1 Theme

Comparing the results of the Partial Least Squares Regression from various data mining tools (free: Tanagra, R; commercial: SIMCA-P, SPAD, and SAS).

Comparing the behavior of tools is always a good way to improve them.

To check and validate the implementation of methods. The validation of the implemented algorithms is an essential point for data mining tools. Even if two programmers use the same references (books, articles), the programming choice can modify the behavior of the approach (behaviors according to the interpretation of the convergence conditions for instance). The analysis of the source code is possible solution. But, if it is often available for free software, this is not the case for commercial tools. Thus, the only way to check them is to compare the results provided by the tools on a benchmark dataset[1]. If there are divergences, we must explain them by analyzing the formulas used.

To improve the presentation of results. There are certain standards to observe in the production of reports, consensus initiated by reference books and / or leader tools in the field. Some ratios should be presented in a certain way. Users need reference points.

Our programming of the PLS approach is based on the Tenenhaus book (1998) [2] which, itself, make reference to the SIMCA-P[3] tool. Using the access to a limited version of this software (version 11), we have check the results provided by Tanagra on various datasets. We show here the results of the study on the CARS dataset. We extend the comparison to other data mining tools.

We have also much used the Garson website[4] for the description of the PLS regression method, particularly to understand the tables and the figures provided by the various tools.

As a reminder, the goal of the PLS Regression is to explain / predict the values of one or more target attributes (the dependents) from the values of one or more explanatory variables (the predictors). All the variables are continuous or considered as such.

2 Dataset

We use the CARS_PLS_REGRESSION.XLS data file in this tutorial (Excel file format - http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/cars_pls_regression.xls).

We want to explain the costs indicators of cars (price, consumption, symboling) from their characteristics (engine size, fuel type, etc.).

There are 20 instances into the data file (Figure 1). It is a sample drawn from the Automobile Data Set available on the UCI Machine Learning Repository Server[5].

[1] In my case, I try often to reproduce the formulas in a spreadsheet application. This allows me to check all the intermediate results.

[2] M. Tenenhaus, « La régression PLS – Théorie et Pratique », Technip, 1998.

[3] SIMCA-P for Multivariate Data Analysis. http://www.umetrics.com/default.asp/pagename/software_simcap/c/3

[4] D. Garson, « Partial Least Squares Regression », from *Statnotes: Topics in Multivariate Analysis.* Retrieved 05/18/2008 from http://www2.chass.ncsu.edu/garson/pa765/statnote.htm.

[5] http://archive.ics.uci.edu/ml/datasets/Automobile. It describes some cars in 1985, this is the reason why some features may appear to be strange today.

Numéro	diesel	twodoors	sportsstyle	wheelbase	length	width	height	curbweight	enginesize	horsepower	horse_per_w	conscity	price	symboling
1	0	1	0	97	172	66	56	2209	109	85	0.0385	8.7	7975	2
2	0	0	0	100	177	66	54	2337	109	102	0.0436	9.8	13950	2
3	0	0	0	116	203	72	57	3740	234	155	0.0414	14.7	34184	-1
4	0	1	1	103	184	68	52	3016	171	161	0.0534	12.4	15998	3
5	0	0	0	101	177	65	54	2765	164	121	0.0438	11.2	21105	0
6	0	1	0	90	169	65	52	2756	194	207	0.0751	13.8	34028	3
7	1	0	0	105	175	66	54	2700	134	72	0.0267	7.6	18344	0
8	0	0	0	108	187	68	57	3020	120	97	0.0321	12.4	11900	0
9	0	0	1	94	157	64	51	1967	90	68	0.0346	7.6	6229	1
10	0	1	0	95	169	64	53	2265	98	112	0.0494	9.0	9298	1
11	1	0	0	96	166	64	53	2275	110	56	0.0246	6.9	7898	0
12	0	1	0	100	177	66	53	2507	136	110	0.0439	12.4	15250	2
13	0	1	1	94	157	64	51	1876	90	68	0.0362	6.4	5572	1
14	0	0	0	95	170	64	54	2024	97	69	0.0341	7.6	7349	1
15	0	1	1	95	171	66	52	2823	152	154	0.0546	12.4	16500	1
16	0	0	0	103	175	65	60	2535	122	88	0.0347	9.8	8921	-1
17	0	0	0	113	200	70	53	4066	258	176	0.0433	15.7	32250	0
18	0	0	0	95	165	64	55	1938	97	69	0.0356	7.6	6849	1
19	1	0	0	97	172	66	56	2319	97	68	0.0293	6.4	9495	2
20	0	0	0	97	172	66	56	2275	109	85	0.0374	8.7	8495	2

Figure 1 – Dataset: the predictors (green) and the dependents (blue)

3 The PLS Regression

Partial least squares (PLS) is sometimes called "Projection to Latent Structures" because of its general strategy. The X variables (the predictors) are reduced to principal components t_h (says also factors or latent variables), as are the Y variables (the dependents). The components of X are used to predict the scores on the Y components u_h (PLS responses), and the predicted Y component scores are used to predict the actual values of the Y variables. In constructing the principal components of X, the PLS algorithm iteratively maximizes the strength of the relation of successive pairs of X and Y component scores (u_h, t_h) by maximizing the covariance of each X-score with the Y variables. This strategy means that while the original X variables may be multicollinear, the X components used to predict Y will be orthogonal. The number of components t_h must not exceed the number of predictors (Garson, PLS Regression).

4 PLS Regression with TANAGRA and SIMCA-P

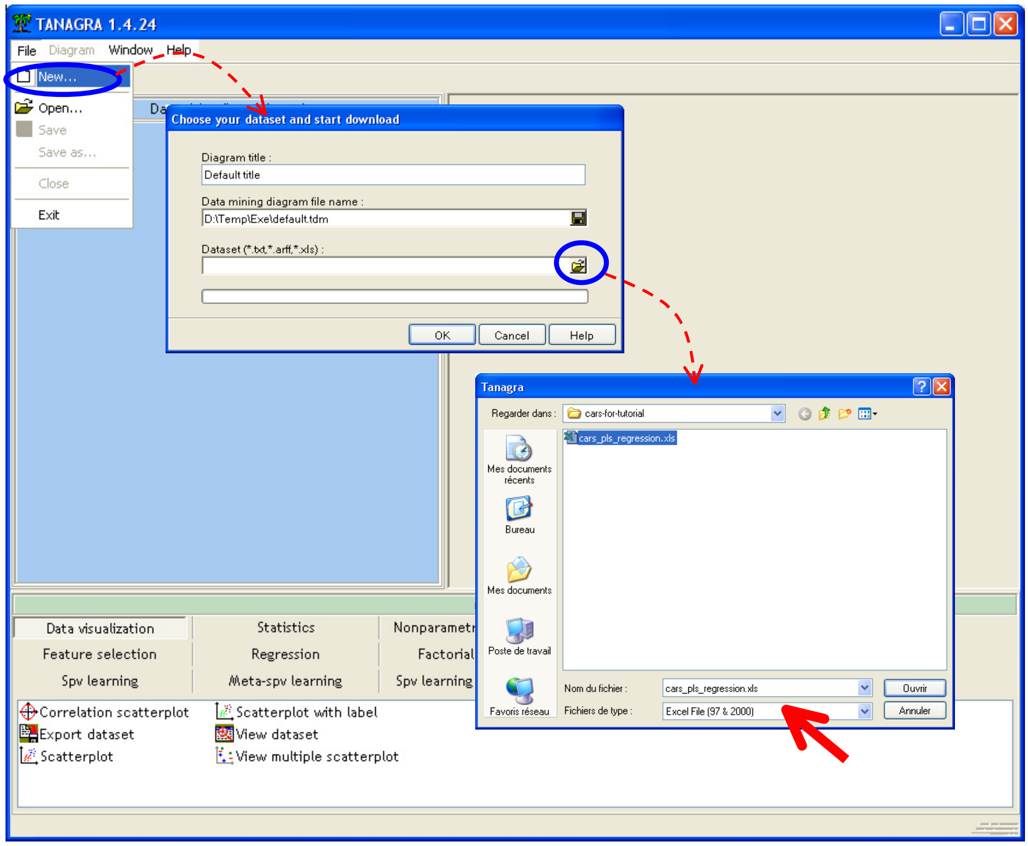
In this section, we detail the outputs of **TANAGRA**. We will compare them to those of **SIMCA-P**. We observe that the results are exactly identical. This suggests that the calculations are based on the same formulas but also that the underlying programming choices are similar (accuracy of the calculations, etc.).

4.1 Importing the data file and creating a diagram

We launch Tanagra. We click on the FILE / NEW menu to create a diagram and import the CARS_PLS_REGRESSION.XLS data file[6].

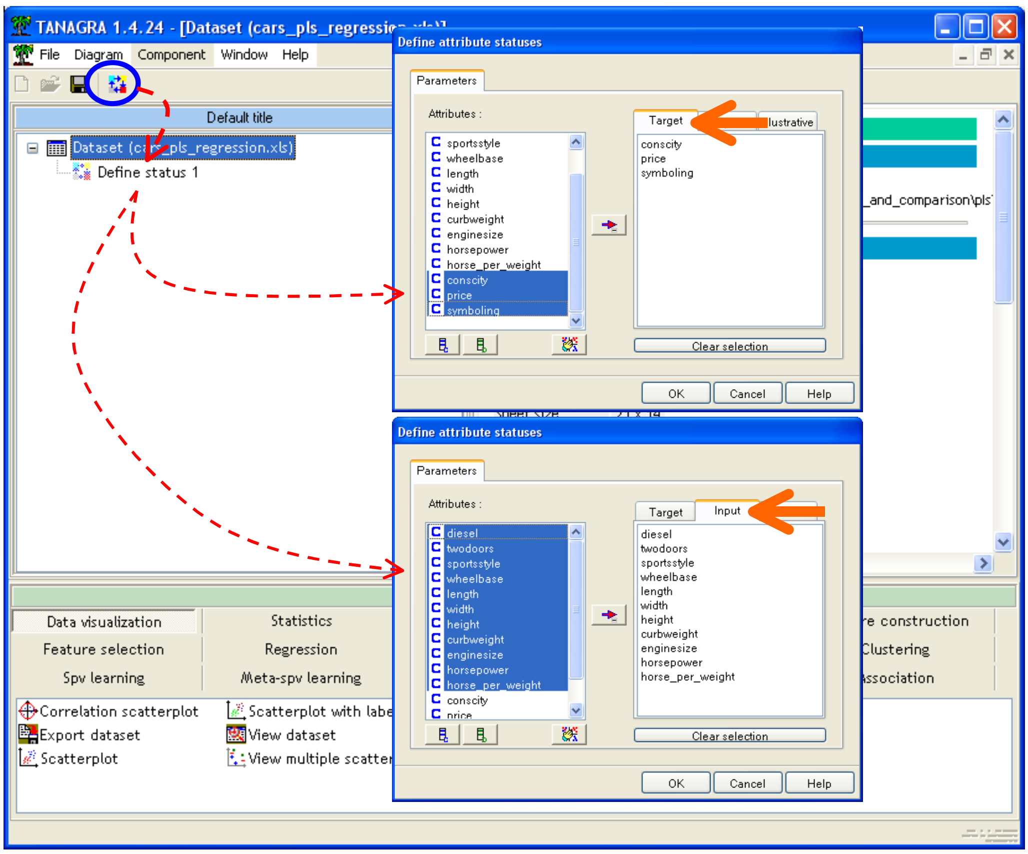
[6] There are various ways to import a XLS data file. We can use the add-on for Excel (http://data-mining-tutorials.blogspot.com/2010/08/sipina-add-in-for-excel.html, http://data-mining-tutorials.blogspot.com/2010/08/tanagra-add-in-for-office-2007-and.html) or, as we do in this tutorial, directly import the dataset (http://data-mining-tutorials.blogspot.com/2008/10/excel-file-format-direct-importation.html). In this last case, the dataset must not be opened in the spreadsheet application. The values must be in the first sheet. The first row corresponds to the name of the variables.

The direct importation is faster than the use of the “tanagra.xla add-on. But, on the other hand, Tanagra can handle only the XLS format here (up to Excel 2003).



4.2 Dependent and independent variables

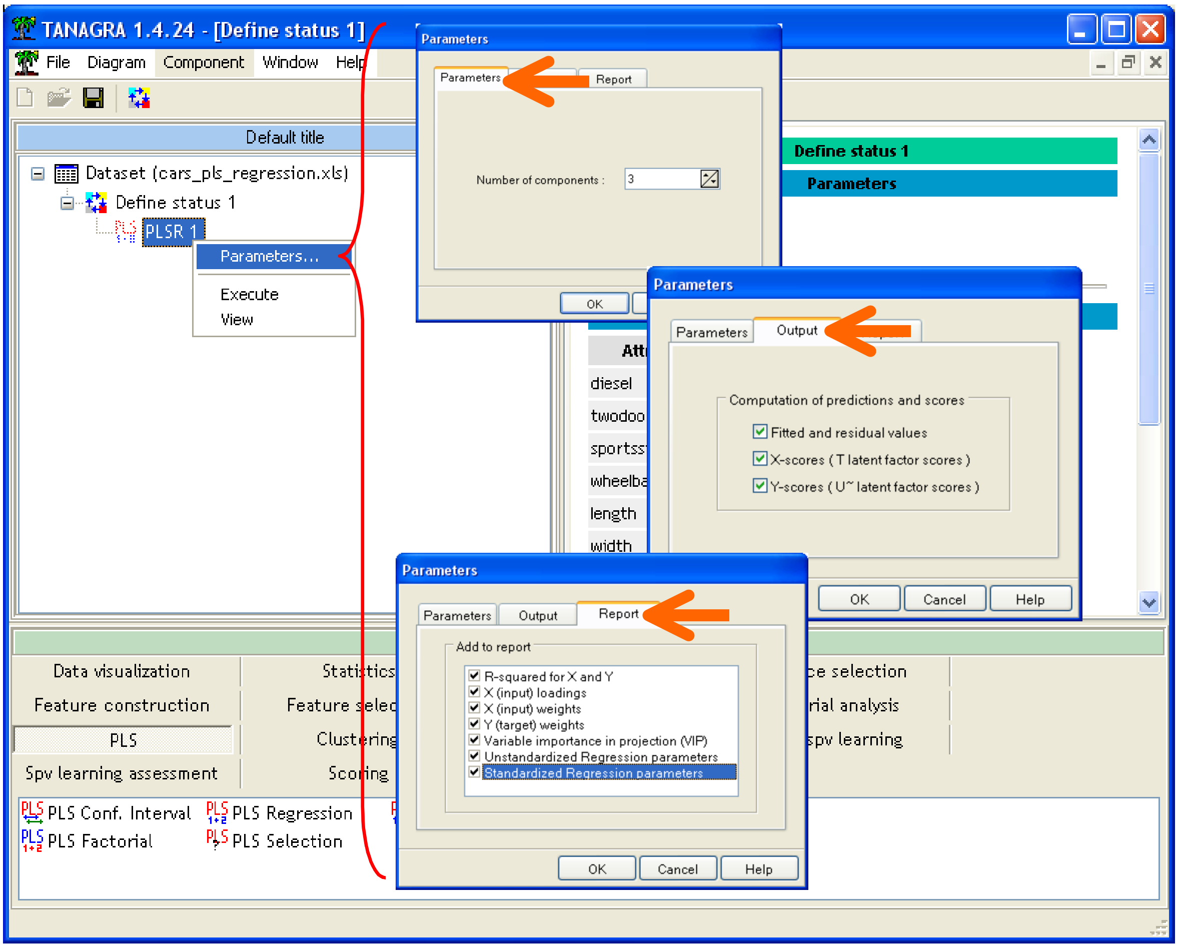
CONSCITY, PRICE and SYMBOLING are the target attributes. The others are the predictors. We use the DEFINE STATUS component to specify the type of the variables in the analysis.



4.3 Settings of the PLS Regression

We add the **PLSR** component (PLS tab) into the diagram. We click on the PARAMETERS menu. Into the first tab (Parameters), we set the number of extracted factors to 3. The number of factors does not exceed the number of predictors. Into the second tab (Output), we specify the new columns generated by the tool. They are available for other calculations in the subsequent part of the diagram. Last, into the third tab (Report), we set the tables which will be included into the report.

We click on the OK button to validate these settings. Then, we click on the VIEW contextual menu.



4.4 Description of the output

Our presentation draws heavily from the text of Garson, dedicated to the description of the results of SPSS (http://faculty.chass.ncsu.edu/garson/PA765/pls.htm).

4.4.1 Proportion of variance explained by latent factors

This table describes the proportion of variance explained by the latent factors, for the predictors (X) and the dependent variables (Y) (Figure 2 and Figure 3).

On the one hand, it shows the quality of the representation of the predictors. The cumulative proportion is 100% if we use all the factors (equal to the number of predictors). For our dataset, we note that the three first factors explain 81.927% of the variance of the predictors.

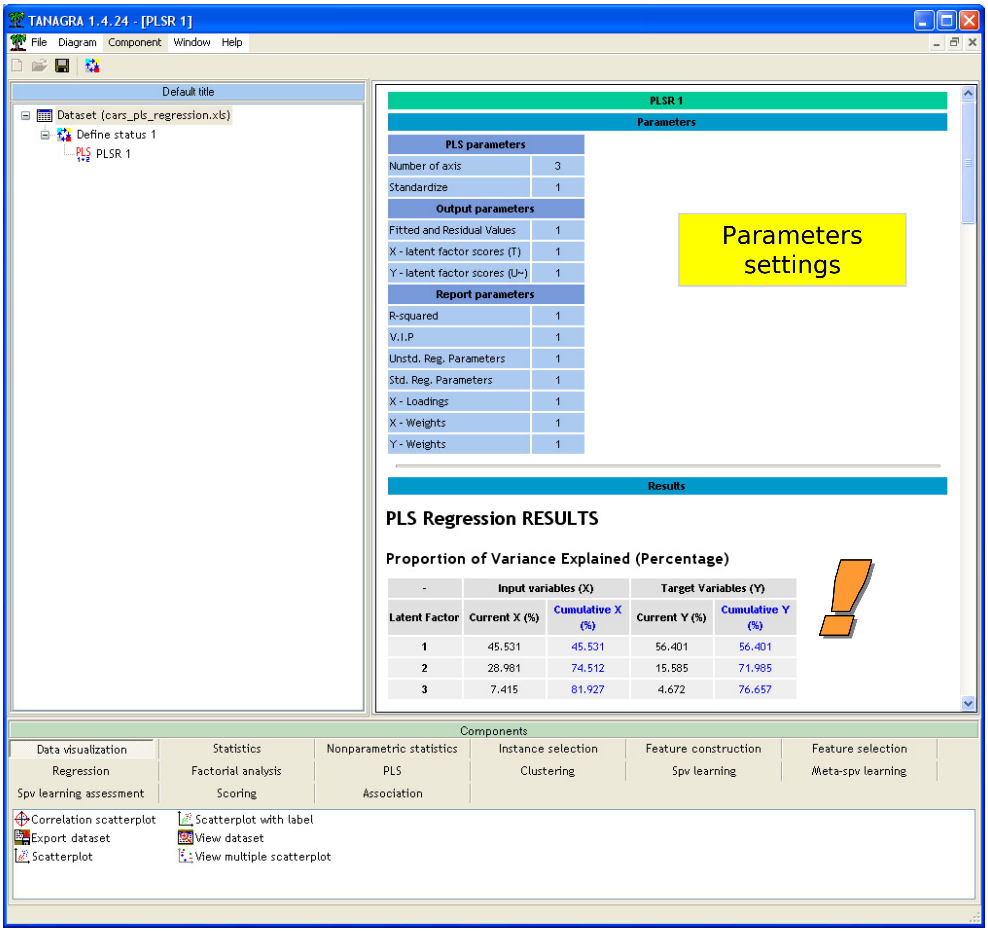


Figure 2 – Tanagra – Proportion of variance explained by latent factor

On the other hand, this table shows the predictive power of the factors. If we use all the factors (equal to the number of predictors), we obtain the R-squared of the linear multiple regression.

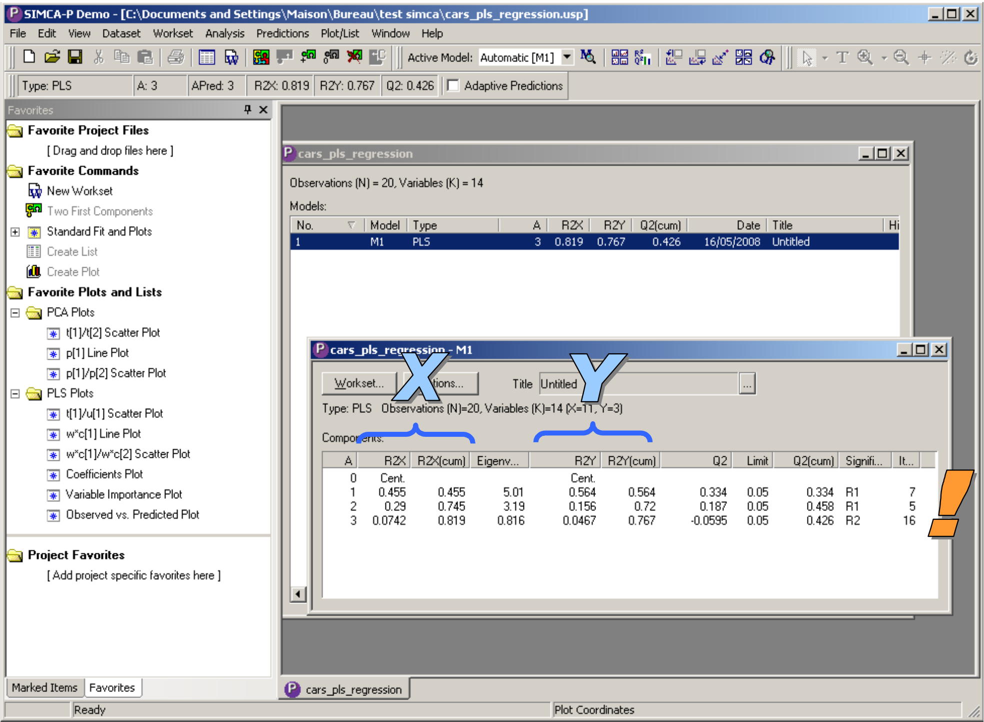


Figure 3 - SIMCA-P - Proportion of variance explained by latent factor

If the cumulative proportion is equal to 1, it means that we can predict exactly the values of the dependent variables from the selected factors. For our dataset, we observe that our model is rather good, 76.657% of the variance of the dependent variables are explained by the first three factors.

4.4.2 Proportion of explained variance for each variable – R^2

These tables (Figure 4) examine in more detail the preceding one. It shows the squared the correlation of each variable with the latent factor i.e. the proportion of variance explained for each variable. This is an important tool for the interpretation and the comprehension of the factors.

- For the first factor, we observe that LENGTH, CURBWEIGHT, ENGINESIZE, and to a lesser extent, HORESPOWER, are important. But we have not the direction of the influence at this stage.
- About the dependent variables, always for the first factor, we observe that CONSCITY and PRICE are well explained. There are thus a form of connection, which remains to be determined, between the predictors above and these dependent variables.
- For the second factor, which represents 28.98% of the variance of the predictors, the variables TWODOORS, HORSE_PER_WEIGHT and HEIGHT are important.
- For the dependent variables, this factor explain 15.56% of their variance, it is mainly in relation with SYMBOLING.
- In the last row, we have the proportion of variance explained by each factor (Figure 2).
- The third factor is hard to understand. It explains a minor part of the variance (7.42%) anyway.

R-squared

Input(s) vs. X-Scores

-	R-squared			Cumulative R-squared		
Input	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
diesel	0.0871	0.1904	0.4243	0.0871	0.2775	0.7018
twodoors	0.0012	0.6790	0.0152	0.0012	0.6802	0.6954
sportsstyle	0.0189	0.2612	0.2483	0.0189	0.2801	0.5284
wheelbase	0.5527	0.3602	0.0393	0.5527	0.9129	0.9522
length	0.8236	0.1376	0.0027	0.8236	0.9613	0.9639
width	0.7692	0.0961	0.0118	0.7692	0.8652	0.8771
height	0.0157	0.5017	0.0173	0.0157	0.5174	0.5347
curbweight	0.9244	0.0258	0.0026	0.9244	0.9502	0.9528
enginesize	0.8967	0.0037	0.0240	0.8967	0.9005	0.9244
horsepower	0.7035	0.2546	0.0155	0.7035	0.9581	0.9737
horse_per_weight	0.2155	0.6775	0.0146	0.2155	0.8930	0.9076
Total Exp.	0.4553	0.2898	0.0742	0.4553	0.7451	0.8193

Target(s) vs. X-Scores

-	R-squared			Cumulative R-squared		
Target	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
conscity	0.8836	0.0407	0.0030	0.8836	0.9243	0.9273
price	0.7790	0.0168	0.1129	0.7790	0.7958	0.9087
symboling	0.0294	0.4100	0.0242	0.0294	0.4394	0.4637
Total Exp.	0.5640	0.1558	0.0467	0.5640	0.7199	0.7666

Figure 4 - Tanagra – R^2 with the factors t_h – Independent and dependent variables

Summary [M1]

	1	2	3	4	5	6	7
1	Var ID (Primary)		R2VY	R2VY(cum)	Q2VY	Q2 limit	Q2VY(cum)
2	Total	Comp 1	0.5640	0.5640	0.3340	0.0500	0.3340
3		Comp 2	0.1558	0.7199	0.1868	0.0500	0.4584
4		Comp 3	0.0467	0.7666	-0.0595	0.0500	0.4262
5							
6	conscity	Comp 1	0.8836	0.8836	0.7999	0.0500	0.7999
7		Comp 2	0.0407	0.9243	0.2820	0.0500	0.8564
8		Comp 3	0.0030	0.9273	-0.0425	0.0500	0.8503
9							
10	price	Comp 1	0.7790	0.7790	0.5834	0.0500	0.5834
11		Comp 2	0.0168	0.7958	-0.3478	0.0500	0.5417
12		Comp 3	0.1129	0.9087	0.3452	0.0500	0.6999
13							
14	symboling	Comp 1	0.0294	0.0294	-0.3812	0.0500	-0.1000
15		Comp 2	0.4100	0.4394	0.2971	0.0500	0.2268
16		Comp 3	0.0242	0.4637	-0.2092	0.0500	0.1495
17							

Figure 5 – SIMCA-P – R² of the dependent variables with t_h

SIMCA-P: Menu ANALYSIS / SUMMARY / MODEL OVERVIEW LIST (Figure 5)

4.4.3 LOADINGS of predictors – Ph Vector

The loadings represent the "correlation" between the factors and the predictors. They supplement the values provided by the R2 table (Figure 4). It specifies the direction of the association. In practice, one considers that an absolute value upper than 0.4 reflects a significant association. But the most important is to obtain an interesting interpretation of the results.

Note that the LOADINGS do not exactly correspond to correlation coefficients. However, they allow to position the variables in the same way with regard to the factors, and this is the most important for the interpretation. We will focus primarily on variables with high absolute value.

We observe that the first factor describes the cars with the same characteristics about LENGTH, WIDTH, CURBWEIGHT and ENGINESIZE. For the second factor, we observe the association between TWODOORS and HORSE_PER_WEIGHT, antinomic with HEIGHT.

TANAGRA

X-loadings (Model Effect Loadings - Vector Ph)

Input	Factor1	Factor2	Factor3
diesel	-0.1326	-0.2496	-0.7368
twodoors	0.0154	0.4714	0.1393
sportsstyle	-0.0617	0.2924	0.5636
wheelbase	0.3342	-0.3433	0.2242
length	0.4080	-0.2122	0.0586
width	0.3942	-0.1773	0.1231
height	0.0563	-0.4052	0.1488
curbweight	0.4322	-0.0919	-0.0577
enginesize	0.4257	0.0349	-0.1751
horsepower	0.3770	0.2886	-0.1410
horse_per_weight	0.2087	0.4708	-0.1367

SIMCA-P

General List [M1]

	1	2	3	4
1	Var ID (Primary)	M1.p[1]	M1.p[2]	M1.p[3]
2	diesel	-0.1326	-0.2496	0.7368
3	twodoors	0.0154	0.4714	-0.1393
4	sportsstyle	-0.0617	0.2924	-0.5636
5	wheelbase	0.3342	-0.3433	-0.2242
6	length	0.4080	-0.2122	-0.0586
7	width	0.3942	-0.1773	-0.1231
8	height	0.0563	-0.4052	-0.1488
9	curbweight	0.4322	-0.0919	0.0577
10	enginesize	0.4257	0.0349	0.1751
11	horsepower	0.3770	0.2886	0.1410
12	horse_per_weigh	0.2087	0.4708	0.1367

Figure 6 - X-Loadings - Ph Vector (Association between INPUT variables - Factors)

SIMCA-P: Menu ANALYSIS / LOADINGS / LINE PLOT, select the « p » series.

4.4.4 WEIGHTS for dependent variables (TARGET) – Ch Vector

They indicate how much the dependent variables are correlated to the PLS responses (Figure 7). It enables to determine which are the variables that are well explained by the PLS responses (u_h). Here also, these are not really the correlation coefficient, but the interpretation is the same.

Thereafter, we must make the connection between these characteristics with the predictors.

TANAGRA

Y-Weights (Dependent Variable Weights - Vector Ch)

Target	Factor 1	Factor 2	Factor 3
conscity	0.4225	0.1154	0.0623
price	0.3968	0.0742	-0.3800
symboling	-0.0771	0.3663	-0.1760

SIMCA-P

General List [M1]

	1	2	3	4
1	Var ID (Primary)	M1.c[1]	M1.c[2]	M1.c[3]
2	conscity	0.4225	0.1154	-0.0623
3	price	0.3968	0.0742	0.3800
4	symboling	-0.0771	0.3663	0.1760

Figure 7 – Weights for dependent variables – PLS Responses

SIMCA-P: Menu ANALYSIS / LOADINGS / LINE PLOT, « c » serie.

4.4.5 WEIGHTS for the predictors (INPUT) – Wh and Wh* vectors

They indicate how much the predictors are correlated with the PLS Responses (u_h) (Figure 8). We observe that WEIGHTS and LOADINGS (section 4.4.3) are quite similar and serve similar interpretive uses. The vectors Wh*, contrary to Wh, are directly related to the predictors. They can interpret them easily.

TANAGRA

X-Weights (Model Effect Weights - Vector Wh and Wh*)

-	Wh			Wh*		
Input	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
diesel	-0.1396	-0.1605	-0.6992	-0.1396	-0.1731	-0.6691
twodoors	0.0262	0.4727	0.2439	0.0262	0.4750	0.1750
sportsstyle	-0.0723	0.1570	0.5264	-0.0723	0.1505	0.5077
wheelbase	0.2873	-0.3950	0.2310	0.2873	-0.3691	0.2727
length	0.3725	-0.1732	0.0704	0.3725	-0.1396	0.0761
width	0.3560	-0.1094	0.0001	0.3560	-0.0773	-0.0025
height	0.0142	-0.3782	0.1794	0.0142	-0.3769	0.2327
curbweight	0.4278	-0.1395	0.0468	0.4278	-0.1009	0.0448
enginesize	0.4490	-0.0117	-0.1792	0.4490	0.0288	-0.2005
horsepower	0.4144	0.3025	-0.1251	0.4144	0.3399	-0.1896
horse_per_weight	0.2645	0.5235	-0.1845	0.2645	0.5474	-0.2728

SIMCA-P

General List [M1]

	1	2	3	4	5	6	7
1	Var ID (Primary)	M1.w[1]	M1.w[2]	M1.w[3]	M1.w*[1]	M1.w*[2]	M1.w*[3]
2	diesel	-0.1396	-0.1605	0.6992	-0.1396	-0.1731	0.6691
3	twodoors	0.0262	0.4727	-0.2439	0.0262	0.4750	-0.1750
4	sportsstyle	-0.0723	0.1570	-0.5264	-0.0723	0.1505	-0.5077
5	wheelbase	0.2873	-0.3950	-0.2310	0.2873	-0.3691	-0.2727
6	length	0.3725	-0.1732	-0.0704	0.3725	-0.1396	-0.0761
7	width	0.3560	-0.1094	-0.0001	0.3560	-0.0773	0.0025
8	height	0.0142	-0.3782	-0.1794	0.0142	-0.3769	-0.2327
9	curbweight	0.4278	-0.1395	-0.0468	0.4278	-0.1009	-0.0448
10	enginesize	0.4490	-0.0117	0.1792	0.4490	0.0288	0.2005
11	horsepower	0.4144	0.3025	0.1251	0.4144	0.3399	0.1896
12	horse_per_weight	0.2645	0.5235	0.1845	0.2645	0.5474	0.2728

Figure 8 – Weight of the predictors related to the PLS responses

SIMCA-P: Menu ANALYSIS / LOADINGS / LINE PLOT, « w » et « w* » series.

4.4.6 Variable Importance in Projection for independent variables (VIP)

The VIP table reflects the relative importance of the predictors, through the H first factors, in the prediction model. We consider often that a predictor is significant when (VIP > 1). On the hand, for small value of VIP (< 0.8), we can consider that the predictor is not relevant. We can remove the variable from the model.

In our table (Figure 9), because we select the first three factors, we analyze the third column.

TANAGRA

VIP (Variable Importance in Projection)

Input	Factor 1	Factor 2	Factor 3
diesel	0.4630	0.4788	0.7369
twodoors	0.0870	0.7334	0.7383
sportsstyle	0.2399	0.3222	0.5322
wheelbase	0.9530	1.0408	1.0261
length	1.2354	1.1257	1.0924
width	1.1807	1.0586	1.0259
height	0.0472	0.5851	0.5857
curbweight	1.4187	1.2741	1.2353
enginesize	1.4893	1.3184	1.2860
horsepower	1.3746	1.3032	1.2670
horse_per_weight	0.8774	1.1206	1.0964

SIMCA-P

VIP List [M1]

	1	2
1	Var ID (Primary)	M1.VIP[3]
2	enginesize	1.2860
3	horsepower	1.2670
4	curbweight	1.2353
5	horse_per_weight	1.0964
6	length	1.0924
7	wheelbase	1.0261
8	width	1.0259
9	twodoors	0.7383
10	diesel	0.7369
11	height	0.5857
12	sportsstyle	0.5322

Figure 9 – Variable importance in projection

SIMCA-P: Menu ANALYSIS / VARIABLE IMPORTANCE / LIST. We can only display the VIP according to the number of selected factors. The variables are sorted according the VIP.

4.4.7 Unstandardized regression coefficients

This table provides the estimated unstandardized regression coefficients, one column for each dependent variable (Figure 10). We can use them directly for the prediction on new instances.

TANAGRA

Unstandardized Regression Parameters (by target variable)

-	Target Variable(s)		
Input	conscity	price	symboling
diesel	-0.946269	4688.641651	0.208085
twodoors	0.450889	-393.099643	0.337465
sportsstyle	0.129202	-4734.098001	-0.081700
wheelbase	0.040765	-23.277199	-0.035576
length	0.036369	86.825747	-0.009458
width	0.192261	596.478952	-0.030636
height	-0.029355	-454.791393	-0.093726
curbweight	0.000863	2.341868	-0.000159
enginesize	0.010825	49.410891	0.000274
horsepower	0.013530	56.182058	0.003425
horse_per_weight	39.666768	201165.156382	23.329061
constant	-18.187248	-39378.332031	12.198230

SIMCA-P

Coefficient List [M1]

	1	2	3	4
1	Var ID (Primary)	M1.Coeff[3](conscity)	M1.Coeff[3](price)	M1.Coeff[3](symboling)
2	$constant	-18.187222	-39379.011719	12.198215
3	diesel	-0.946268	4688.631836	0.208090
4	twodoors	0.450888	-393.075836	0.337465
5	sportsstyl	0.129203	-4734.107422	-0.081705
6	wheelbase	0.040765	-23.277510	-0.035576
7	length	0.036369	86.826599	-0.009458
8	width	0.192261	596.485535	-0.030635
9	height	-0.029355	-454.787842	-0.093727
10	curbweight	0.000863	2.341853	-0.000159
11	enginesize	0.010825	49.410564	0.000274
12	horsepower	0.013530	56.182014	0.003425
13	horse_per_	39.666752	201165.500000	23.329159

Figure 10 – Unstandardized Regression Coefficients

Because, the variables are not in the same unit, we cannot use these coefficients to compare the relative influence of the predictors in the prediction.

SIMCA-P: Menu ANALYSIS / COEFFICIENTS / LIST. Select the UNSCALED coefficients.

4.4.8 Standardized Regression Coefficients

This table provides the standardized regression coefficients (Figure 11). We can use them to compare the relative importance of the variables for the prediction.

TANAGRA

Standardized Regression Parameters (by target variable)

-	Target Variable(s)				
-	Average		10.0517	14579.5000	1.0000
-	Std.Deviation		2.8730	9231.7001	1.1698
Input	Average	Std.Deviation	conscity	price	symboling
diesel	0.1500	0.3663	-0.120663	0.186062	0.065167
twodoors	0.3500	0.4894	0.076801	-0.020838	0.141172
sportsstyle	0.2000	0.4104	0.018456	-0.210452	-0.028662
wheelbase	99.5450	6.7530	0.095818	-0.017027	-0.205370
length	174.6500	11.5355	0.146026	0.108493	-0.093263
width	65.8000	2.1121	0.141344	0.136468	-0.055314
height	54.0000	2.2480	-0.022970	-0.110748	-0.180118
curbweight	2570.6500	572.3690	0.171900	0.145197	-0.077838
enginesize	134.5500	47.9215	0.180569	0.256491	0.011226
horsepower	106.1500	43.0034	0.202525	0.261709	0.125914
horse_per_weight	0.0406	0.0114	0.157935	0.249261	0.228124

SIMCA-P

Coefficient List [M1]

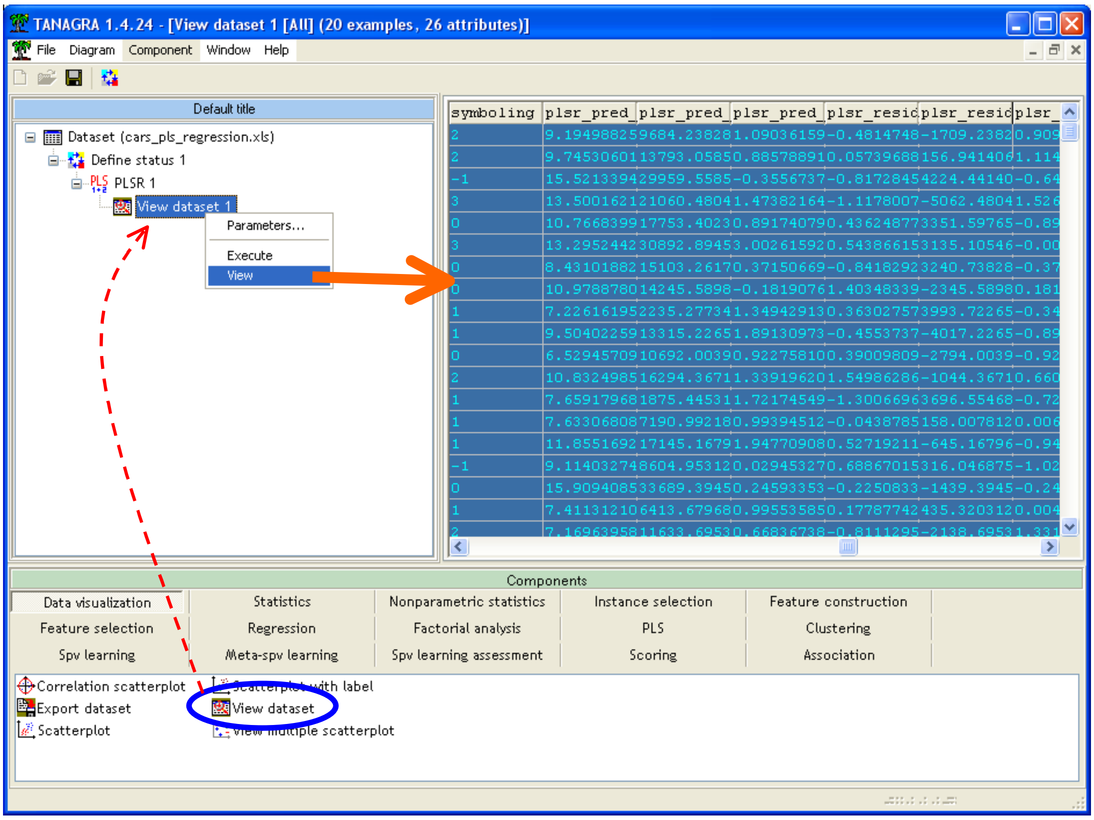
	1	2	3	4
1	Var ID (Primary)	M1.CoeffCS[3](conscity)	M1.CoeffCS[3](price)	M1.CoeffCS[3](symboling)
2	diesel	-0.120663	0.186062	0.065168
3	twodoors	0.076801	-0.020836	0.141172
4	sportsstyl	0.018456	-0.210453	-0.028664
5	wheelbase	0.095818	-0.017027	-0.205371
6	length	0.146026	0.108494	-0.093263
7	width	0.141344	0.136470	-0.055314
8	height	-0.022970	-0.110747	-0.180118
9	curbweight	0.171900	0.145196	-0.077838
10	enginesize	0.180569	0.256489	0.011226
11	horsepower	0.202525	0.261709	0.125914
12	horse_per_	0.157935	0.249262	0.228125

Figure 11 – Standardized Regression Coefficients

SIMCA-P: Menu ANALYSIS / COEFFICIENTS / LIST. Select SCALED & CENTERED coefficients.

4.4.9 Predictions and residuals

When we have set the parameters of PLSR, we have selected all the options of the OUTPUT tab (section 4.3). In this case, PLSR provides various new data columns for the subsequent branches of the diagram. To visualize them, we use the VIEW DATASET (DATA VISUALIZATION tab) component. Into the visualization grid, we observe the original dataset, and the new columns provided by the model: the factors scores, the PLS responses, the predictions and the residuals.



About the prediction, two columns are generated: one for the predicted values, the other for the residuals (PLSR_PRED_TARGET_VARIABLE_NAME and PLSR_RESIDUAL_VARIABLE_NAME) (Figure 12). These values are computed on the learning set i.e. the selected instances for the modelization. But, it can also be computed for the test set i.e. the instances which are not used during the modelization process. We can copy the values from the visualization grid to a spreadsheet application for subsequent calculations or simply for a better display.

We observe the same value (CONSCITY) with SIMCA-P (col. 4, prediction; col. 2, residual) (Figure 13). This last one uses a separated presentation for each dependent variable. We note that SIMCA-P provides also the variance of the prediction. It can be useful when we want to compute the confidence interval of the prediction.

	Prédiction			Résidus		
examples	ed_conscity_1	_pred_price_1	_symboling_1	al_conscity_1	sidual_price_1	_symboling_1
1	9.195	9684.238	1.090	-0.481	-1709.238	0.910
2	9.745	13793.059	0.886	0.057	156.941	1.114
3	15.521	29959.559	-0.356	-0.817	4224.441	-0.644
4	13.500	21060.480	1.474	-1.118	-5062.480	1.526
5	10.767	17753.402	0.892	0.436	3351.598	-0.892
6	13.295	30892.895	3.003	0.544	3135.105	-0.003
7	8.431	15103.262	0.372	-0.842	3240.738	-0.372
8	10.979	14245.590	-0.182	1.403	-2345.590	0.182
9	7.226	2235.277	1.349	0.363	3993.723	-0.349
10	9.504	13315.227	1.891	-0.455	-4017.227	-0.891
11	6.529	10692.004	0.923	0.390	-2794.004	-0.923
12	10.832	16294.367	1.339	1.550	-1044.367	0.661
13	7.659	1875.445	1.722	-1.301	3696.555	-0.722
14	7.633	7190.992	0.994	-0.044	158.008	0.006
15	11.855	17145.168	1.948	0.527	-645.168	-0.948
16	9.114	8604.953	0.029	0.689	316.047	-1.029
17	15.909	33689.395	0.246	-0.225	-1439.395	-0.246
18	7.411	6413.680	0.996	0.178	435.320	0.004
19	7.170	11633.695	0.668	-0.811	-2138.695	1.332
20	8.757	10007.340	0.716	-0.043	-1512.340	1.284

Figure 12 – Tanagra, predictions and residuals

Residuals List [M1]

	1	2	3	4
1	Obs ID (Primary)	Observed-Pred.	M1.YVar(conscity)	M1.YPred[3](conscity)
2	1	-0.481	8.714	9.195
3	2	0.057	9.803	9.745
4	3	-0.817	14.704	15.521
5	4	-1.118	12.382	13.500
6	5	0.436	11.203	10.767
7	6	0.544	13.839	13.295
8	7	-0.842	7.589	8.431
9	8	1.403	12.382	10.979
10	9	0.363	7.589	7.226
11	10	-0.455	9.049	9.504
12	11	0.390	6.920	6.529
13	12	1.550	12.382	10.832
14	13	-1.301	6.359	7.659
15	14	-0.044	7.589	7.633
16	15	0.527	12.382	11.855
17	16	0.689	9.803	9.114
18	17	-0.225	15.684	15.909
19	18	0.178	7.589	7.411
20	19	-0.811	6.359	7.170
21	20	-0.043	8.714	8.757
22		RMSEE	0.844	

Figure 13 - SIMCA-P – Residuals, variances of prediction, predictions

4.4.10 Factor scores for predictors (Scores X, Vecteur « t_h »)

TANAGRA

examples	plsr_t_1_1	plsr_t_2_1	plsr_t_3_1
1	-0.8663	0.2919	0.5479
2	-0.1672	-0.3067	-0.0102
3	4.9653	-1.9137	0.4263
4	2.1854	1.9583	0.8166
5	0.6839	-0.2210	-0.2338
6	1.8294	4.1295	-1.9338
7	-0.5753	-2.1501	-1.1698
8	1.1884	-2.0594	0.9338
9	-2.6462	0.6861	0.8899
10	-0.9021	1.7776	-0.2343
11	-2.2872	-1.4091	-1.5550
12	0.3808	0.9137	0.0872
13	-2.6206	1.7195	1.2210
14	-1.8943	-0.3876	0.0527
15	0.6998	2.6030	0.5076
16	-0.3893	-1.9024	0.9250
17	5.0939	-0.8270	-0.2904
18	-2.0754	-0.4084	0.0810
19	-1.7067	-1.7503	-1.2840
20	-0.8963	-0.7439	0.2222

SIMCA-P

General List [M1]

	1	2	3	4
1	Obs ID (Primary)	M1.t[1]	M1.t[2]	M1.t[3]
2	1	-0.8663	0.2919	-0.5479
3	2	-0.1672	-0.3067	0.0102
4	3	4.9653	-1.9137	-0.4263
5	4	2.1854	1.9583	-0.8166
6	5	0.6839	-0.2210	0.2338
7	6	1.8294	4.1295	1.9338
8	7	-0.5753	-2.1501	1.1698
9	8	1.1884	-2.0594	-0.9338
10	9	-2.6462	0.6861	-0.8899
11	10	-0.9021	1.7776	0.2343
12	11	-2.2872	-1.4091	1.5550
13	12	0.3808	0.9137	-0.0872
14	13	-2.6206	1.7195	-1.2210
15	14	-1.8943	-0.3876	-0.0527
16	15	0.6998	2.6030	-0.5076
17	16	-0.3893	-1.9024	-0.9250
18	17	5.0939	-0.8270	0.2904
19	18	-2.0754	-0.4084	-0.0810
20	19	-1.7067	-1.7503	1.2840
21	20	-0.8963	-0.7439	-0.2222

Figure 14 - " t_h " vectors for Tanagra and SIMCA-P

SIMCA-P: Menu ANALYSIS / SCORES / LINE PLOT, « t » series.

"SCORES X" are the factors scores computed from the predictors. The used formula is

$$t_h = X \times w_h *$$

The projection of the instances in this new representation space enables to better understand some special cases (outliers) or to detect possible groups.

4.4.11 PLS Responses for dependent variables (Scores Y, « $\tilde{u}_h$ » vectors)

"SCORES Y" are the PLS Responses scores computed from the dependent variables

$$\tilde{u}_h = Y \times c_h$$

TANAGRA – Vecteur « $\tilde{u}$»

examples	plsr_utilde_1_1	plsr_utilde_2_1	plsr_utilde_3_1
1	-1.5986	1.3483	0.5151
2	-0.3790	1.9482	-0.7250
3	4.8511	-1.8419	-2.2595
4	0.7953	4.7797	-1.7224
5	1.5083	-1.4015	-0.5196
6	3.6883	6.1095	-5.6858
7	-0.3933	-2.4954	-0.3230
8	0.8585	-1.5756	1.7366
9	-2.1089	-1.0852	1.6193
10	-1.0954	-0.5408	1.0912
11	-1.9944	-3.2200	1.9944
12	0.8940	2.6938	-0.7113
13	-2.7209	-1.4428	1.6212
14	-1.9682	-1.0264	1.3621
15	1.2440	0.7128	-0.1589
16	-0.4328	-4.4560	2.9478
17	4.8369	0.3605	-2.5362
18	-2.0310	-1.0527	1.4769
19	-2.4206	0.8099	-0.1189
20	-1.5333	1.3757	0.3957

SIMCA-P – Vecteur « u »

General List [M1]

	1	2	3	4
1	Obs ID (Primary)	M1.u[1]	M1.u[2]	M1.u[3]
2	1	-1.5986	1.6312	-0.1186
3	2	-0.3790	2.0028	0.9746
4	3	4.8511	-3.4632	0.1017
5	4	0.7953	4.0662	-0.5627
6	5	1.5083	-1.6249	0.2020
7	6	3.6883	5.5121	2.5856
8	7	-0.3933	-2.3075	1.7039
9	8	0.8585	-1.9636	-1.4895
10	9	-2.1089	-0.2212	-0.3102
11	10	-1.0954	-0.2463	-1.3811
12	11	-1.9944	-2.4732	0.0918
13	12	0.8940	2.5694	0.0402
14	13	-2.7209	-0.5872	-0.8207
15	14	-1.9682	-0.4079	-0.0060
16	15	1.2440	0.4843	-1.5150
17	16	-0.4328	-4.3289	-1.7999
18	17	4.8369	-1.3028	-0.2193
19	18	-2.0310	-0.3750	0.0011
20	19	-2.4206	1.3672	2.0088
21	20	-1.5333	1.6683	0.5133

Figure 15 - " $\tilde{u}$ " vectors – Tanagra; "u" vectors - SIMCA-P

SIMCA-P: Menu ANALYSIS / SCORES / LINE PLOT, « u » series.

Except the first column, Tanagra and SIMCA-P do not provide the same values. The reason is that Tanagra computes the scores directly from the dependent variables (see the formula above). The values are easier to interpret.

4.4.12 Some charts

PLS Regression is also a factor analysis approach. We can construct various graphical representations of the individuals or the variables. They enable to better understand the associations or the contrasts between the variables and the individuals.

Variables Charts. They are based on the LOADINGS and WEIGHTS. For Tanagra, we cannot design directly the plots. But we can copy the values from the data visualization grid (COMPONENT / COPY RESULTS menu) to a spreadsheet application and construct all the charts we want.

Below, we display the SIMCA-P chart from Wh* and Ch for the first two factors. The graphical representation is especially interesting when we have a large number of variables.

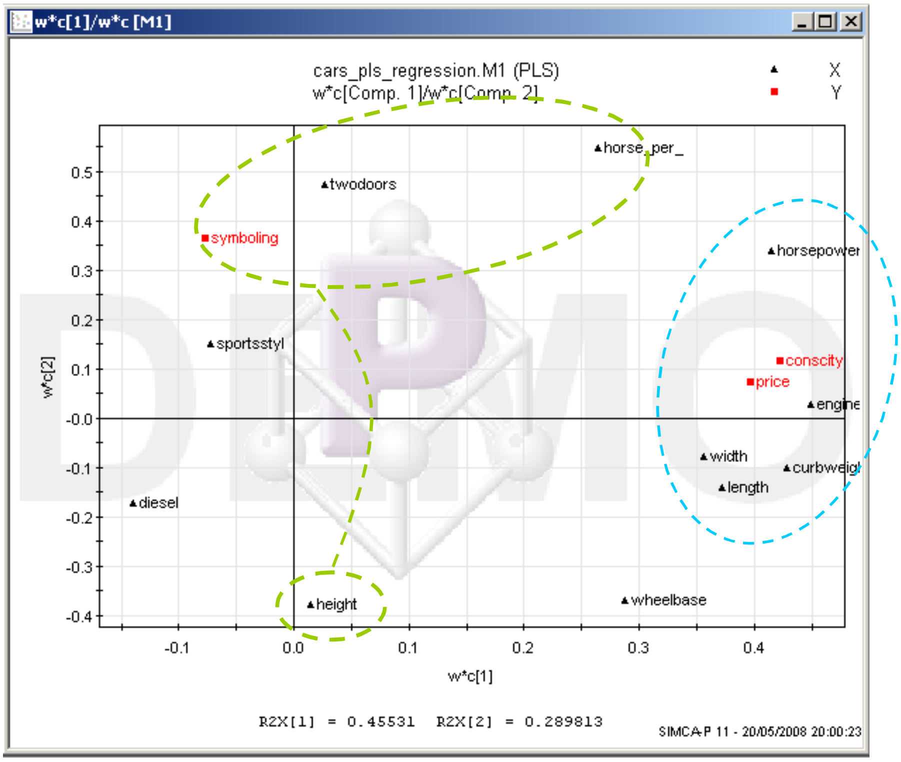
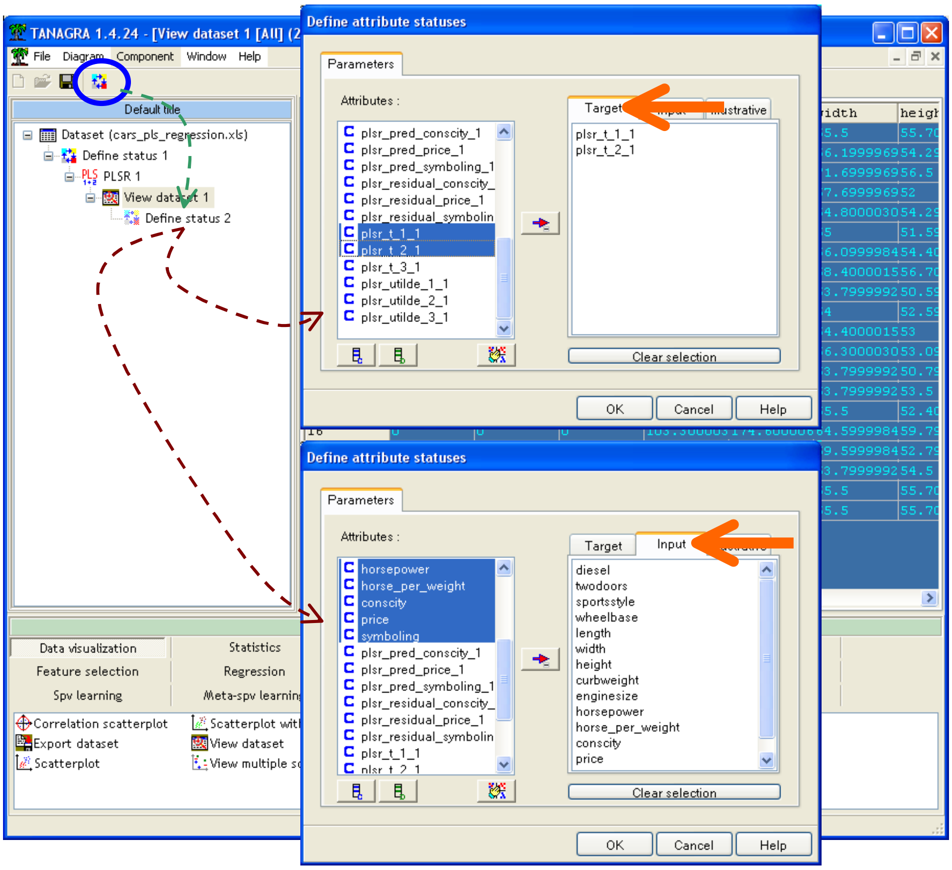


Figure 16 – Variables chart - SIMCA-P - w*c[1] vs. w*c[2]

Correlations between variables and factors. Because Tanagra provides automatically the factor scores, we can calculate explicitly the correlations between the factors and the variables (target and/or input). We add the DEFINE STATUS component into the diagram. We set as TARGET the two first factors. We set as INPUT all the variables of the dataset (predictors and dependent variables) (DIESEL…SYMBOLING).

Note: We can add to the INPUT ones other variables which are not used during the modelization phase. It can be useful when we want to study the behavior of illustrative variables, which are not directly related to the analysis.



Then we add the CORRELATION SCATTERPLOT component (DATA VISUALIZATION tab) (Figure 17).

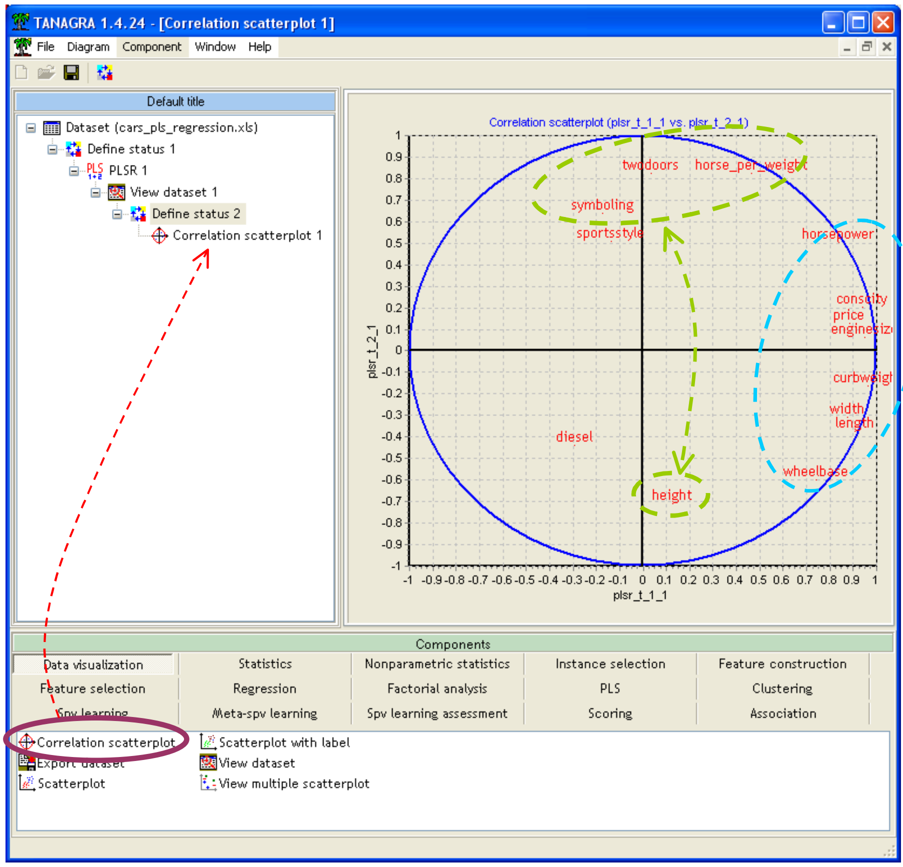


Figure 17 – Correlations between factors and variables – Two first factors

We observe that the relative coordinates between the variables (Figure 17) is very similar the ones in the variables charts based on Wh* and Ch vectors (Figure 16). Both enable to understand the associations between the variables.

Individuals representation. A very informative aspect of factorial methods is the ability to position the instances in order to analyze proximities or to delimit groups. To create scatter plots, we add the SCATTERPLOT WITH LABEL component (DATA VISUALIZATION tab) into the diagram. We can select the factors. The points are labeled with the instance number. On our dataset, we can then observe the particular influence of some instances in the plots of the factors (t1, t2) (Figure 18), or for a factor versus a PLS response (t1, $\tilde{u}_1$) (Figure 19).

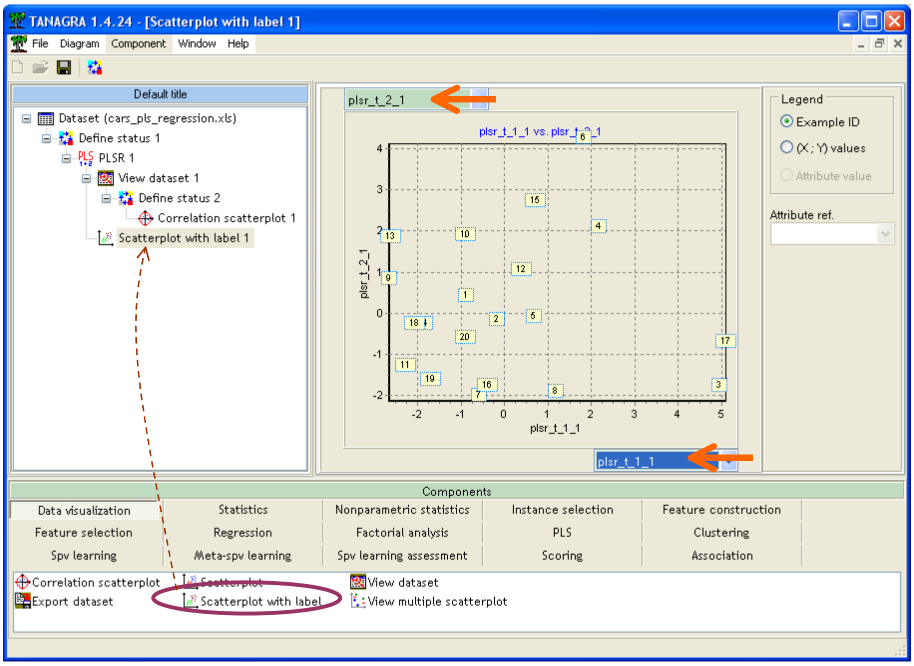


Figure 18 – Scatter plot for the factors (t1, t2)

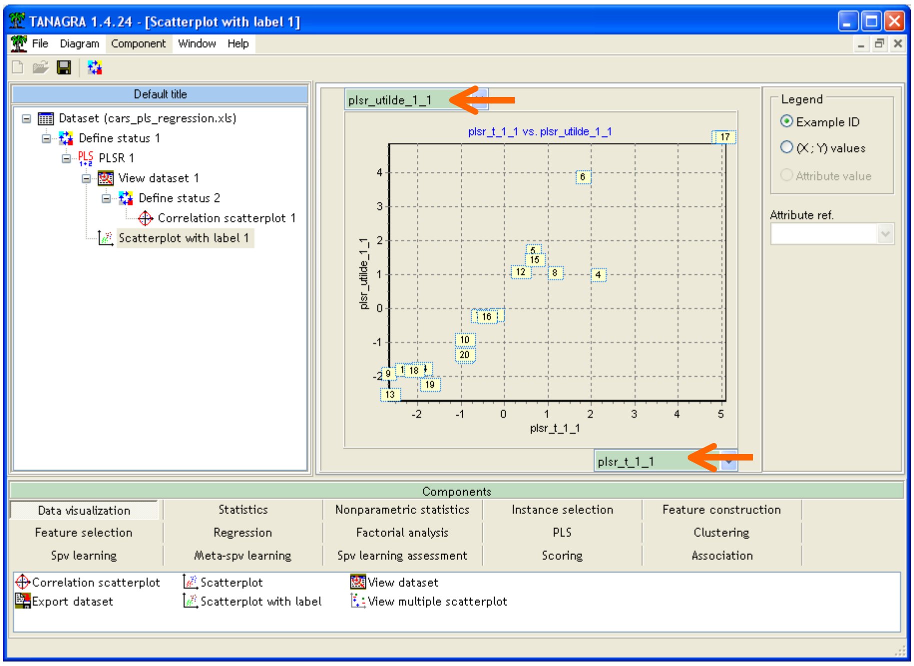


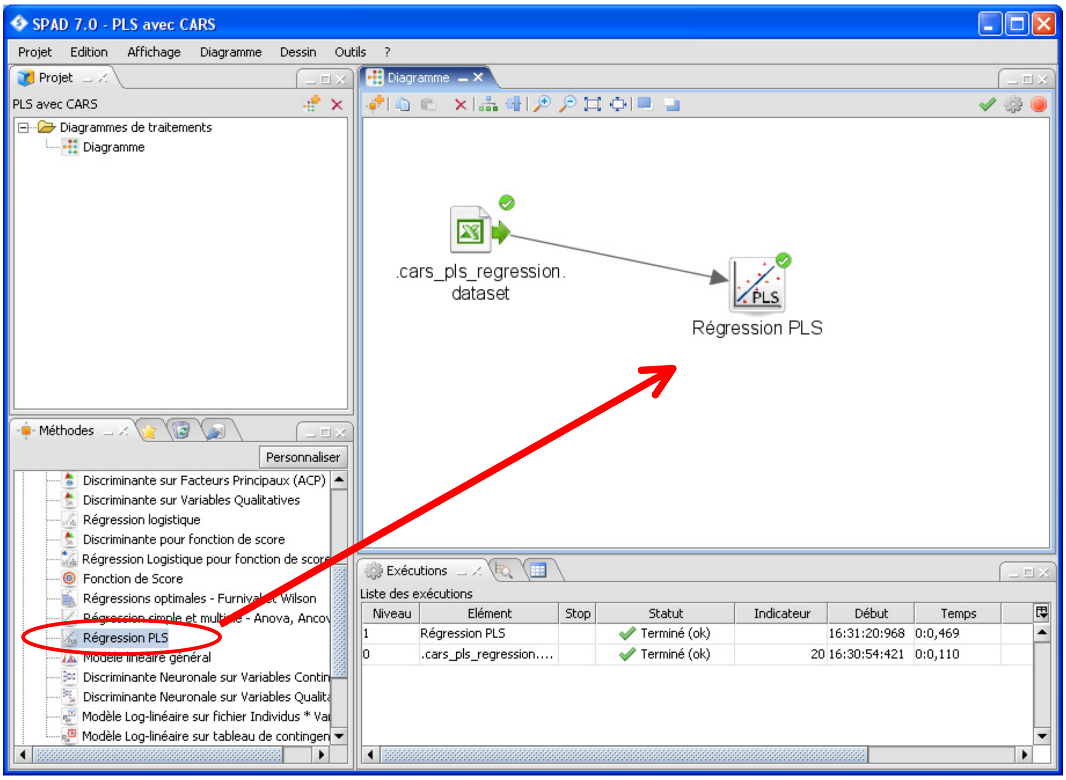
Figure 19 – Scatter plot for a factor and a PLS response (t1, $\tilde{u}_1$)

5 PLS Regression with other tools

5.1 PLS Regression with SPAD

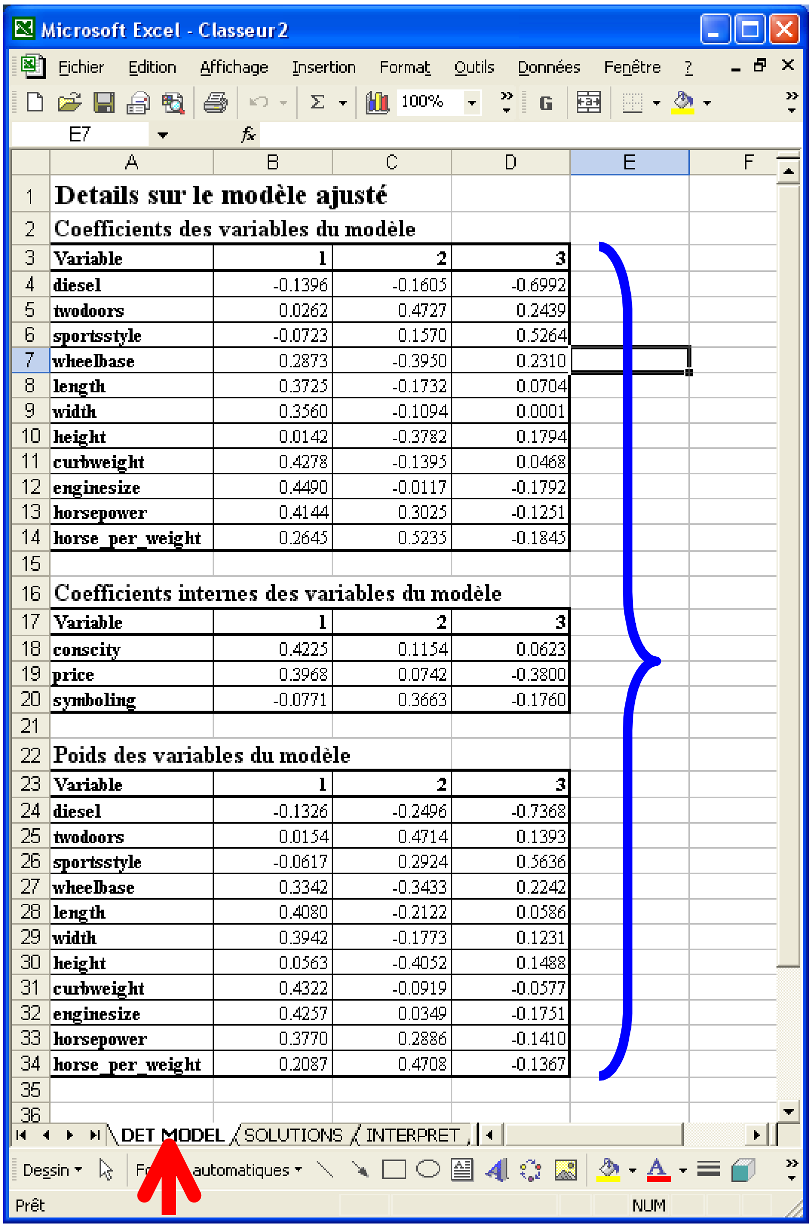
SPAD is a very popular French data mining tool which has a long history. It was one of the first French tools which provided algorithms for exploratory data analysis for personal computer in 70's.

Spad 7.0 incorporates a PLS Regression component (http://www.coheris.fr/en/page/produits/SPAD-data-mining.html). We create a new diagram. Then we import the data file. Last we add the PLS Regression component. We see below the main window of the software.



Several tables are generated. They are automatically loaded in the Excel spreadsheet application. The results are consistent with those computed from Tanagra or SIMCA-P.

Into the **DET MODEL** sheet, we observe: the WEIGHTS of the predictors ("Coefficients des variables du modèle", Wh vector, see Figure 8); the WEIGHTS of the dependent variables ("Coefficients internes des variables du modèle", Ch vector, see Figure 7); the LOADINGS of the predictors ("Poids des variables du modèle", Ph vector, see Figure 6).



Details sur le modèle ajusté

Coefficients des variables du modèle

Variable	1	2	3
diesel	-0.1396	-0.1605	-0.6992
twodoors	0.0262	0.4727	0.2439
sportsstyle	-0.0723	0.1570	0.5264
wheelbase	0.2873	-0.3950	0.2310
length	0.3725	-0.1732	0.0704
width	0.3560	-0.1094	0.0001
height	0.0142	-0.3782	0.1794
curbweight	0.4278	-0.1395	0.0468
enginesize	0.4490	-0.0117	-0.1792
horsepower	0.4144	0.3025	-0.1251
horse_per_weight	0.2645	0.5235	-0.1845

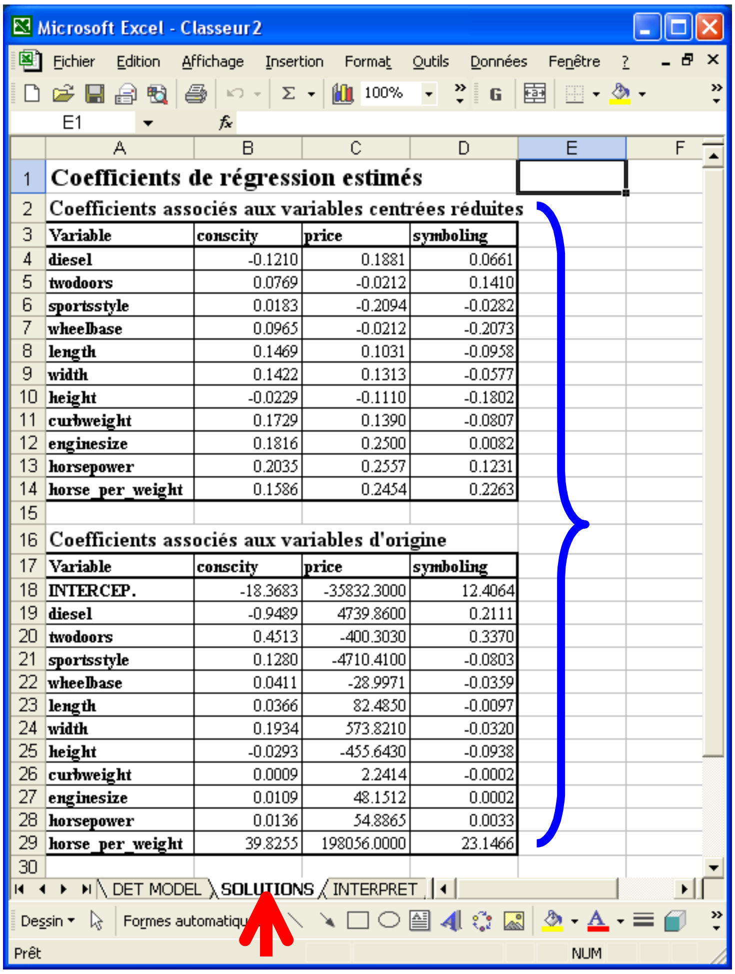
Coefficients internes des variables du modèle

Variable	1	2	3
conscity	0.4225	0.1154	0.0623
price	0.3968	0.0742	-0.3800
symboling	-0.0771	0.3663	-0.1760

Poids des variables du modèle

Variable	1	2	3
diesel	-0.1326	-0.2496	-0.7368
twodoors	0.0154	0.4714	0.1393
sportsstyle	-0.0617	0.2924	0.5636
wheelbase	0.3342	-0.3433	0.2242
length	0.4080	-0.2122	0.0586
width	0.3942	-0.1773	0.1231
height	0.0563	-0.4052	0.1488
curbweight	0.4322	-0.0919	-0.0577
enginesize	0.4257	0.0349	-0.1751
horsepower	0.3770	0.2886	-0.1410
horse_per_weight	0.2087	0.4708	-0.1367

Into the **SOLUTIONS** sheet, we observe the estimated coefficients of the model: "Coefficients associées aux variables centrées réduites" are the standardized coefficients (see Figure 11); "Coefficients associés aux variables d'origine" are the unstandardized coefficients (see Figure 10).



Coefficients de régression estimés

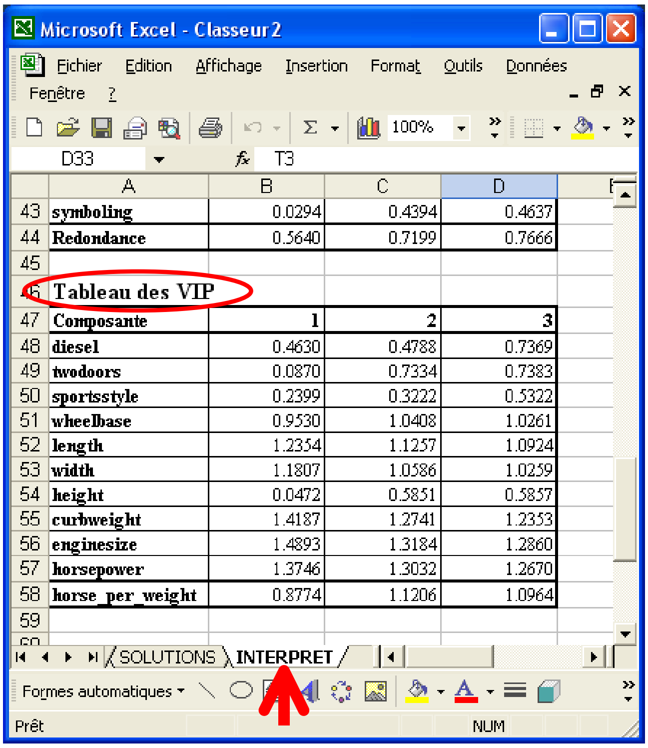
Coefficients associés aux variables centrées réduites

Variable	conscity	price	symboling
diesel	-0.1210	0.1881	0.0661
twodoors	0.0769	-0.0212	0.1410
sportsstyle	0.0183	-0.2094	-0.0282
wheelbase	0.0965	-0.0212	-0.2073
length	0.1469	0.1031	-0.0958
width	0.1422	0.1313	-0.0577
height	-0.0229	-0.1110	-0.1802
curbweight	0.1729	0.1390	-0.0807
enginesize	0.1816	0.2500	0.0082
horsepower	0.2035	0.2557	0.1231
horse_per_weight	0.1586	0.2454	0.2263

Coefficients associés aux variables d'origine

Variable	conscity	price	symboling
INTERCEP.	-18.3683	-35832.3000	12.4064
diesel	-0.9489	4739.8600	0.2111
twodoors	0.4513	-400.3030	0.3370
sportsstyle	0.1280	-4710.4100	-0.0803
wheelbase	0.0411	-28.9971	-0.0359
length	0.0366	82.4850	-0.0097
width	0.1934	573.8210	-0.0320
height	-0.0293	-455.6430	-0.0938
curbweight	0.0009	2.2414	-0.0002
enginesize	0.0109	48.1512	0.0002
horsepower	0.0136	54.8865	0.0033
horse_per_weight	39.8255	198056.0000	23.1466

Into the **INTERPRET** sheet, we observe various tables which are useful for the interpretation, for instance we observe below the variable importance in projection table (VIP, see Figure 9).



symboling	0.0294	0.4394	0.4637
Redondance	0.5640	0.7199	0.7666

Tableau des VIP

Composante	1	2	3
diesel	0.4630	0.4788	0.7369
twodoors	0.0870	0.7334	0.7383
sportsstyle	0.2399	0.3222	0.5322
wheelbase	0.9530	1.0408	1.0261
length	1.2354	1.1257	1.0924
width	1.1807	1.0586	1.0259
height	0.0472	0.5851	0.5857
curbweight	1.4187	1.2741	1.2353
enginesize	1.4893	1.3184	1.2860
horsepower	1.3746	1.3032	1.2670
horse_per_weight	0.8774	1.1206	1.0964

5.2 PLS Regression with SAS

We have used the PROC PLS procedure of SAS 9.1. The description of the procedure is available online (http://support.sas.com/91doc/docMainpage.jsp; search: PROC PLS).

The PROC PLS is very informative. To obtain results comparable to the other tools, we set the following settings: METHOD = PLS and ALGORITHM = NIPALS. We use the following commands for our dataset.

```
%let cost = conscity price symboling;
%let charac = diesel twodoors sportsstyle wheelbase length width height curbweight
enginesize horsepower horse_per_weight;
proc pls data=cars method=pls (algorithm=nipals) details nfac=3;
model &cost=&charac;
output out=pls         predicted=predY1-predY3
                       yresidual=resY1-resY3
                       xscore=xsc1-xsc3
                       yscore=ysc1-ysc3;
run;
```

First, we obtain the proportion of explained variance by the factors, for the predictors and for the dependent variables. The results are coherent with those of Tanagra and SIMCA-P (see Figure 2 and Figure 3).

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors

Number of Extracted Factors	Model Effects Current	Model Effects Total	Dependent Variables Current	Dependent Variables Total
1	45.5310	45.5310	56.4009	56.4009
2	28.9813	74.5123	15.5845	71.9855
3	7.4150	81.9273	4.6715	76.6570

Then, we have the LOADINGS (see Figure 6). The organization of the values is a little different.

Model Effect Loadings

Number of Extracted Factors	diesel	twodoors	sportsstyle	wheelbase	length	width
1	-0.131855	0.015262	-0.061349	0.332197	0.405519	0.391885
2	-0.244389	0.461515	0.286265	-0.336150	-0.207791	-0.173595
3	-0.721275	0.136361	0.551743	0.219469	0.057399	0.120507

Number of Extracted Factors	height	curbweight	enginesize	horsepower	horse_per_weight
1	0.055946	0.429607	0.423140	0.374797	0.207439
2	-0.396712	-0.090016	0.034193	0.282591	0.460990
3	0.145641	-0.056496	-0.171430	-0.138054	-0.133844

We note that the values seem not consistent with those of our reference tools (Tanagra and SIMCA-P). This is the same case for the WEIGTHS below (in comparison to the values into the Figure 7).

Dependent Variable Weights

Number of Extracted Factors	conscity	price	symboling
1	0.722639	0.678529	-0.131878
2	0.295016	0.189718	0.936468
3	0.147212	-0.897485	-0.415751

The differences are explained by the normalization process used by SAS. Indeed, if we calculate the following formula for the first factor: $0.722369^2 + 0.678529^2 + (-0.131878)^2 = 1$. If we normalize the values provided by Tanagra and SIMCA-P, we obtain the same coefficients.

5.3 PLS Regression with R – The PLS package

We used the 2.6.0 version of R (http://www.r-project.org/). We installed and implemented the PLS package (http://cran.r-project.org/web/packages/pls/index.html). The main reference is the paper published into the "Journal of Statistical Software" (http://www.jstatsoft.org/v18/i02). We set the following commands.

```
#clear all
rm(list=ls())

#************************
#some packages
#************************
#xls data file handling
library(xlsReadWrite)
#pls regression
library(pls)

#downloading the dataset
setwd("directory of the dataset")
cars.data <- read.xls(file="cars_pls_regression.xls",rowNames=FALSE,sheet=1)
#checking the variables
summary(cars.data)
#subdivide the dataset into matrix Y and X
Y <- as.matrix(cars.data[,12:14])
X <- as.matrix(cars.data[,1:11])

#pls regression: 3 factors, nipals
cars.pls <- mvr(Y ~ X, ncomp = 3, method = "oscorespls", scale = TRUE)
summary(cars.pls)
```

First, we obtain a summary of the main results. Here we observe divergences compared with the results of the other tools.

```
> summary(cars.pls)
Data:    X dimension: 20 11
         Y dimension: 20 3
Fit method: oscorespls
Number of components considered: 3
TRAINING: % variance explained
           1 comps  2 comps  3 comps
X           45.333    66.30    81.73      [1]
conscity    87.988    89.17    91.94
price       80.482    88.47    92.72      [2]
symboling    2.546    33.95    41.78
```

About the proportion of variances explained by the factors **[1]**, the values (45.33%, 66.30% and 81.73%) are not the same as the other tools (see for instance Figure 3) where we have (45.53%, 74.51%, 81.93%). The deviation is singularly problematic for the second factor.

If we observe the cumulative proportion of variance explained by the factors for each dependent variable **[2]**, the values are also in contradiction to the other tools (see Figure 5).

We can obtain various indicators (loadings, estimated parameters, residuals, etc.) with the following commands.

```
#loadings for X
loadings(cars.pls)

#weights for X
loading.weights(cars.pls)

#weights for Y
Yloadings(cars.pls)

#regression coefficients
coef(cars.pls)

#prediction
fitted(cars.pls)

#residuals
residuals(cars.pls)
```

About the estimated standardized regression coefficients, we have (see Figure 11).

```
> coef(cars.pls)
, , 3 comps

                    conscity       price    symboling
diesel           -0.26017692  1557.55068  0.002923431
twodoors          0.24626479  -889.76672  0.149642873
sportsstyle       0.06821250 -1820.00863  0.057453989
wheelbase         0.27648511   -34.11222 -0.186698496
length            0.44138223   462.08244 -0.119974109
width             0.41972747   578.93918 -0.115984611
height           -0.09896129 -1246.64032 -0.203604184
curbweight        0.50771044  1634.46156 -0.045387443
enginesize        0.51005416  3122.95037  0.029842667
horsepower        0.57798583  2488.51319  0.121556434
horse_per_weight  0.43161789  2200.81123  0.198713642
```

I think that the divergence with the other tools is mainly caused by a different normalization mechanism. This is highlighted by the authors in their paper (http://www.jstatsoft.org/v18/i02/paper; last paragraph, page 3). This makes comparisons difficult.

6 Conclusion

In this tutorial, we introduced the PLSR component of Tanagra (from 1.4.24 version). We have mainly improved the presentation of the results in order to be comparable to the state-of-the-art tools such as SIMCA-P or SAS. We note that the most of the tools provide the same results when they are applied to the same dataset.