1. Topic

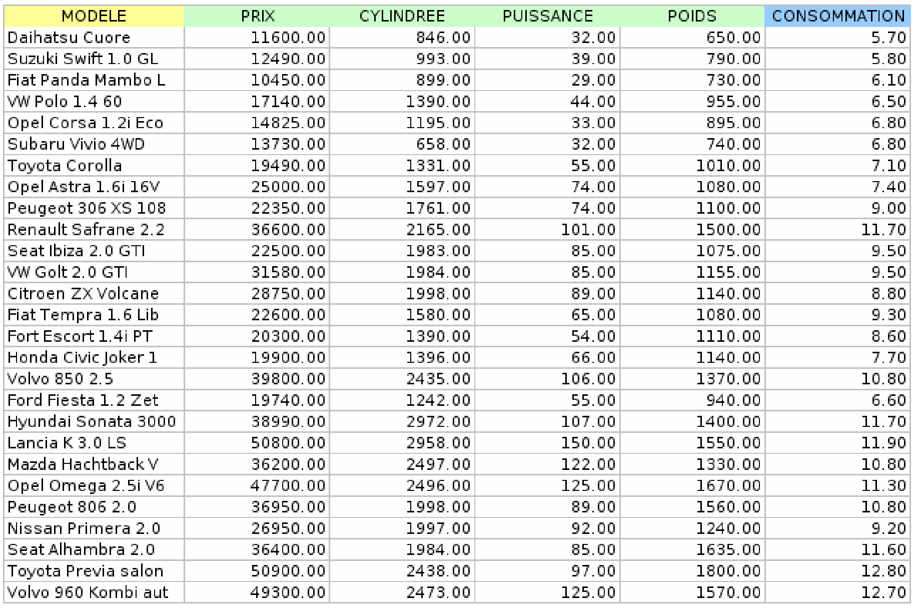
Dealing with multicollinearity in multiple regression.

Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. In this situation the coefficient estimates may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole; it only affects calculations regarding individual predictors. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with others (Wikipedia - http://en.wikipedia.org/wiki/Multicollinearity). Sometimes the signs of the coefficients are inconsistent with the domain knowledge; sometimes, explanatory variables which seems individually significant are invalidated when we add other variables.

There are two steps when we want to treat this kind of problem: (1) detecting the presence of the collinearity; (2) implementing solutions in order to obtain more consistent results.

In this tutorial, we study three approaches to avoid the multicollinearity problem: the variable selection; the regression on the latent variables provided by PCA (principal component analysis); the PLS regression (partial least squares).

2. Dataset



MODELE	PRIX	CYLINDREE	PUISSANCE	POIDS	CONSOMMATION
Daihatsu Cuore	11600.00	846.00	32.00	650.00	5.70
Suzuki Swift 1.0 GL	12490.00	993.00	39.00	790.00	5.80
Fiat Panda Mambo L	10450.00	899.00	29.00	730.00	6.10
VW Polo 1.4 60	17140.00	1390.00	44.00	955.00	6.50
Opel Corsa 1.2i Eco	14825.00	1195.00	33.00	895.00	6.80
Subaru Vivio 4WD	13730.00	658.00	32.00	740.00	6.80
Toyota Corolla	19490.00	1331.00	55.00	1010.00	7.10
Opel Astra 1.6i 16V	25000.00	1597.00	74.00	1080.00	7.40
Peugeot 306 XS 108	22350.00	1761.00	74.00	1100.00	9.00
Renault Safrane 2.2	36600.00	2165.00	101.00	1500.00	11.70
Seat Ibiza 2.0 GTI	22500.00	1983.00	85.00	1075.00	9.50
VW Golt 2.0 GTI	31580.00	1984.00	85.00	1155.00	9.50
Citroen ZX Volcane	28750.00	1998.00	89.00	1140.00	8.80
Fiat Tempra 1.6 Lib	22600.00	1580.00	65.00	1080.00	9.30
Fort Escort 1.4i PT	20300.00	1390.00	54.00	1110.00	8.60
Honda Civic Joker 1	19900.00	1396.00	66.00	1140.00	7.70
Volvo 850 2.5	39800.00	2435.00	106.00	1370.00	10.80
Ford Fiesta 1.2 Zet	19740.00	1242.00	55.00	940.00	6.60
Hyundai Sonata 3000	38990.00	2972.00	107.00	1400.00	11.70
Lancia K 3.0 LS	50800.00	2958.00	150.00	1550.00	11.90
Mazda Hachtback V	36200.00	2497.00	122.00	1330.00	10.80
Opel Omega 2.5i V6	47700.00	2496.00	125.00	1670.00	11.30
Peugeot 806 2.0	36950.00	1998.00	89.00	1560.00	10.80
Nissan Primera 2.0	26950.00	1997.00	92.00	1240.00	9.20
Seat Alhambra 2.0	36400.00	1984.00	85.00	1635.00	11.60
Toyota Previa salon	50900.00	2438.00	97.00	1800.00	12.80
Volvo 960 Kombi aut	49300.00	2473.00	125.00	1570.00	12.70

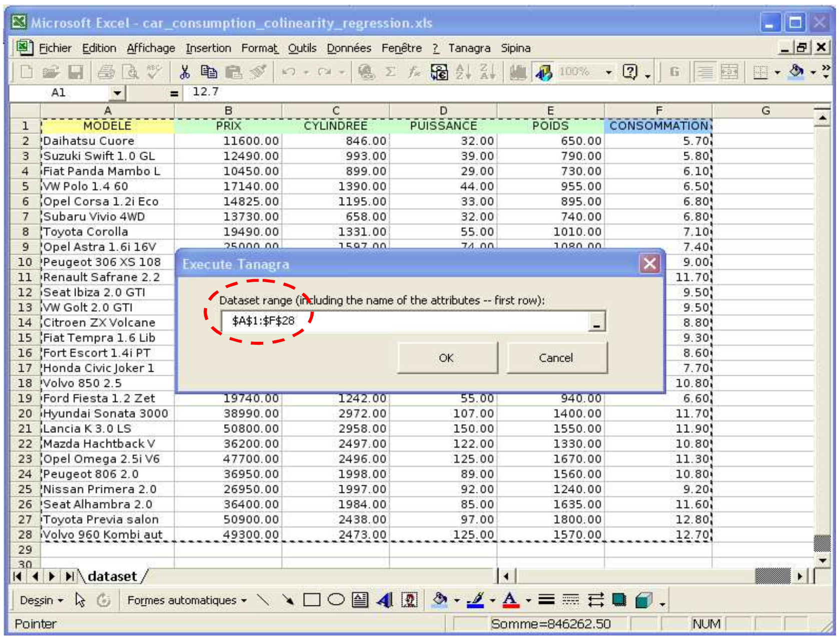
Figure 1 – Dataset - Dependent variable: "CONSOMMATION"

There 27 instances into the car_consumption_colinearity_regression.xls data file. The goal is to predict the consumption of cars (CONSOMMATION) from various characteristics (price, engine size, horsepower and weight).

3. Multiple regression and the multicollinearity problem

3.1. Creating a diagram and importing the data file

The easiest way to launch Tanagra and to import the dataset is to load the file into Excel spreadsheet. We select the dataset. We click on the TANAGRA / EXECUTE TANAGRA menu[1].


Microsoft Excel - car_consumption_colinearity_regression.xls
Fichier Edition Affichage Insertion Format Outils Données Fenêtre ? Tanagra Sipina
A1 = 12.7
dataset
Dessin Formes automatiques
Pointer Somme=846262.50 NUM

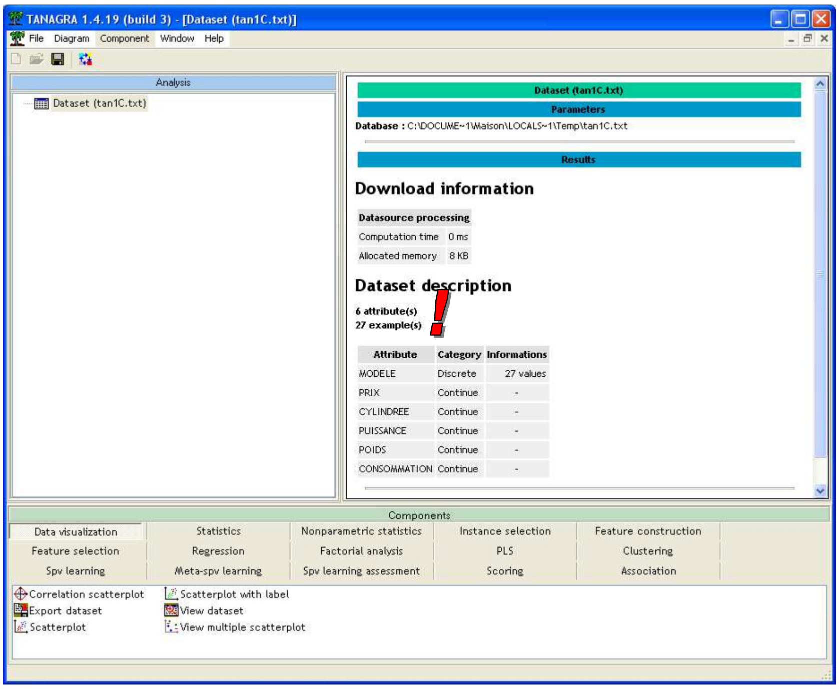
	A	B	C	D	E	F
1	MODELE	PRIX	CYLINDREE	PUISSANCE	POIDS	CONSOMMATION
2	Daihatsu Cuore	11600.00	846.00	32.00	650.00	5.70
3	Suzuki Swift 1.0 GL	12490.00	993.00	39.00	790.00	5.80
4	Fiat Panda Mambo L	10450.00	899.00	29.00	730.00	6.10
5	VW Polo 1.4 60	17140.00	1390.00	44.00	955.00	6.50
6	Opel Corsa 1.2i Eco	14825.00	1195.00	33.00	895.00	6.80
7	Subaru Vivio 4WD	13730.00	658.00	32.00	740.00	6.80
8	Toyota Corolla	19490.00	1331.00	55.00	1010.00	7.10
9	Opel Astra 1.6i 16V	25000.00	1597.00	74.00	1080.00	7.40
10	Peugeot 306 XS 108					9.00
11	Renault Safrane 2.2					11.70
12	Seat Ibiza 2.0 GTI					9.50
13	VW Golt 2.0 GTI					9.50
14	Citroen ZX Volcane					8.80
15	Fiat Tempra 1.6 Lib					9.30
16	Fort Escort 1.4i PT					8.60
17	Honda Civic Joker 1					7.70
18	Volvo 850 2.5					10.80
19	Ford Fiesta 1.2 Zet	19740.00	1242.00	55.00	940.00	6.60
20	Hyundai Sonata 3000	38990.00	2972.00	107.00	1400.00	11.70
21	Lancia K 3.0 LS	50800.00	2958.00	150.00	1550.00	11.90
22	Mazda Hachtback V	36200.00	2497.00	122.00	1330.00	10.80
23	Opel Omega 2.5i V6	47700.00	2496.00	125.00	1670.00	11.30
24	Peugeot 806 2.0	36950.00	1998.00	89.00	1560.00	10.80
25	Nissan Primera 2.0	26950.00	1997.00	92.00	1240.00	9.20
26	Seat Alhambra 2.0	36400.00	1984.00	85.00	1635.00	11.60
27	Toyota Previa salon	50900.00	2438.00	97.00	1800.00	12.80
28	Volvo 960 Kombi aut	49300.00	2473.00	125.00	1570.00	12.70

Execute Tanagra

Dataset range (including the name of the attributes -- first row): $A$1:$F$28

OK Cancel

Tanagra is launched. The dataset is now available in the root of the diagram.



Dataset (tan1C.txt)

Parameters

Database : C:\DOCUME~1\Maison\LOCALS~1\Temp\tan1C.txt

Results

Download information

Datasource processing	
Computation time	0 ms
Allocated memory	8 KB

Dataset description

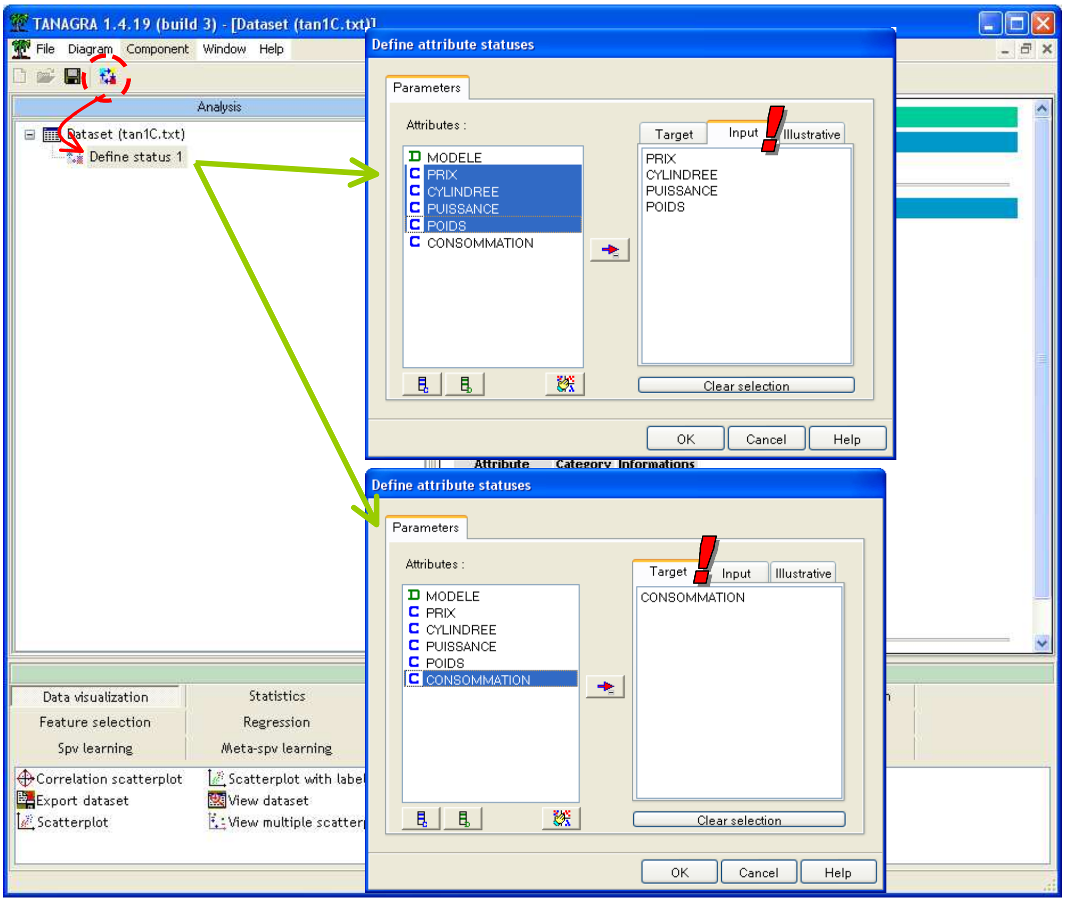
6 attribute(s)
27 example(s)

Attribute	Category	Informations
MODELE	Discrete	27 values
PRIX	Continue	-
CYLINDREE	Continue	-
PUISSANCE	Continue	-
POIDS	Continue	-
CONSOMMATION	Continue	-

[1] See the following tutorial for the installation and the utilization of the Tanagra add-on for Excel - http://data-mining-tutorials.blogspot.com/2008/10/excel-file-handling-using-add-in.html

3.2. Multiple linear regression

In a first time, we intended to perform a multiple regression analysis using all the explanatory variables. By using the shortcut into the toolbar, we insert the DEFINE STATUS component. We set CONSOMMATION as TARGET; PRIX, CYLINDREE, PUISSANCE and POIDS as INPUT.

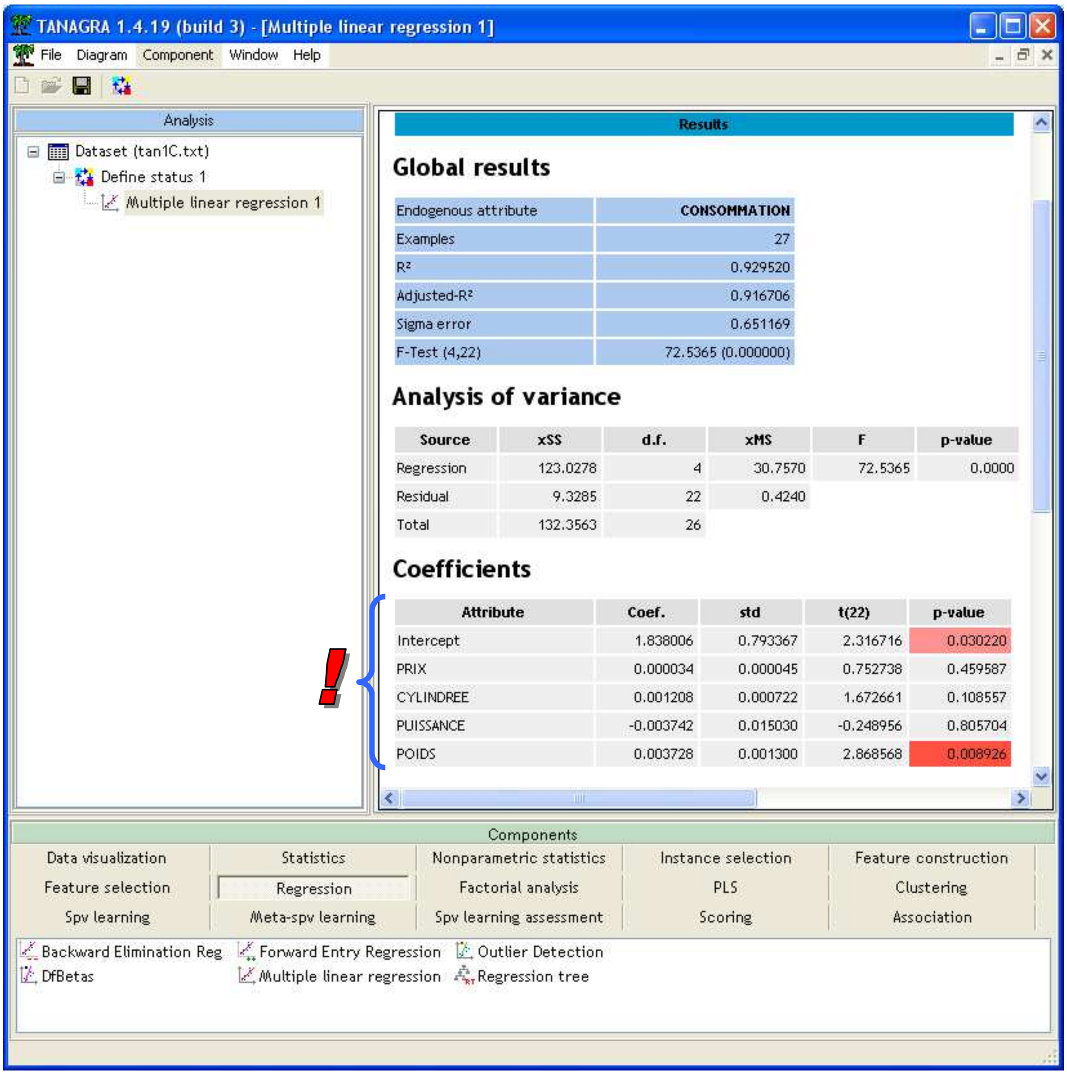


Input: PRIX, CYLINDREE, PUISSANCE, POIDS

Target: CONSOMMATION

We add the MULTIPLE LINEAR REGRESSION component (REGRESSION tab). We click on the VIEW menu to obtain the results. The model seems very good. The coefficient of determination R2 is **0.9295** (http://en.wikipedia.org/wiki/Coefficient_of_determination). We are rather confident about the quality of the model.

But, when we consider the coefficients of the model, some results seem strange. Only the weight is significant for the explanation of the consumption. The sign is positive, when the weight of the car increases, the consumption increases also. It seems natural. But, neither the horsepower nor the engine size seems to influence the consumption? It is unusual. It suggests that two cars with the same weight have similar consumption, even if the engine size of the second is 4 times bigger than the first. Although not a great expert, this last result does not correspond at all with what we know about cars.



Results

Global results

Endogenous attribute	CONSOMMATION
Examples	27
R²	0.929520
Adjusted-R²	0.916706
Sigma error	0.651169
F-Test (4,22)	72.5365 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	123.0278	4	30.7570	72.5365	0.0000
Residual	9.3285	22	0.4240		
Total	132.3563	26			

Coefficients

Attribute	Coef.	std	t(22)	p-value
Intercept	1.838006	0.793367	2.316716	0.030220
PRIX	0.000034	0.000045	0.752738	0.459587
CYLINDREE	0.001208	0.000722	1.672661	0.108557
PUISSANCE	-0.003742	0.015030	-0.248956	0.805704
POIDS	0.003728	0.001300	2.868568	0.008926

Figure 2 – Regression using all explanatory variables

3.3. Detecting the multicollinearity

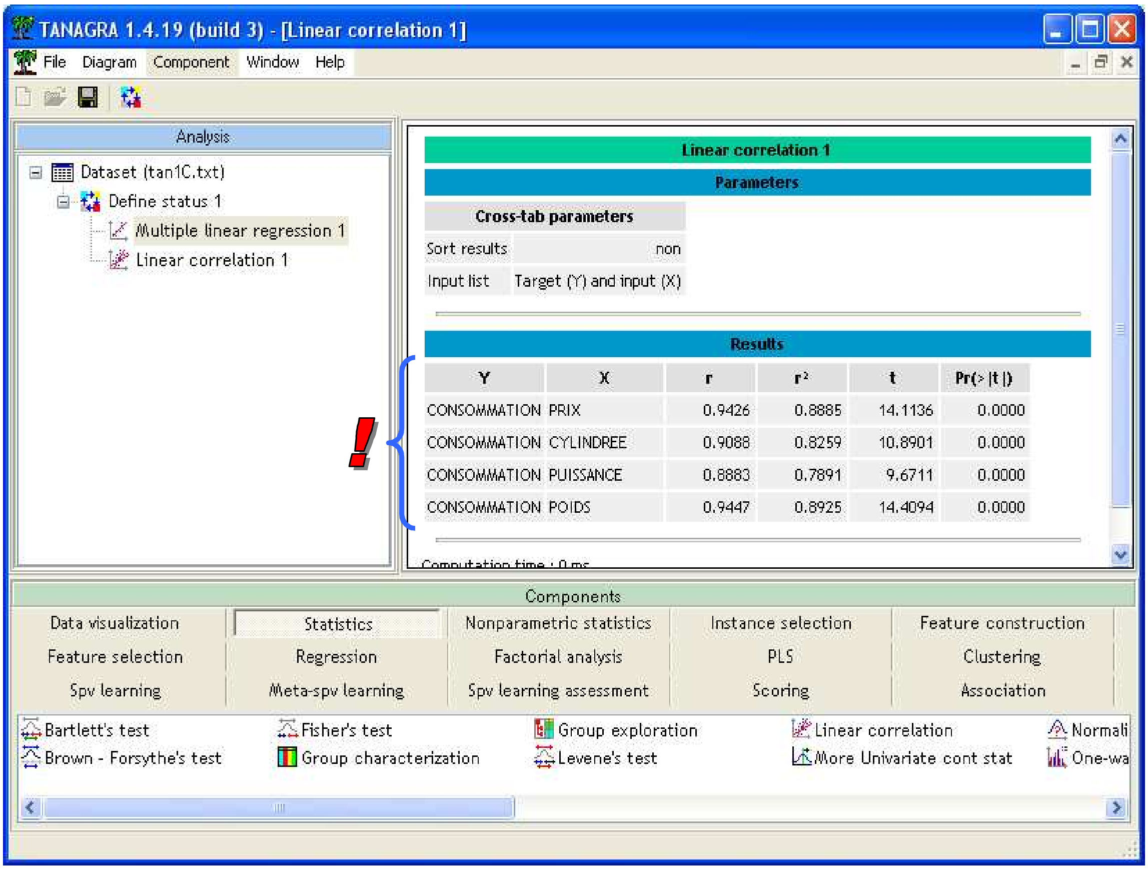
We suspect a multicollinearity phenomenon here. We know for instance that the engine size and the horsepower are often highly correlated. It influences the results in a different ways. The model is very unstable; a small change in the dataset (by removing or adding instances) causes a large modification of the estimated parameters. The sign and the values of the coefficients are inconsistent with the domain knowledge. For instance, it seems here that the horsepower has a negative influence on the consumption. We know that this cannot be true. Lastly, some variables that we know they are relevant according the domain knowledge are not significant into the regression.

In short, we have an excellent model (according the R2) but unusable because we can not draw a meaningful interpretation of the coefficients. It is impossible to understand the causal mechanism of the phenomenon studied.

We use very simple calculations to detect the multicollinearity problem.

Sign consistency. The first strategy is very basic. We check if the sign of the coefficient is consistent with the sign of the correlation of each explanatory variable with the target variable (computed individually). If some of them are inconsistent, it means that other variables interfere in the association between the explanatory variable and the dependent variable.

To compute the correlation between each independent variable and the dependent variable, we add the LINEAR CORRELATION tool (STATISTICS tab) behind DEFINE STATUS 1. We use the default settings.



Linear correlation 1

Parameters

Cross-tab parameters	
Sort results	non
Input list	Target (Y) and input (X)

Results

Y	X	r	r²	t	Pr(>|t|)
CONSOMMATION	PRIX	0.9426	0.8885	14.1136	0.0000
CONSOMMATION	CYLINDREE	0.9088	0.8259	10.8901	0.0000
CONSOMMATION	PUISSANCE	0.8883	0.7891	9.6711	0.0000
CONSOMMATION	POIDS	0.9447	0.8925	14.4094	0.0000

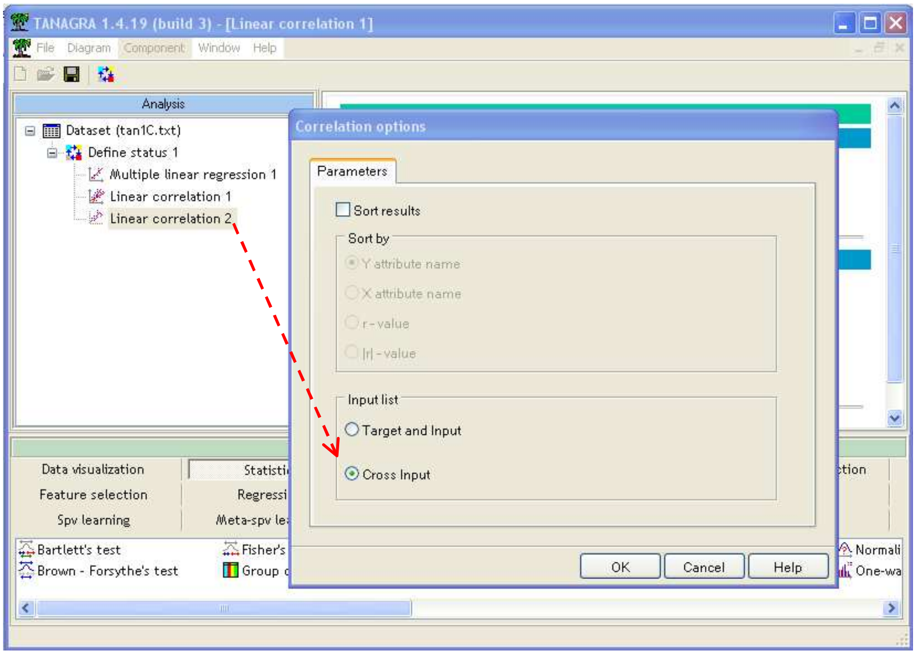
Computation time : 0 ms

Figure 3 – Correlation – Dependent vs. each independent variable

Each explanatory variable is highly correlated with the dependent variable (≥ 0.8883). We note also that there is a problem about PUISSANCE (horsepower). The correlation is positive, but the sign of its coefficient into the regression is negative. Another variable probably interferes with PUISSANCE.

Klein's rule. We compute the square of the correlation for each couple of explanatory variables. If one or more of the values are higher than (or at less near) the coefficient of determination (R2) of the regression, there is probably a multicollinearity problem. The advantage here is that we can identify the variables which are redundant in the regression.

We add again the LINEAR CORRELATION component, but we modify the settings by clicking on the PARAMETERS contextual menu. We set the CROSS INPUT option in the INPUT LIST section. Tanagra computes the correlation between the INPUT variables.



Correlation options

Parameters

☐ Sort results

Sort by

○ Y attribute name

○ X attribute name

○ r - value

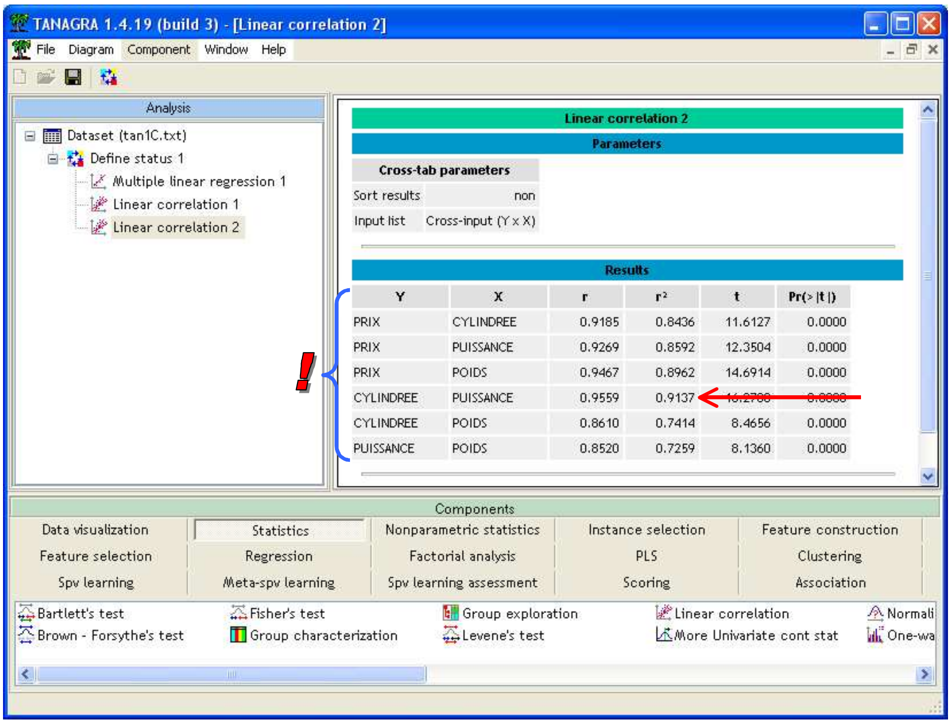
○ |r| - value

Input list

○ Target and Input

● Cross Input

We confirm the settings and we click on the VIEW menu. In the visualization window, we observe that all the explanatory variables are highly correlated each other.



Linear correlation 2

Parameters

Cross-tab parameters	
Sort results	non
Input list	Cross-input (Y x X)

Results

Y	X	r	r²	t	Pr(>|t|)
PRIX	CYLINDREE	0.9185	0.8436	11.6127	0.0000
PRIX	PUISSANCE	0.9269	0.8592	12.3504	0.0000
PRIX	POIDS	0.9467	0.8962	14.6914	0.0000
CYLINDREE	PUISSANCE	0.9559	0.9137	16.2700	0.0000
CYLINDREE	POIDS	0.8610	0.7414	8.4656	0.0000
PUISSANCE	POIDS	0.8520	0.7259	8.1360	0.0000

Figure 4 - Cross correlation between the explanatory variables

We note among others than the square of the correlation between PUISSANCE (horsepower) and CYLINDREE (engine size) is very close to the coefficient of determination of the regression (0.9137 vs. 0.9295).

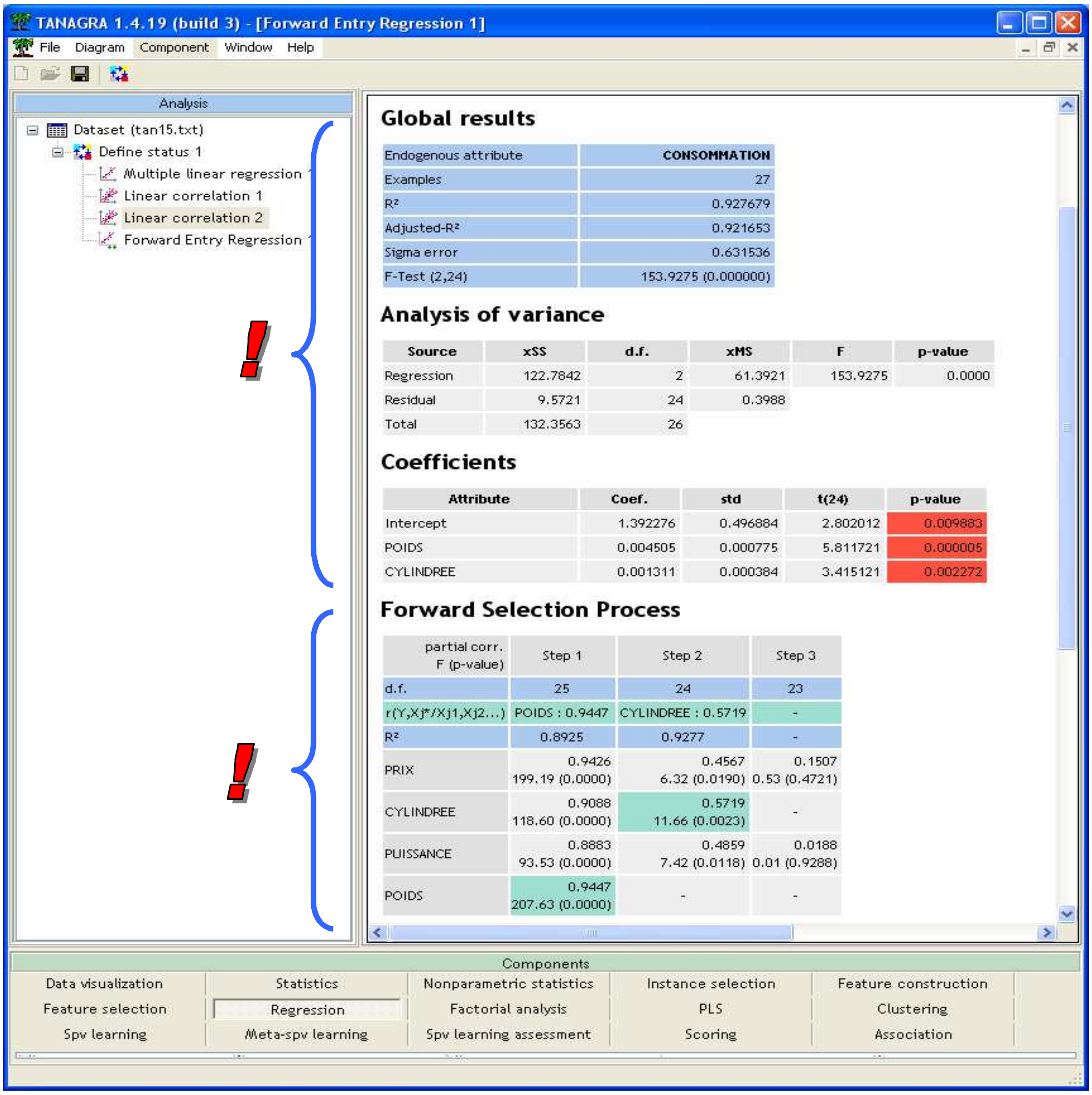
All these symptoms suggest that there is a problem of collinearity in our study. We must adopt an appropriate strategy if we want to get usable results.

4. Variable selection

The selection of explanatory variables is not really a direct solution to collinearity. Even in the absence of collinearity between independent variables, reducing the dimensionality of the problem studied is always beneficial. It helps to identify relevant variables and give an interpretable result. In the context of multicollinearity problem, it can especially remove redundant variables which interfere in the regression. This is this characteristic that is interesting here.

We use a forward search. At each step, we search the most relevant explanatory variable according the absolute value of the correlation coefficient. We must take into account the influence of already selected variables i.e. we use a partial correlation.

We add the FORWARD ENTRY REGRESSION component (REGRESSION tab) into the diagram. We click on the VIEW menu. In addition to the standard output of the regression, we obtain more detailed results about the selection process.



Global results

Endogenous attribute	CONSOMMATION
Examples	27
R²	0.927679
Adjusted-R²	0.921653
Sigma error	0.631536
F-Test (2,24)	153.9275 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	122.7842	2	61.3921	153.9275	0.0000
Residual	9.5721	24	0.3988		
Total	132.3563	26			

Coefficients

Attribute	Coef.	std	t(24)	p-value
Intercept	1.392276	0.496884	2.802012	0.009883
POIDS	0.004505	0.000775	5.811721	0.000005
CYLINDREE	0.001311	0.000384	3.415121	0.002272

Forward Selection Process

partial corr. F (p-value)	Step 1	Step 2	Step 3
d.f.	25	24	23
r(Y,Xj*/Xj1,Xj2...)	POIDS : 0.9447	CYLINDREE : 0.5719	-
R²	0.8925	0.9277	-
PRIX	0.9426 199.19 (0.0000)	0.4567 6.32 (0.0190)	0.1507 0.53 (0.4721)
CYLINDREE	0.9088 118.60 (0.0000)	0.5719 11.66 (0.0023)	-
PUISSANCE	0.8883 93.53 (0.0000)	0.4859 7.42 (0.0118)	0.0188 0.01 (0.9288)
POIDS	0.9447 207.63 (0.0000)	-	-

Figure 5 – FORWARD selection process

We observe that:

- The selected explanatory variables are POIDS (weigth) and CYLINDREE (engine size). They are very significant.
- Compared to the initial regression, despite the elimination of two variables, the proportion of explained variance remains very good with a coefficient of determination of R2 = 0.9277 (R2 = 0.9295 for the model with 4 variables).
- POIDS and CYLINDREE have both a positive influence on the consumption i.e. when the weight (or the engine size) increases, the consumption increases also. This is rather consistent to the domain knowledge.
- And the signs of the coefficients are in adequation to the sign of the correlation coefficient computed individually (Figure 3).
- When we consider the cross-correlation (Figure 4), we note that these variables are the less correlated among the explanatory variable. We have r² = 0.7417, it is largely lower than the coefficient of determination of the regression.
- About the selection steps (Figure 5), the used significance level is 5%[2]:
  - At the first step, the variable which is the mostly correlated (absolute value) to the dependent variable is POIDS (r = 0.9447). The test statistic F for the significance is 207.63. The p-value of the test is <0.0001. Because the p-value is lower than the significance level, we add the variable into the regression.
  - At the next step, we search the variable which is the most correlated with the dependent variable, by removing the effect of the already selected variables. The partial correlation[3,4] for CYLINDREE is 0.5719. We note that it is clearly lower than the direct correlation which was 0.9088 (Figure 3). The partial correclation is significant (p-value = 0.0118). The variable is selected.
  - At the thirs step, we observe that PRIX is the most partially correlated variable (0.1507). It is significant (p-value = 0.4721). The selection process is stoped.

There are other search strategies (backward, stepwise). They give very similar results in the most of situations. Actually, there is not really an optimal subset of explanatory variables. It is more informative to examine their influence on the regression.

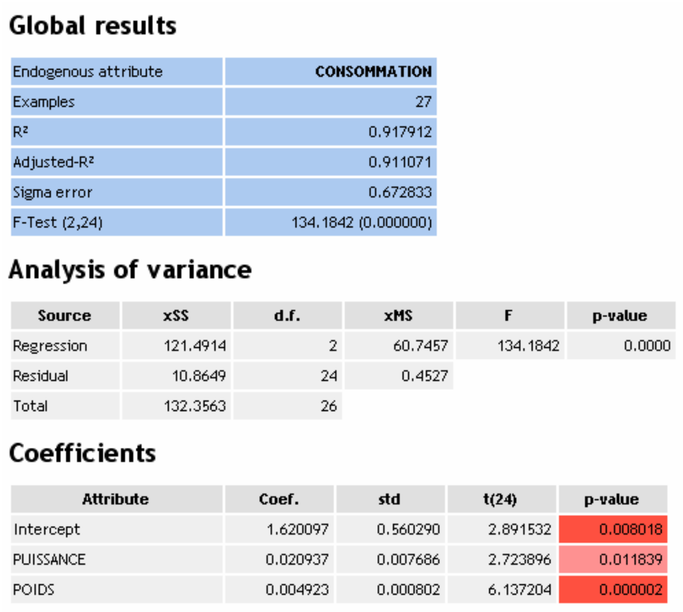
Let us consider the second step of the selection process. CYLINDREE (partial-r = 0.5719) is in competition with PUISSANCE (horsepower, partial-r = 0.4859). It seems better; it is added in the selected subset. In the next steps, PUISSANCE is forevermore excluded. This does not mean that PUISSANCE has no influence on the consumption. If we perform a regression with POIDS and PUISSANCE, we observe that the model (Figure 6 – R2 = 0.9179) is almost as good as the selected model above (Figure 5 – R2= 0.9277), and PUISSANCE is significant (p-value = 0.0118).

[2] http://en.wikipedia.org/wiki/Statistical_significance

[3] http://faculty.chass.ncsu.edu/garson/PA765/partialr.htm

[4] http://en.wikipedia.org/wiki/Partial_correlation

Thus, we should be used carefully the variable selection process. Of course, it removes the irrelevant explanatory variables. But it removes (masks) also variables which are strongly associated with the dependent variable, but redundant with some already selected variables. It is really important to analyze attentively the results provided by the selection process in order to differentiate these two kinds of explanatory variables (irrelevant or redundant).



Global results

Endogenous attribute	CONSOMMATION
Examples	27
R²	0.917912
Adjusted-R²	0.911071
Sigma error	0.672833
F-Test (2,24)	134.1842 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	121.4914	2	60.7457	134.1842	0.0000
Residual	10.8649	24	0.4527		
Total	132.3563	26			

Coefficients

Attribute	Coef.	std	t(24)	p-value
Intercept	1.620097	0.560290	2.891532	0.008018
PUISSANCE	0.020937	0.007686	2.723896	0.011839
POIDS	0.004923	0.000802	6.137204	0.000002

Figure 6 – Regression with POIDS and PUISSANCE

5. Regression from the factors of PCA

Principal component analysis is a variable reduction procedure. From the original variables, it computes a small number of articial variables called "principal components" or "factors" or "latent variables". These new variables are uncorrelated. They can be used as predictof in subsequent analysis[5,6]. In our context, we use them as explanatory variables in the regression analysis.

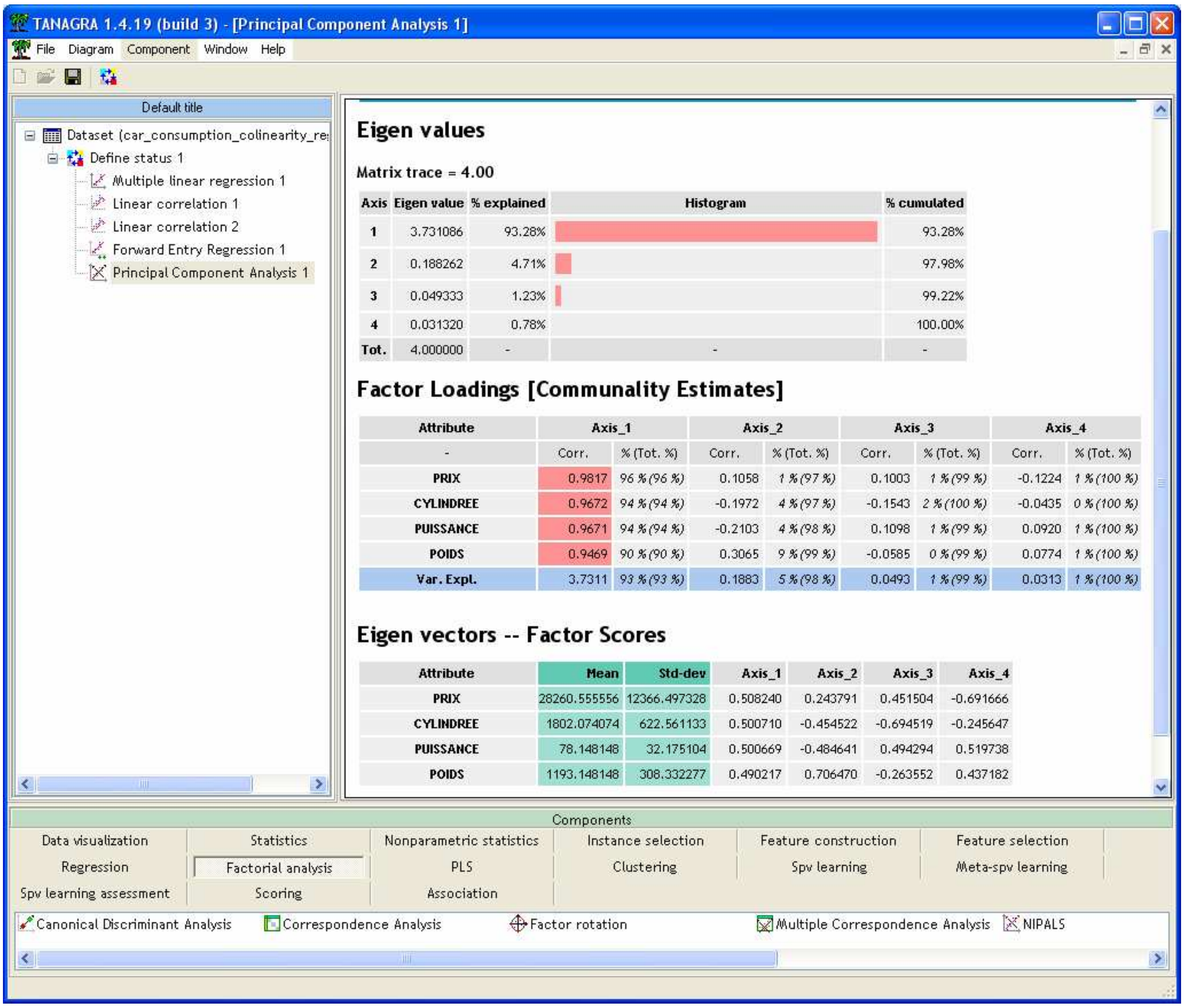
Thus, the regression process is organized as follows: (1) we compute the factors from the explanatory variables; (2) we use some of them as new explanatory variable in the regression analysis; (3) we obtain the coefficients of the linear combination on the original variable from the results of the regression and the PCA.

5.1. The principal component analysis (PCA)

To perform a PCA with Tanagra, we add the PRINCIPAL COMPONENT ANALYSIS (FACTORIAL ANALYSIS tab). We click on the VIEW menu, we obtain the following results.

[5] http://support.sas.com/publishing/pubcat/chaps/55129.pdf

[6] About its implementation with Tanagra, see for instance http://data-mining-tutorials.blogspot.com/2009/04/principal-component-analysis-pca.html



Eigen values

Matrix trace = 4.00

Axis	Eigen value	% explained	Histogram	% cumulated
1	3.731086	93.28%		93.28%
2	0.188262	4.71%		97.98%
3	0.049333	1.23%		99.22%
4	0.031320	0.78%		100.00%
Tot.	4.000000	-	-	-

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2		Axis_3		Axis_4	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
PRIX	0.9817	96 % (96 %)	0.1058	1 % (97 %)	0.1003	1 % (99 %)	-0.1224	1 % (100 %)
CYLINDREE	0.9672	94 % (94 %)	-0.1972	4 % (97 %)	-0.1543	2 % (100 %)	-0.0435	0 % (100 %)
PUISSANCE	0.9671	94 % (94 %)	-0.2103	4 % (98 %)	0.1098	1 % (99 %)	0.0920	1 % (100 %)
POIDS	0.9469	90 % (90 %)	0.3065	9 % (99 %)	-0.0585	0 % (99 %)	0.0774	1 % (100 %)
Var. Expl.	3.7311	93 % (93 %)	0.1883	5 % (98 %)	0.0493	1 % (99 %)	0.0313	1 % (100 %)

Eigen vectors -- Factor Scores

Attribute	Mean	Std-dev	Axis_1	Axis_2	Axis_3	Axis_4
PRIX	28260.555556	12366.497328	0.508240	0.243791	0.451504	-0.691666
CYLINDREE	1802.074074	622.561133	0.500710	-0.454522	-0.694519	-0.245647
PUISSANCE	78.148148	32.175104	0.500669	-0.484641	0.494294	0.519738
POIDS	1193.148148	308.332277	0.490217	0.706470	-0.263552	0.437182

Figure 7 – Results of the principal component analysis (PCA)

The visualization window is divided in three parts.

- EIGEN VALUES[7]. The eigenvalue for a given factor measures the variance in all the variables which is accounted for by that factor. The ratio of eigenvalues is the ratio of explanatory importance of the factors with respect to the variables. We observe that 93.28% of the variance is explained by the first factor on our dataset. It means, among others, that the explanatory variables are highly redundant. The second factor explains only 4.71% of the global variance, the others are negligible.
- FACTOR LOADINGS. The factor loadings, also called component loadings in PCA, are the correlation coefficients between the variables (rows) and factors (columns). Analogous to Pearson's r, the squared factor loading is the percent of variance in that indicator variable explained by the factor. We observe here that all the variables are correlated with the first factor. It is rather natural: cars with a large engine are also heavy and powerful; last, they are also costly (price).

[7] See http://faculty.chass.ncsu.edu/garson/PA765/factor.htm for detailed description.

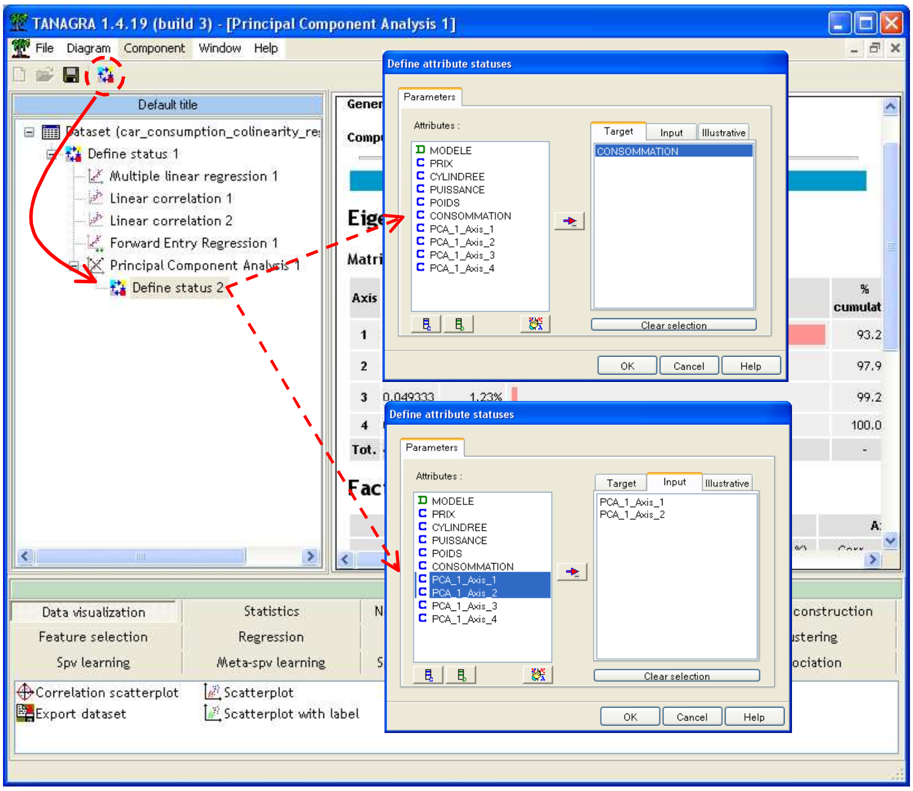
- EIGEN VECTORS - FACTOR SCORES. This last table gives the coefficients which are useful to compute the factor scores of an instance. The factor score is the coordinate on the representation space defined by the factors. About the first car (Daihatsu Core -), its coordinate on the first factor is computed as follows

$$
\begin{aligned}
v_1 &= 0.51 \times \left(\frac{11600-28260.6}{12366.5}\right) + 0.50 \times \left(\frac{846-1802.1}{622.6}\right) + 0.50 \times \left(\frac{32-78.1}{32.2}\right) + 0.49 \times \left(\frac{650-1193.1}{308.3}\right) \\
&= -3.035
\end{aligned}
$$

5.2. Regression from the factors of the PCA

A crucial problem is the determination of the number of factors used for the regression. The easiest way is to use all the factors. But this solution has a drawback. We know that some factors have a very weak power, their proportion of variance explained (Eigen value) is low. They are very unstable; indeed, they correspond to residual information of the explanatory variable. By using only the relevant factors, we perform a kind of regularization by smoothing the information provided by the learning instances.

For our dataset, taking into account the PCA results, we retain only the first two factors. We insert the DEFINE STATUS component into the diagram. We set as TARGET the dependent variable for the regression, as INPUT the two first factors PCA_1_AXIS_1 and PCA_1_AXIS_2.



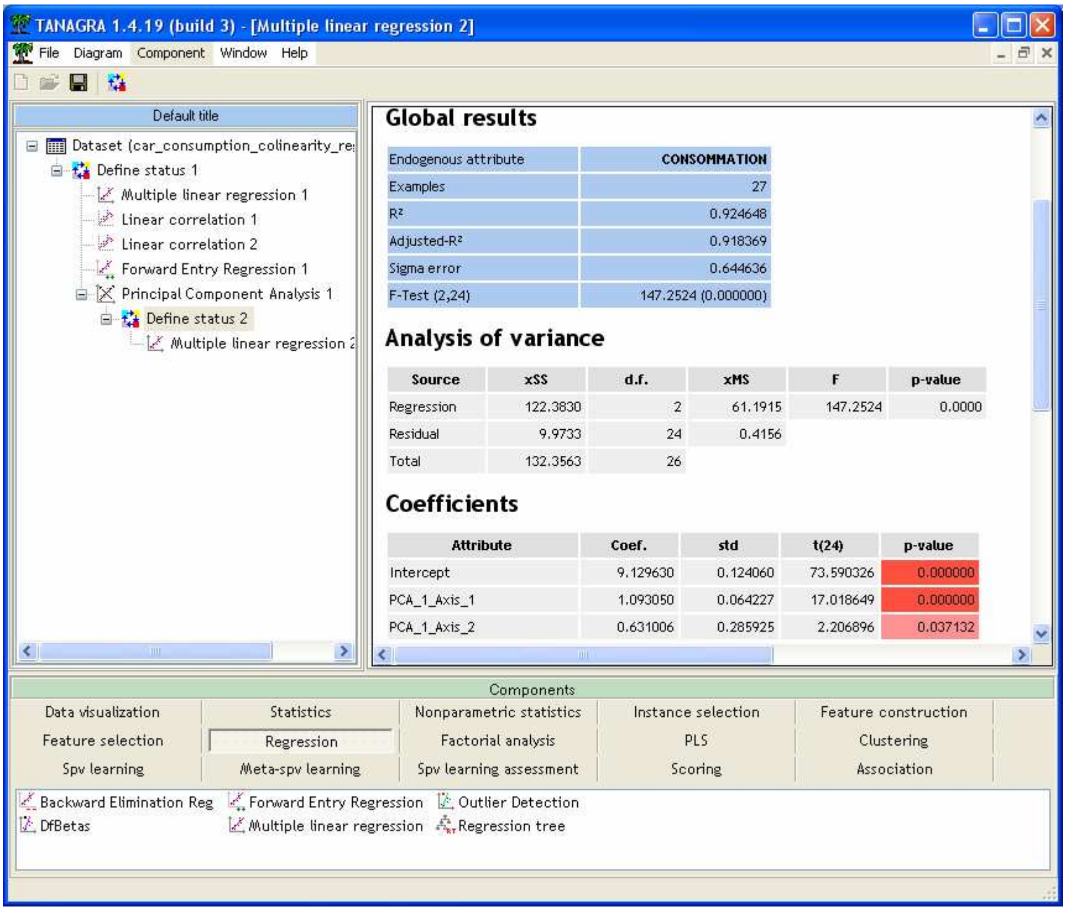
Define attribute statuses — Attributes: MODELE, PRIX, CYLINDREE, PUISSANCE, POIDS, CONSOMMATION, PCA_1_Axis_1, PCA_1_Axis_2, PCA_1_Axis_3, PCA_1_Axis_4

Target: CONSOMMATION

Define attribute statuses — Attributes: MODELE, PRIX, CYLINDREE, PUISSANCE, POIDS, CONSOMMATION, PCA_1_Axis_1, PCA_1_Axis_2, PCA_1_Axis_3, PCA_1_Axis_4

Input: PCA_1_Axis_1, PCA_1_Axis_2

Then we insert again the MULTIPLE LINEAR REGRESSION component. We click on the VIEW menu.



Global results

Endogenous attribute	CONSOMMATION
Examples	27
R²	0.924648
Adjusted-R²	0.918369
Sigma error	0.644636
F-Test (2,24)	147.2524 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	122.3830	2	61.1915	147.2524	0.0000
Residual	9.9733	24	0.4156		
Total	132.3563	26			

Coefficients

Attribute	Coef.	std	t(24)	p-value
Intercept	9.129630	0.124060	73.590326	0.000000
PCA_1_Axis_1	1.093050	0.064227	17.018649	0.000000
PCA_1_Axis_2	0.631006	0.285925	2.206896	0.037132

Figure 8 - Regression on the two first factors

The model seems good since the coefficient of determination is R2 = 0.9246, similar to this obtained from the regression on POIDS ans CYLINDREE (Figure 5, R² = 0.9276). The two factors are both significant.

Note: *We have attempted to perform the regression on all the factors. It appears that the two last ones (third and fourth) are not significant in the regression. This corroborates the choice to use only the two first factors in the regression phase*.

5.3. Obtaining the coefficients of the linear combination from the initial variables

At this step, our model is not easy to deploy on unseen cases because we must handle many linear combinations: those which define the factors, the coefficients of the regression on the selected factors. Obviously, it is more convenient to have a single linear combination defined on the original explanatory variables. The interpretation of the results, understanding the associations between each explanatory variable and the dependent variable, is easier.

For this, we use the results of the regression (Figure 8) in conjunction with those of the PCA (Figure 7). Let V1 and V2 the factors of the PCA, m_x and σ_x are the mean and the standard deviation of the explanatory variable X. The equation of the linear combination from the original explanatory variables is computed as follows.

$$
\begin{aligned}
y &= 9.13 + 1.09 \times V_1 + 0.63 \times V_2 \\
&= 9.13 + 1.09 \times \left[0.51 \times \left(\frac{prix - m_{prix}}{\sigma_{prix}} \right) + 0.50 \times \left(\frac{cylindree - m_{cylindree}}{\sigma_{cylindree}} \right) + \cdots \right] \\
&\quad + 0.63 \left[0.24 \times \left(\frac{prix - m_{prix}}{\sigma_{prix}} \right) - 0.45 \times \left(\frac{cylindree - m_{cylindree}}{\sigma_{cylindree}} \right) + \cdots \right]
\end{aligned}
$$

Then we obtain the equation defined on the standardized variables.

$$
\begin{aligned}
y &= 9.13 \\
&\quad + 0.7094 \times \left(\frac{prix - m_{prix}}{\sigma_{prix}} \right) \\
&\quad + 0.2605 \times \left(\frac{cylindree - m_{cylindree}}{\sigma_{cylindree}} \right) \\
&\quad + 0.2414 \times \left(\frac{puissance - m_{puissance}}{\sigma_{puissance}} \right) \\
&\quad + 0.9816 \times \left(\frac{poids - m_{poids}}{\sigma_{poids}} \right)
\end{aligned}
$$

Let us consider the coefficients of this linear combination. Because they are defined on the standardized variables, the influences of each explanatory variable on the dependent variable are directly comparable. We observe that all the variables have a positive influence on the consumption. We note also that the weight (poids) has the strongest influence on the consumption. When the weight increases of one standard deviation, the consumption increases of 0.9816 times of its standard deviation.

By inserting the estimated values of the mean and the standard deviation, we obtain the unstandardized coefficients of the model defined from the original explanatory variables.

$$y = 2.36954 + 0.00006 \times prix + 0.00042 \times cylindree + 0.00750 \times puissance + 0.00318 \times poids$$

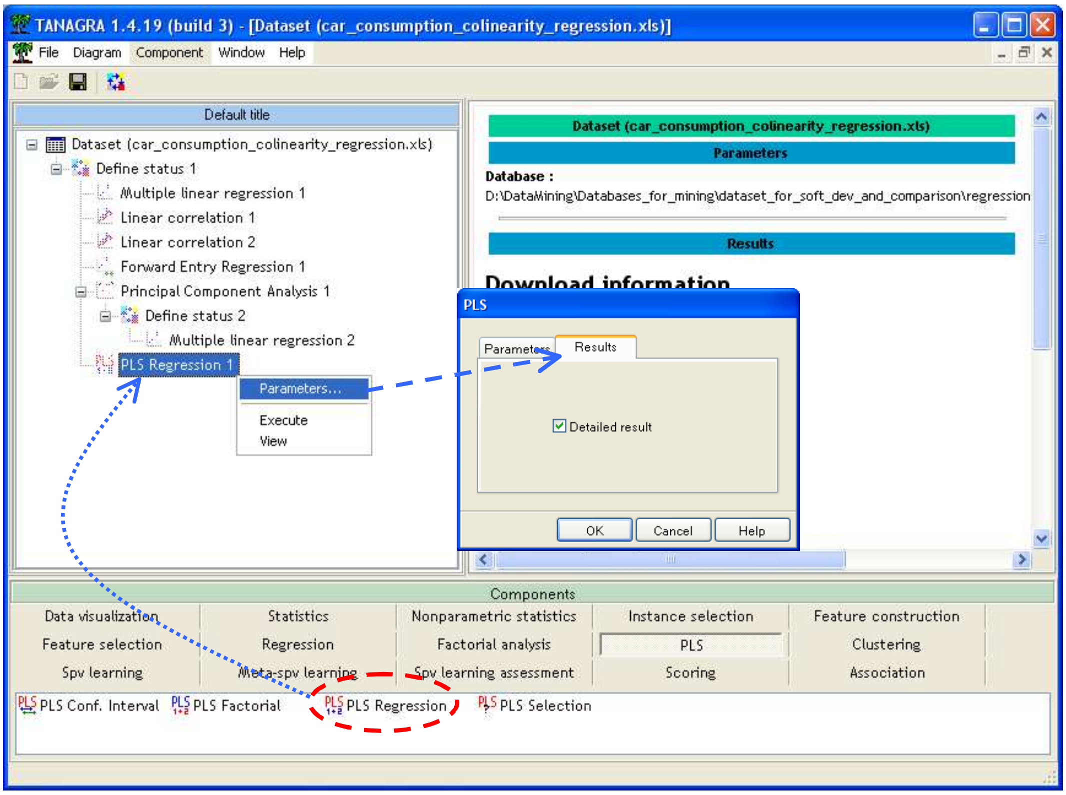
6. PLS Regression

“Partial least squares” (PLS) is sometimes called "Projection to Latent Structures" because of its general strategy. The explanatory variables are reduced to principal components, as are the Y variables (the approach can handle one or more dependent variables). The components of X are used to predict the scores on the Y components, and the predicted Y component scores are used to predict the actual values of the Y variables. While the original X variables may be multicollinear, the X components used to predict Y will be orthogonal (http://faculty.chass.ncsu.edu/garson/PA765/pls.htm). Thus, this approach is particularly convenient for treating the multicollinearity problem.

Like for the PCA + Regression approach, we can smooth the used information by selecting only the relevant factors for the construction of the final model.

6.1. Reading the output of the PLS regression

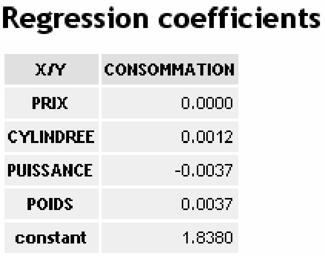
We add the PLS REGRESSION component (PLS tab) into the diagram, behind DEFINE STATUS 1. We click on the PARAMETERS contextual menu. With the default settings, the algorithm computes automatically L = MIN (5, number of input variables) factors. So, we obtain 4 factors on our dataset. We activate the RESULTS tab. We select the DETAILED RESULTS option.



☑ Detailed result

We validate our choice and we click on the VIEW menu.

Regression coefficients. We obtain the regression parameter estimates (Figure 9). We can use them to predict the value of the dependent variable of an unlabeled case. **Because we use all the factors (corresponding to the number of explanatory variables), we note that we have the same coefficients of the standard multiple linear regression** (Figure 2).



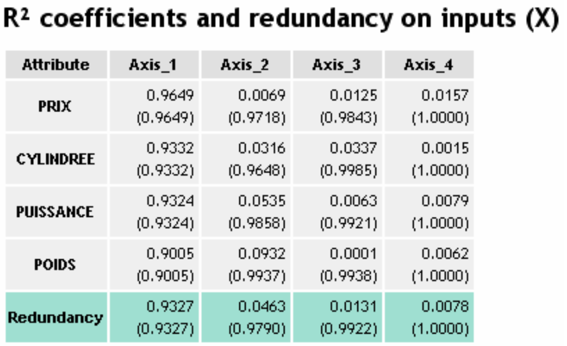
Regression coefficients

X/Y	CONSOMMATION
PRIX	0.0000
CYLINDREE	0.0012
PUISSANCE	-0.0037
POIDS	0.0037
constant	1.8380

Figure 9 - Coefficients de la régression PLS - 4 axes

Proportion of variance explained by latent factors for X (Redundancy). It described the part of the variance of the explanatory variables explained by the latent factors (Figure 10). When we use all the factors (the last column of the table), we explain all the variance of each explanatory variable (in brackets, the cumulative part of variance explained). We note here that from the third factor, the part of variance explained is negligible whatever the independent variable. It suggests that only the two first factors are enough to build an efficient model.

The last row (redundancy) describes the part of variance of all the explanatory variables explained by the latent factors. Like for the PCA, we observe that the two first factors explains about 98% of the global variance. But, unlike the PCA, the factors are computed by taking into account the values of the dependent variable in the context of the PLS regression.



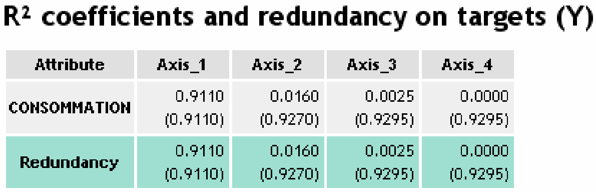
R² coefficients and redundancy on inputs (X)

Attribute	Axis_1	Axis_2	Axis_3	Axis_4
PRIX	0.9649 (0.9649)	0.0069 (0.9718)	0.0125 (0.9843)	0.0157 (1.0000)
CYLINDREE	0.9332 (0.9332)	0.0316 (0.9648)	0.0337 (0.9985)	0.0015 (1.0000)
PUISSANCE	0.9324 (0.9324)	0.0535 (0.9858)	0.0063 (0.9921)	0.0079 (1.0000)
POIDS	0.9005 (0.9005)	0.0932 (0.9937)	0.0001 (0.9938)	0.0062 (1.0000)
Redundancy	0.9327 (0.9327)	0.0463 (0.9790)	0.0131 (0.9922)	0.0078 (1.0000)

Figure 10 - Redundancy for the latent factors – PLS regression

Proportion of variance explained by latent factors for Y (Redundancy). The aim of the PLS Regression is to explain (predict) the values of the dependent variable(s) (Y). This table (Figure 11) describes the proportion of the variance of the dependent variable explained by the latent factors. In a prediction purpose, this table is maybe the most important of the results.

Once again, if we use all the latent factors here, the coefficient of determination (redundancy - Figure 11) is the same as this obtained with the standard linear regression i.e. R2 = 92.95% (Figure 2). But we observe into the same table the quality of the model when we use only some of the latent factors. For instance, if we use only the two first factors, we have already explained R2 = 92.70% of the variance of Y. It suggests that only the two first factors are enough to obtain a good model.



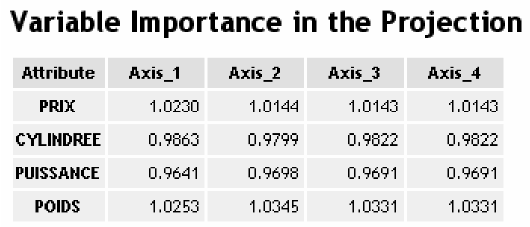
R² coefficients and redundancy on targets (Y)

Attribute	Axis_1	Axis_2	Axis_3	Axis_4
CONSOMMATION	0.9110 (0.9110)	0.0160 (0.9270)	0.0025 (0.9295)	0.0000 (0.9295)
Redundancy	0.9110 (0.9110)	0.0160 (0.9270)	0.0025 (0.9295)	0.0000 (0.9295)

Figure 11 – Redundancy on Y – PLS Regression

Variable Importance in Projection. The last table gives an indication about the contribution of each explanatory variable on the dependent variable, through the latent factors (Figure 12). Like for the regression on the PCA factors, we note that all the explanatory variables have a positive influence on the consumption. POIDS and PRIX seem the most important ones (VIP > 1).

Note: The influence of prices on consumption is not very obvious. If we understand that the price is certainly correlated with the consumption, it seems strange that it has an influence. This would mean that if we reduce artificially the price of cars, their consumption decreased in the same time. A numerical technique can not demonstrate that it is nonsense. This is for this reason that we will always need the domain knowledge to validate the results.



Variable Importance in the Projection

Attribute	Axis_1	Axis_2	Axis_3	Axis_4
PRIX	1.0230	1.0144	1.0143	1.0143
CYLINDREE	0.9863	0.9799	0.9822	0.9822
PUISSANCE	0.9641	0.9698	0.9691	0.9691
POIDS	1.0253	1.0345	1.0331	1.0331

Figure 12 – VIP table – PLS Regression

6.2. Choosing the appropriate number of factors

The determination of the appropriate number of factors is a crucial problem in the PLS regression. Some considerations expressed above, mostly about the redundancies (Figure 10 and Figure 11), seem suggest that the "optimal" number factors could be 2. But this approach is mainly a rule of thumb. From the reading these tables, two statisticians might come to different conclusions.

There is a more stringent approach to determine the right number of latent factors. It is based on the accuracy of the model in a prediction perspective. The PLS SELECTION component is based on the PRESS criterion (Predicted Residual Sum of Squares).

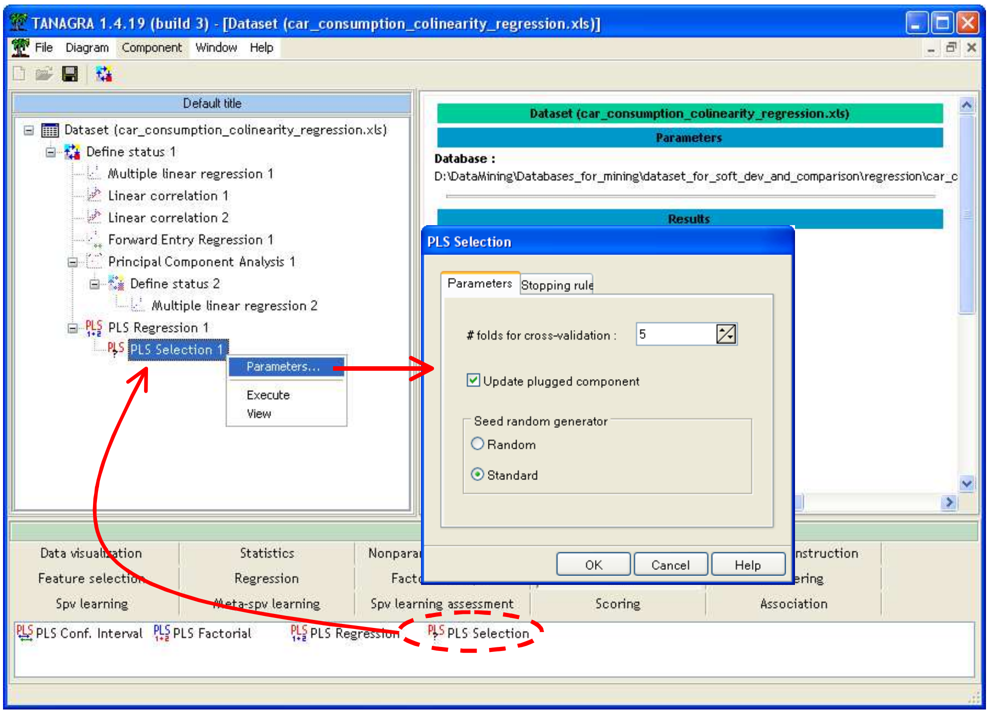
The **PRESS** is computed in a leave-one-out way[8]: (a) we remove the i-th instance from the dataset; (b) we learn the model form the remaining instances; (c) we predict the output of the model for this instance; (d) we compare the predicted and the observed values, (e) we store the square of the error; (f) we perform this procedure for all the instances, and we obtain the sum of the square of the error.

PRESS si more reliable than the **RSS** criterion (Residual Sum of Squares) where the i-th instance is used for the construction of the predictive model.

Another criterion based on the PRESS and RSS can be used in order to determine the right number of factors. **Q2** compares the RSS of the model computed on (h-1) factors to the PRESS of the model based on h factors. The aim is to assess the real contribution of the h-th factor on the quality of the prediction.

We insert the PLS SELECTION (PLS tab) into the diagram. We click on the PARAMETRES contextual menu. We set the following settings.

[8] We describe the process based on leave-one-out here, but we can generalize the approach on a cross-validation framework. The computation is faster while maintaining accuracy. This approach is favored under Tanagra.



# folds for cross-validation : 5

☑ Update plugged component

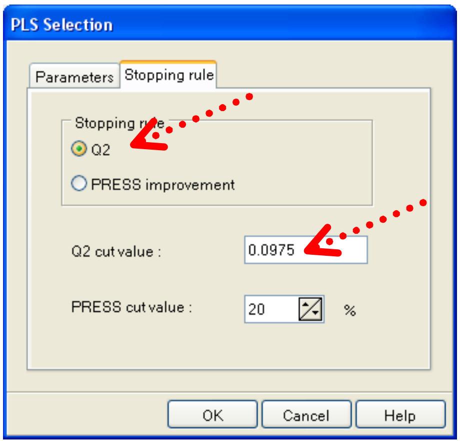
Seed random generator

○ Random

● Standard

Into the first tab “PARAMETERS”:

- We can specify the number of folds for the cross-validation process. The default value is 5 i.e. the dataset is subdivided in 5 folds: 4/5 of the dataset is used during the construction of the model, 1/5 of them for the computation of the error[9].
- « Update plugged component ». If this option is selected, the preceding component is automatically updated with the optimal number of factors after the search process.
- Last, « Seed Random Generator » allows to specify the behavior of the random number generator.



Stopping rule

● Q2

○ PRESS improvement

Q2 cut value : 0.0975

PRESS cut value : 20 %

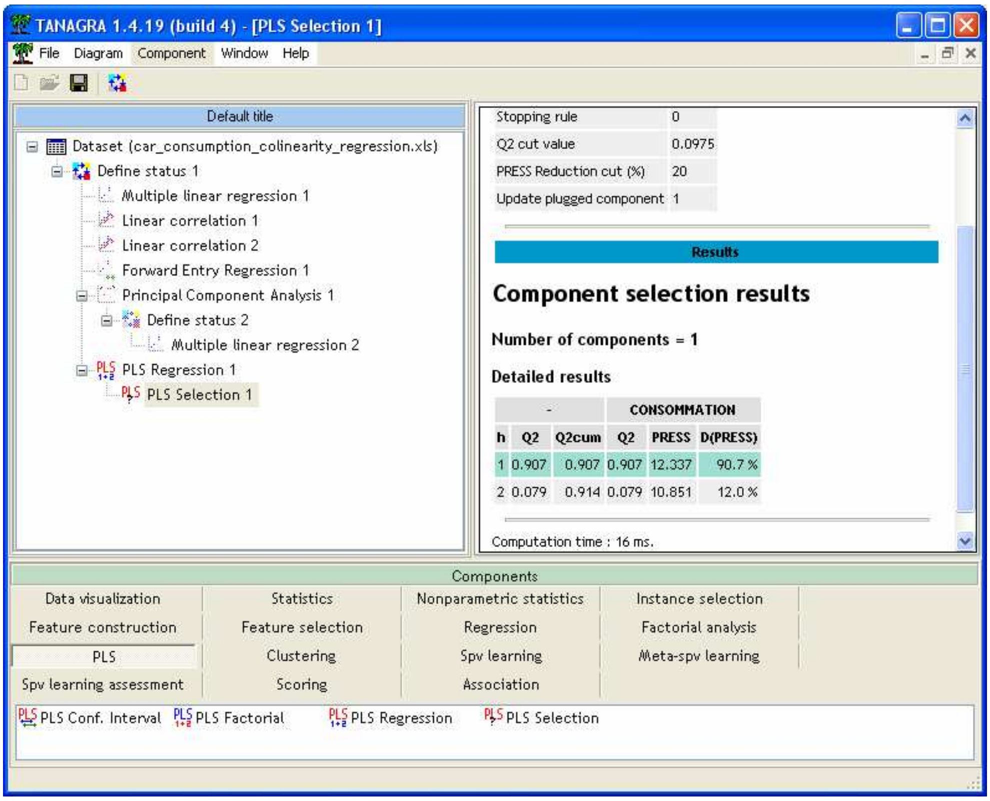
[9] http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29 – See K-fold cross validation.

Into the second tab “STOPPING RULE”, we specify the way to stop the search process i.e. the method of detection of the right number of factors:

- Q2 criterion, the seach is stopped if Q2 > Q2 cut value (default value = 0.05).
- We can use also the PRESS criterion. In this case, we stop the process if the relative decreasing is lower a threshold (default value = 20%) when we add a factor.

On our dataset, we select the Q2 criterion, the used cut value is 0.0975.

We validate these settings and we click on the VIEW menu. We obtain the following results.



Stopping rule	0
Q2 cut value	0.0975
PRESS Reduction cut (%)	20
Update plugged component	1

Results

Component selection results

Number of components = 1

Detailed results

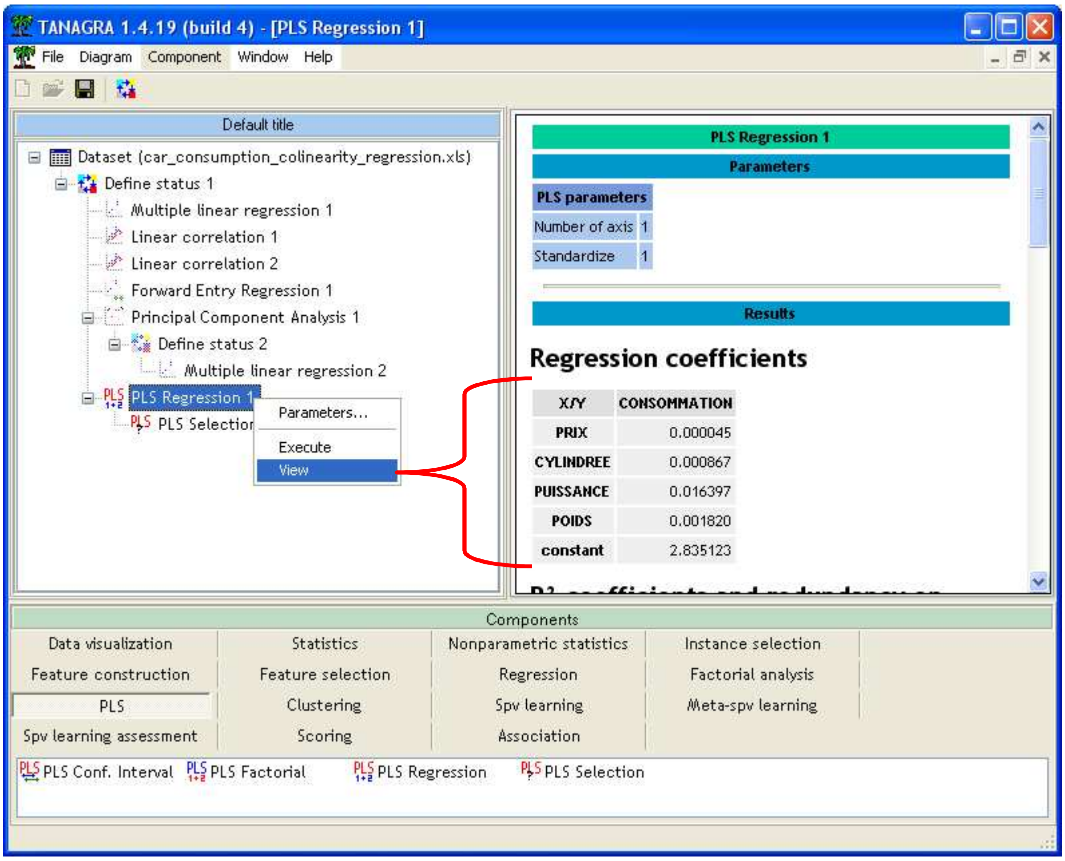
-			CONSOMMATION		
h	**Q2**	**Q2cum**	**Q2**	**PRESS**	**D(PRESS)**
1	0.907	0.907	0.907	12.337	90.7 %
2	0.079	0.914	0.079	10.851	12.0 %

Computation time : 16 ms.

When we select only one factor, the Q2 is 0.907 and the PRESS is 12.337. When we add the 2nd factor, the PRESS decreases to 10.851, but it seems that this diminution is not really significant because the Q2 is only 0.079, lower than our cut value (0.0975).

Thus, we select the model with only one latent factor. We note that the PLS REGRESSION component is automatically updated. We can visualize the new coefficients by clicking on the VIEW menu of PLS REGRESSION 1 into the diagram. The new coefficients are (Figure 13)

$$y = 2.835123 + 0.000045 \times prix + 0.000867 \times cylindree + 0.016397 \times puissance + 0.001820 \times poids$$



PLS Regression 1

Parameters

PLS parameters	
Number of axis	1
Standardize	1

Results

Regression coefficients

X/Y	CONSOMMATION
PRIX	0.000045
CYLINDREE	0.000867
PUISSANCE	0.016397
POIDS	0.001820
constant	2.835123

Figure 13 - Coefficients of PLS Regression based on 1 latent factor

We summarize the unstandardized estimated parameters of the model according the method:

Variable	Standard multiple regression	Regression from factors of PCA	PLS Regression
Constante	1.8380	2.36954	2.835123
Prix	0.0000	0.00006	0.000045
Cylindrée	0.0012	0.00042	0.000867
Puissance	-0.0037	0.00750	0.0016397
Poids	0.0037	0.00318	0.001820

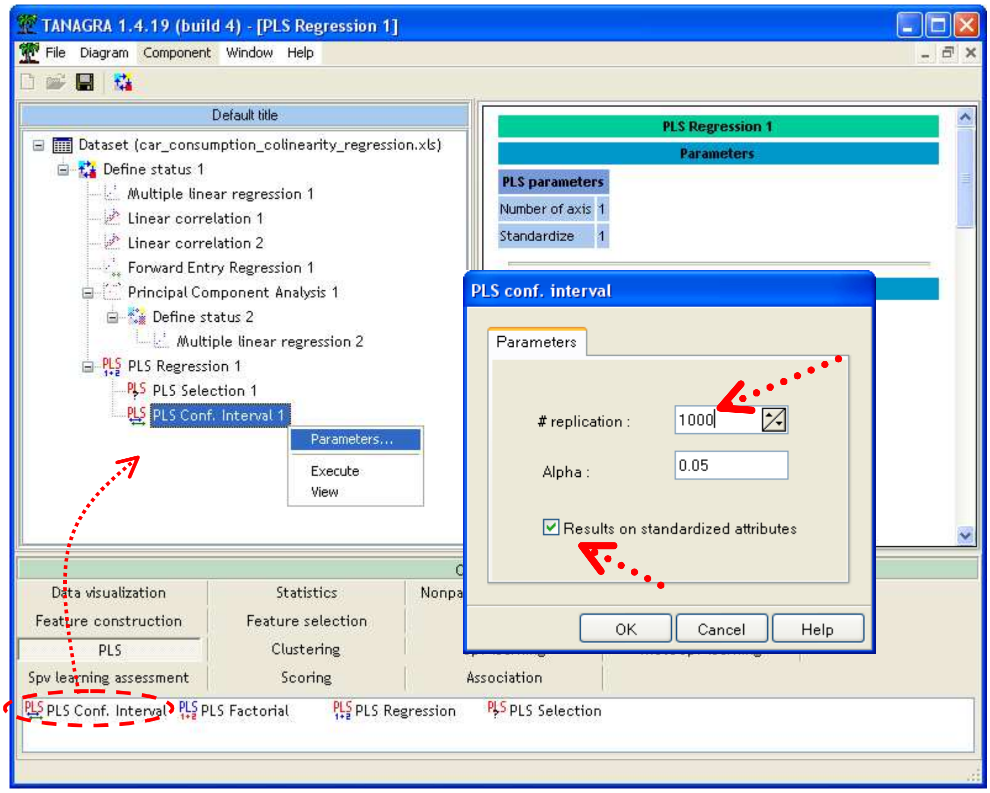
On the one hand, the parameters estimated from the regression of factors of the PCA and the PLS regression are consistent. On the other hand, the coefficients estimated from the standard linear regression are very different for some explanatory variables (e.g. PUISSANCE). We know at this time that this is a consequence of the multicollinearity problem.

6.3. Assessing the estimated parameters from the PLS Regression

One of the main advantages of the standard regression is that we can easily assess the contribution of each explanatory variable by computing the confidence interval of the estimated parameters or by implementing the test of significance. About the first approach, we consider that the contribution of an explanatory variable is significant if the confidence interval of its associated coefficient does not cover the zero value.

Tanagra uses a resampling approach to implement the same assessment for the estimated parameters of the PLS Regression. Roughly speaking, the process can be described as follows: we draw a sample of n instances with replacement (n is the size of the available dataset); we compute the parameters; we repeat this process K times (K is a parameter of the algorithm); we obtain empirically the confidence interval by computing the quantiles. For $1-\alpha$ confidence level, the lower bound corresponds to the $\alpha/2$ -quantile; and the upper bound is the ($1-\alpha/2$)-quantile.

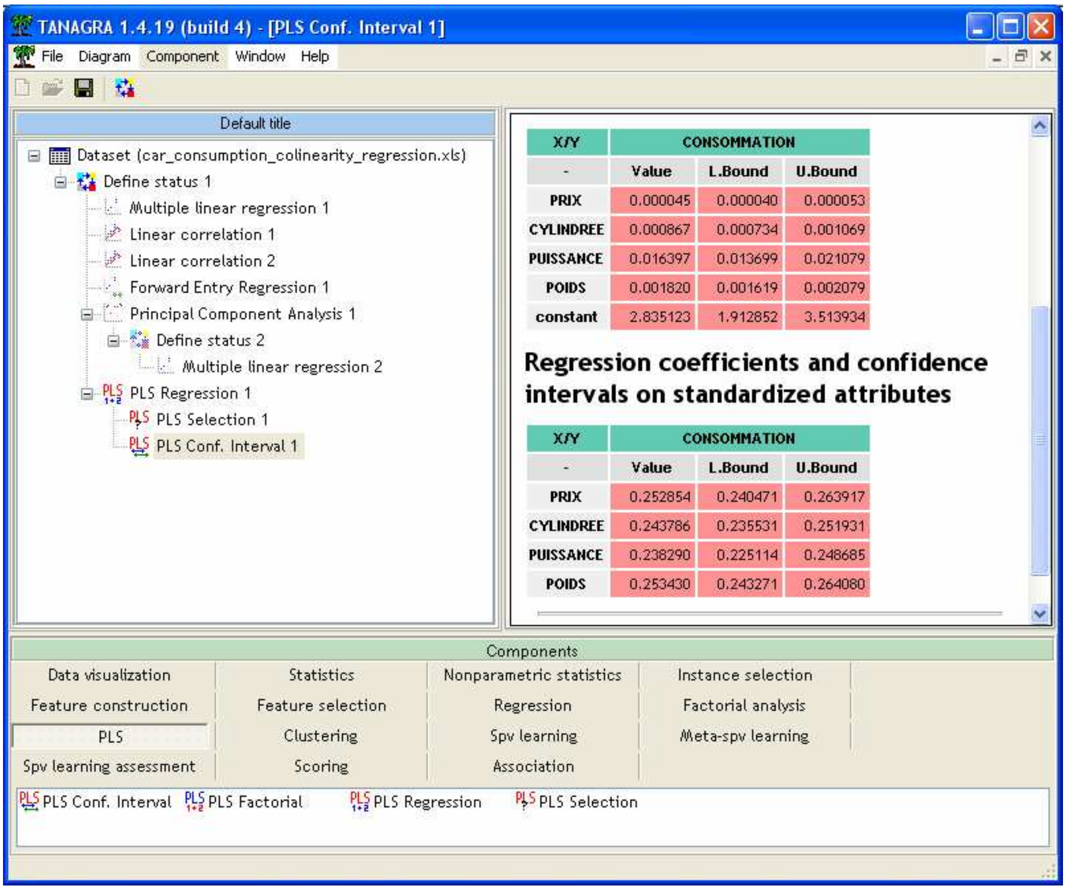
We add PLS CONF INTERVAL (PLS tab) behind PLS REGRESSION 1. That means that it uses the same settings during the resampling process, especially the same number of latent factors. We click on the PARAMETERS menu. We ask K = 1000 replications. We ask also the standardized parameters in order to compare the influence of the explanatory variables.



Number of axis 1
Standardize 1

replication : 1000
Alpha : 0.05
☑ Results on standardized attributes

We click on the VIEW menu to launch the calculations. We obtain both the unstandardized and the standardized estimated parameters. According these last ones, we observe that all the variables are significant; they have a positive influence on the consumption.



X/Y	CONSOMMATION		
-	Value	L.Bound	U.Bound
PRIX	0.000045	0.000040	0.000053
CYLINDREE	0.000867	0.000734	0.001069
PUISSANCE	0.016397	0.013699	0.021079
POIDS	0.001820	0.001619	0.002079
constant	2.835123	1.912852	3.513934

Regression coefficients and confidence intervals on standardized attributes

X/Y	CONSOMMATION		
-	Value	L.Bound	U.Bound
PRIX	0.252854	0.240471	0.263917
CYLINDREE	0.243786	0.235531	0.251931
PUISSANCE	0.238290	0.225114	0.248685
POIDS	0.253430	0.243271	0.264080

Moreover, we note that the coefficients are very stable. This draws our attention if we consider the small size of the learning sample. It is certainly the consequence of the utilization of a few numbers of latent factors.

7. Conclusion

In this tutorial, we show how to deal with the multicollinearity problem in a regression framework.

Of course, the results can be slightly different according the approaches. The most important for us is to understand the outputs of the software, and deduce the correct interpretation of the estimated parameters. The case of PUISSANCE (horsepower) is very interesting. Because it is correlated with CYLINDREE (engine size), we could conclude that it has no influence on the consumption. We observe afterwards, when we use the appropriate approaches, that it is not true.

Nevertheless, whatever the quality of a statistical technique, nothing can replace human expertise. In our problem, it is clear that the price can not be a predictor of consumption, although we can easily understand that they are somehow linked. Neither approach has been able to highlight this nonsense.