 10 rounds result in between 3 and 15 predictors being included in a 5-component model (see last row of Table 1). In addition, as shown above the cutoff line in Table 1, 3 predictors were included in the model at least 80% of the time: CFactor1, Acidity, Sweeteningpower.
- For segment 2, the results over the 10 rounds are much more consistent -- 4 predictors being selected as optimal in 7 of the 10 rounds. In addition, the following 4 predictors were included at least 80% of the time: CFactor1, Acidity, Sweeteningpower, Smellintensity.

We now estimate a new 2-class regression model containing a random intercept along with the 3 predictors Acidity, Sweeteningpower and Smellintensity, with the effect of Smellintensity set to 0 for segment 1. This results in a model with  $R^2 = .4082$ . As shown below (Fig. 11), the effects for the predictors Acidity ( $p=5.2E-9$ ) and Sweeteningpower ( $p=.003$ ) are statistically significant, and the effect for Smellintensity is marginally significant ( $p=.06$ ). The single most important predictor is Acidity, with segment 1 showing an adversity towards orange juices with high acidity levels, while segment 2 preferring such.

CFactor1 : Intercept						
	0.4429	0.4429	27.4783	1.6e-7	0.0000	.
Predictors	Class1	Class2	Wald	p-value	Wald(=)	p-value
Acidity	-1.6791	0.5274	38.1474	5.2e-9	37.3537	9.9e-10
Sweeteningpower	0.0461	-0.0203	11.4917	0.0032	8.1356	0.0044
Smellintensity	0.0000	-1.0379	3.5490	0.060	0.0000	.

Figure 11. Parameters Output

To retrieve the setup for this model:

- From the menus choose: File→Open
- From the 'Files of type' drop down menu, select 'LatentGOLD files (\*.lgf)'
- Open final.lgf

This model can now be estimated, and the Parameters output (Figure 11) as well as additional output reported by Latent GOLD can be viewed. In addition, predictions and classification information can be output to a file using the ClassPred tab (see [Latent GOLD User's Guide](#) for details).

### **References:**

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

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.