1 Topic

The SAS add-in 4.3 for Excel. Comparing the results of some calculations with those of Tanagra.

The connection between a data mining tool and a spreadsheet application such as Excel is a really valuable feature. We benefit from the powerful of the first one, and the popularity and the easy to use of the second one. Many people use a spreadsheet in their data preparation phase. Recently, I have presented an add-in for the connection between R and Excel. In this document, I describe a similar tool for the SAS software.

SAS (http://www.sas.com/) is a popular tool, well-known of the statisticians. But the use of SAS is not really simple for the non-specialist people. We must know the syntax of the commands before to perform a statistical analysis. With the SAS add-in for Excel, some of the SAS drawbacks are alleviated: we do not need to load and organize the dataset into a bank; we do not need to know the command syntax to perform an analysis and set the associated parameters (we use a menu and dialog boxes instead); the results are automatically incorporated in a new sheet of an Excel workbook (the post processing of the results becomes easy).

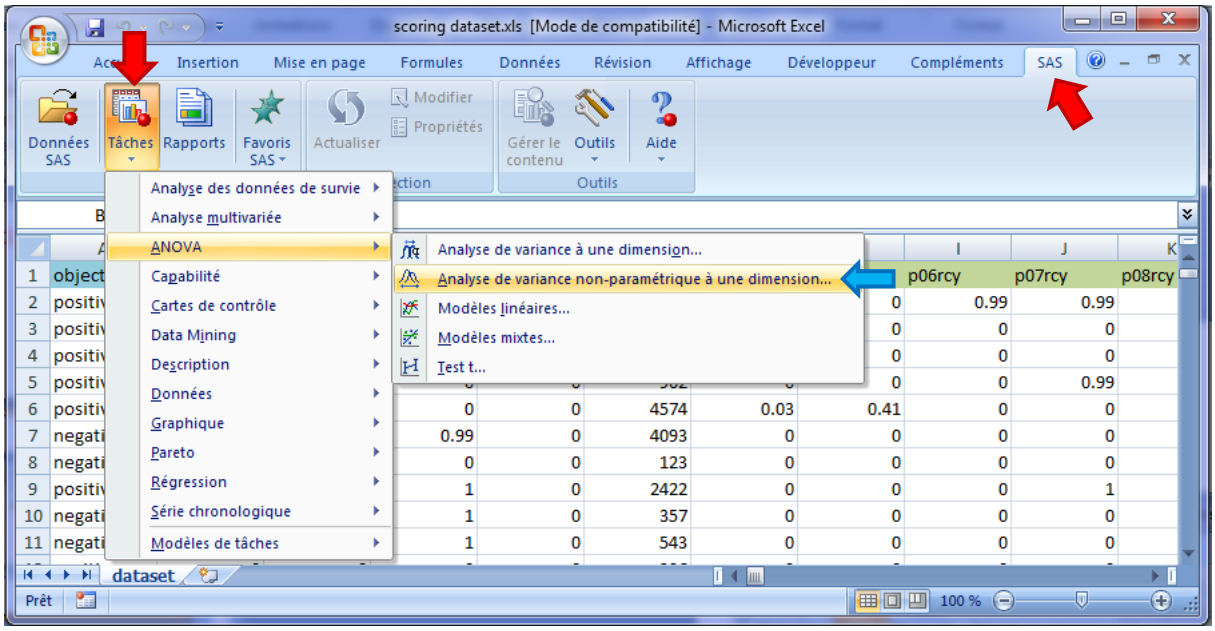
In this tutorial, I describe the behavior of the add-in for various kinds of analyses (nonparametric statistic, logistic regression). We compare the results with those of Tanagra.

2 Dataset

We use the « scoring_dataset.xls » data file. It contains 2158 instances and 201 variables. The "objective" variable describes the individuals which respond positively or not to a marketing campaign. We load the data file into Excel 2007.

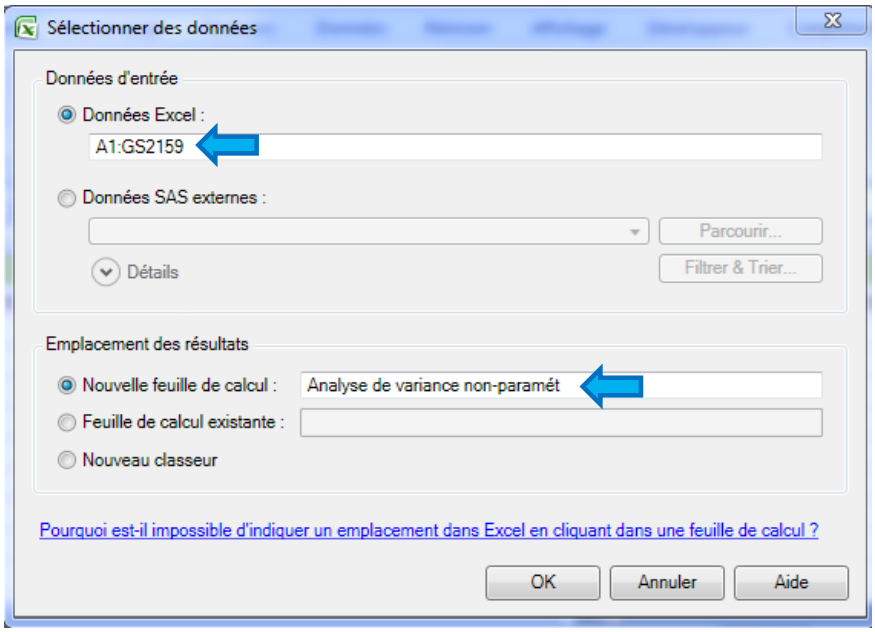
3 The SAS 4.3 add-in

When we launch Excel, a new tab "SAS" appears into the Excel ribbon. The statistical methods are available when we click on the TACHES button (probably TASK in the English version).

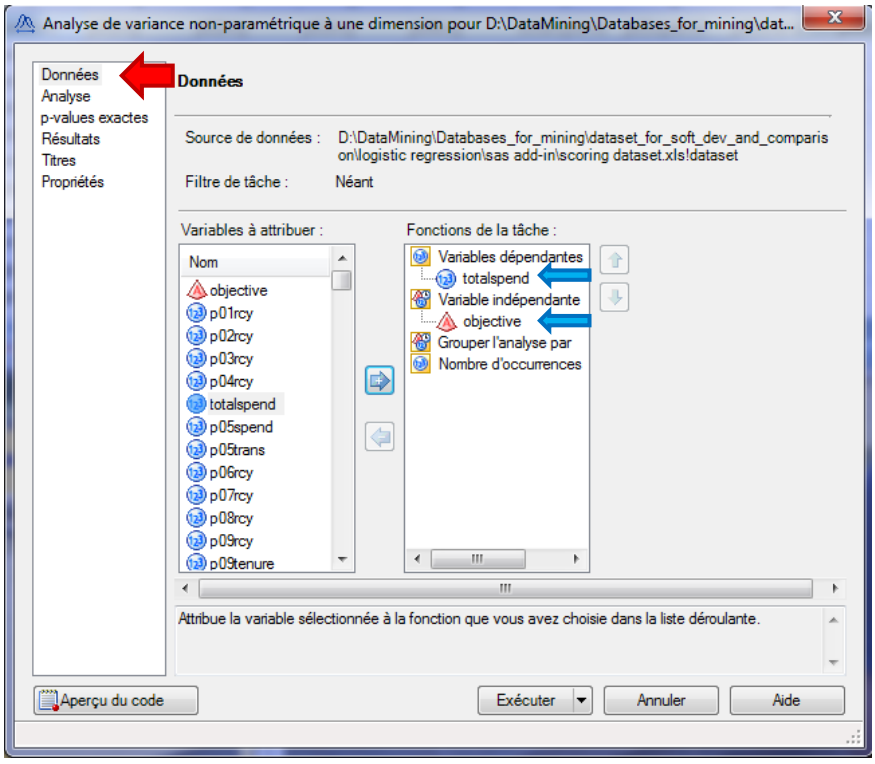


3.1 Non-parametric statistics

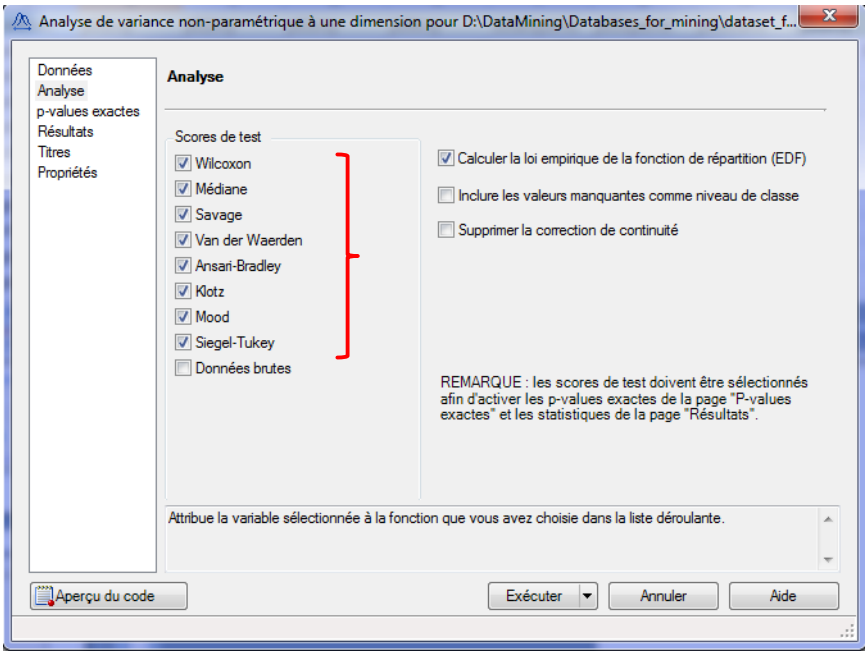
We want to compare the spending of the customers (TOTAL SPEND) according to their response to the marketing campaign. We select first the cells containing the dataset, including the first row which corresponds to the names of the variables. Then, we click on the TACHES / ANOVA / ANALYSE DE VARIANCE NON PARAMETRIQUE A UNE DIMENSION menu. A dialog box appears. We check the coordinates of the selected cells, and we can set the name of the sheet in which the results are incorporated. We validate our settings by clicking on the OK button.



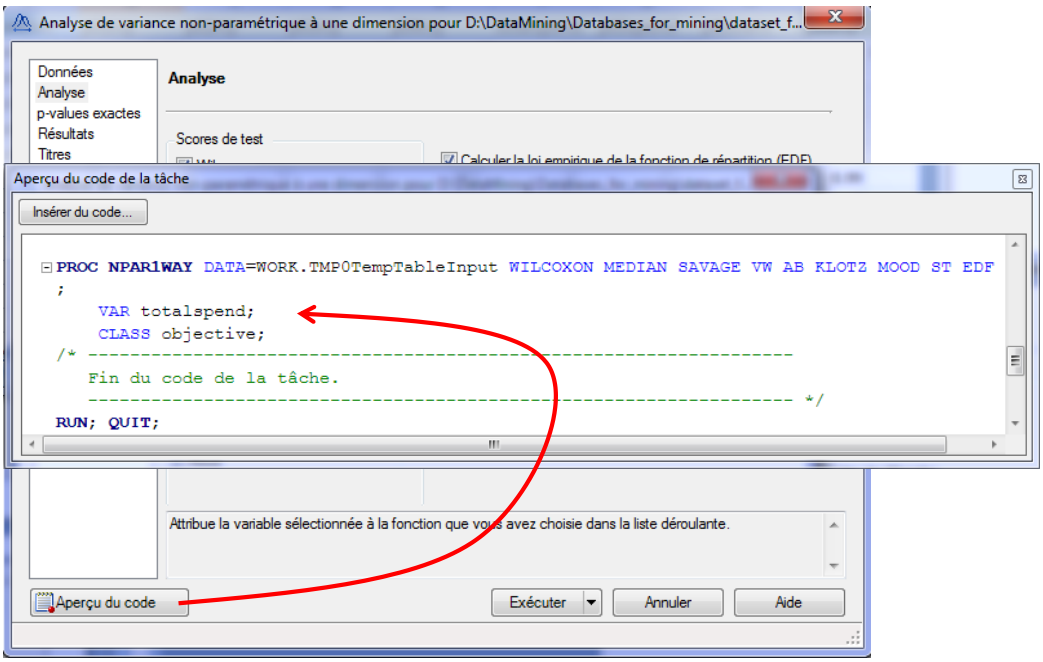
Another dialog box enables to set the parameters of the statistical method. We set the variables used in the analysis: OBJECTIVE is the independent variable, TOTALSPEND is the dependent variable.



In the same dialog box, for ANALYSE we select the tests that we want to perform.



Other options are available. We observe that we can inspect the SAS command by clicking on the "Aperçu du code" button. This is really interesting if we want to learn the SAS command language. Here for instance, we note that SAS uses the NPAR1WAY procedure.



This feature is very similar to the one of the RATTLE package for R[1]. We can moreover refine the analysis by modifying manually the commands. So, we click on the EXECUTER button.

For **Tanagra**, we import the dataset using the tanagra.xla add-in[2]. We set a nonparametric statistic analysis (e.g. http://data-mining-tutorials.blogspot.fr/2008/11/nonparametric-statistics.html). Into the DEFINE STATUS component, we set TOTALSPEND as TARGET and OBJECTIVE as INPUT).

[1] http://data-mining-tutorials.blogspot.fr/2011/08/data-mining-with-r-rattle-package.html

3.1.1 Wilcoxon-Mann-Whitney test

SAS computes the Wilcoxon statistic, Tanagra the Mann-Whitney one[3]. But we have the same standardized Z statistic, |Z| = 9.91233. Because we have a large sample, the continuity correction used by SAS is not perceptible.

Analyse de variance non-paramétrique à une dimension

Scores de Wilcoxon (Sommes du rang) pour la variable totalspend Classés par variable objective					
objective	N	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
positive	1079	1308241	1164780.5	14472.9379	1212.4569
negative	1079	1021320	1164780.5	14472.9379	946.5431
Les scores moyens ont été utilisés pour les liens.					

Test à deux échantillons de Wilcoxon	
Statistique	1308241
Approximation normale	
Z	9.9123
Unilatéral Pr > Z	<.0001
Bilatéral Pr > |Z|	<.0001
Approximation t	
Unilatéral Pr > Z	<.0001
Bilatéral Pr > |Z|	<.0001
Z inclut une correction de continuité de 0.5.	

SAS

Results

		Value	Examples	Average	Rank sum	Rank mean
totalspend	objective	positive	1079	1763.1909	1308241.0	1212.4569
		negative	1079	992.8267	1021320.0	946.5431
		All	2158	1378.0088	2329561.0	1079.5000

TANAGRA

Mann-Whitney U	438660.00000
E(U)	582120.50000
V(U)	209465931.84330
|Z|	9.91233
P(>|Z|)	0.00000

3.1.2 Kruskal-Wallis test

Test de Kruskal-Wallis	
Khi-2	98.2542
DLL	1
Pr > Khi-2	<.0001

SAS

Results

Attribute_Y	Attribute_X	Description					Statistical test		
		Value	Examples	Average	Rank sum	Rank mean	Statistics	Value	Proba
totalspend	objective	positive	1079	1763.1909	1308241.0	1212.4569	Kruskal-Wallis	98.254035	0.000000
		negative	1079	992.8267	1021320.0	946.5431	KW (corr.ties)	98.254236	0.000000
		All	2158	1378.0088	2329561.0	1079.5000			

TANAGRA

[2] http://data-mining-tutorials.blogspot.fr/2010/08/tanagra-add-in-for-office-2007-and.html

[3] http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U

SAS provides the results for the Kruskal-Wallis test[4] with the previous analysis. Tanagra uses a dedicated component. We obtain the same results.

3.1.3 Median test

Two approaches can be used for the median test: the first is based on the ranking, the second on the contingency table (http://en.wikipedia.org/wiki/Median_test). Both SAS and Tanagra provide the results for the two approaches. SAS…

Analyse de variance non-paramétrique à une dimension

Scores médians (Nbre de points au-dessus de la médiane) pour la variable totalspend Classés par variable objective					
objective	N	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
positive	1079	638.333333	539.5	11.609082	0.591597
negative	1079	440.666667	539.5	11.609082	0.408403
Les scores moyens ont été utilisés pour les liens.					

Test à deux échantillons de la médiane	
Statistique	638.3333
Z	**8.5134**
Unilatéral Pr > Z	<.0001
Bilatéral Pr > |Z|	<.0001

Analyse à une dimension de la médiane	
Khi-2	**72.4788**
DLL	1
Pr > Khi-2	<.0001

SAS

…TANAGRA.

Results

Attribute_Y	Attribute_X	Description					Statistical test	
		Value	Examples	Average	Scores sum	Scores mean	Two-Sample Test	
totalspend	objective	positive	1079	1763.1909	638.3333	0.5916	S	440.66667
		negative	1079	992.8267	440.6667	0.4084	E(S)	539.50000
		All	2158	1378.0088	1079.0	0.5000	V(S)	134.77079
							|Z|	8.51345
							p-value	0.00000
							One-way Analysis	
							Chi-Square	72.47882
							d.f.	1
							p-value	0.00000

TANAGRA
« Median test »

[4] http://en.wikipedia.org/wiki/Kruskal%E2%80%93Wallis_one-way_analysis_of_variance

3.1.4 Van der Waerden test

The Van der Waerden test[5] provides also the two kinds of results (based on the Z-statistic and on the chi-squared statistic).

Analyse de variance non-paramétrique à une dimension

SAS

Scores de Van der Waerden (Normal) pour la variable totalspend
Classés par variable objective

objective	N	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
positive	1079	223.95545	0	23.158361	0.207558
negative	1079	-223.95545	0	23.158361	-0.207558

Les scores moyens ont été utilisés pour les liens.

Test à deux échantillons de Van der Waerden	
Statistique	223.9554
Z	9.6706
Unilatéral Pr > Z	<.0001
Bilatéral Pr > |Z|	<.0001

Analyse à une dimension de Van der Waerden	
Khi-2	93.5207
DLL	1
Pr > Khi-2	<.0001

Results

Attribute_Y	Attribute_X
totalspend	objective

Description

Value	Examples	Average	Scores sum	Scores mean
positive	1079	1763.1909	223.9555	0.2076
negative	1079	992.8267	-223.9554	-0.2076
All	2158	1378.0088	0.0	0.0000

TANAGRA

Statistical test

Two-Sample Test	
S	-223.95545
E(S)	0.00000
V(S)	536.30966
|Z|	9.67061
p-value	0.00000
One-way Analysis	
Chi-Square	93.52068
d.f.	1
p-value	0.00000

3.1.5 Savage test

The Savage test is available into SAS only.

Analyse de variance non-paramétrique à une dimension

Scores selon la formule de Savage (Exponentiel) pour la variable totalspend
Classés par variable objective

objective	N	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
positive	1079	216.94123	0	23.18801	0.201058
negative	1079	-216.94123	0	23.18801	-0.201058

Les scores moyens ont été utilisés pour les liens.

Test à deux échantillons de Savage	
Statistique	216.9412
Z	9.3558
Unilatéral Pr > Z	<.0001
Bilatéral Pr > |Z|	<.0001

Analyse à une dimension de Savage	
Khi-2	87.5301
DLL	1
Pr > Khi-2	<.0001

[5] http://en.wikipedia.org/wiki/Van_der_Waerden_test

3.1.6 Siegel and Tukey test

The Siegel and Tukey test[6] is available into SAS only also. From here, we compare the differences in scale of the distributions (differences in location previously).

Analyse de variance non-paramétrique à une dimension

Scores Siegel-Tukey pour la variable totalspend Classés par variable objective					
objective	N	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
positive	1079	1140335.6	1164780.5	14472.8912	1056.84486
negative	1079	1189225.4	1164780.5	14472.8912	1102.15514
Les scores moyens ont été utilisés pour les liens.					

Test à deux échantillons de Siegel-Tukey	
Statistique	1140335.601
Z	-1.689
Unilatéral Pr < Z	0.0456
Bilatéral Pr > |Z|	0.0912
Z inclut une correction de continuité de 0.5.	

Analyse à une dimension de Siegel-Tukey	
Khi-2	2.8528
DLL	1
Pr > Khi-2	0.0912

3.1.7 Ansari-Bradley test

The Ansari-Bradley test is present both in SAS and TANAGRA.

Scores Ansari-Bradley pour la variable totalspend Classés par variable objective					
objective	N	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
positive	1079	570436.667	582660	7236.44447	528.67161
negative	1079	594883.333	582660	7236.44447	551.32839
Les scores moyens ont été utilisés pour les liens.					

SAS

Test à deux échantillons de Ansari-Bradley	
Statistique	570436.6667
Z	-1.6891
Unilatéral Pr < Z	0.0456
Bilatéral Pr > |Z|	0.0912

Analyse à une dimension de Ansari-Bradley	
Khi-2	2.8532
DLL	1
Pr > Khi-2	0.0912

Results

Attribute_Y	Attribute_X	Description					Statistical test	
		Value	Examples	Average	Scores sum	Scores mean	Two-Sample Test	
		positive	1079	1763.1909	570436.6666	528.6716	S	570436.66663
		negative	1079	992.8267	594883.3333	551.3284	E(S)	582659.99994
		All	2158	1378.0088	1165320.0	540.0000	V(S)	52366128.47201
totalspend	objective						|Z|	1.68914
							p-value	0.09119
							One-way Analysis	
							Chi-Square	2.85318
							d.f.	1
							p-value	0.09119

TANAGRA

[6] http://en.wikipedia.org/wiki/Siegel%E2%80%93Tukey_test

3.1.8 Klotz test

The Klotz test is a nonparametric test for scale differences.

SAS

Scores Klotz Scores pour la variable totalspend Classés par variable objective					
objective	N	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
positive	1079	1131.69048	1072.29604	32.118911	1.048833
negative	1079	1012.9016	1072.29604	32.118911	0.938741
Les scores moyens ont été utilisés pour les liens.					

Test à deux échantillons de Klotz	
Statistique	1131.6905
Z	1.8492
Unilatéral Pr > Z	0.0322
Bilatéral Pr > |Z|	0.0644

Analyse à une dimension de Klotz	
Khi-2	3.4196
DLL	1
Pr > Khi-2	0.0644

TANAGRA

Results

Attribute_Y	Attribute_X	Description					Statistical test	
		Value	Examples	Average	Scores sum	Scores mean	Two-Sample Test	
		positive	1079	1763.1909	1131.6905	1.0488	S	1012.90160
		negative	1079	992.8267	1012.9016	0.9387	E(S)	1072.29604
		All	2158	1378.0088	2144.6	0.9938	V(S)	1031.62442
totalspend	objective						|Z|	1.84920
							p-value	0.06443
							One-way Analysis	
							Chi-Square	3.41956
							d.f.	1
							p-value	0.06443

3.1.9 Mood test

The Mood test described here is intended for the comparison of the scales (MOOD SCALE TEST). The Mood's runs test, present also in Tanagra, has another goal (MOOD RUNS TEST).

SAS

Scores Mood pour la variable totalspend Classés par variable objective					
objective	N	Somme des scores	Attendue sous H0	Ecart-type sous H0	Score moyen
positive	1079	431767608	418738590	8064157.45	400155.337
negative	1079	405709571	418738590	8064157.45	376005.163
Les scores moyens ont été utilisés pour les liens.					

Test à deux échantillons de Mood	
Statistique	431767608.4
Z	1.6157
Unilatéral Pr > Z	0.0531
Bilatéral Pr > |Z|	0.1062

Analyse à une dimension de Mood	
Khi-2	2.6104
DLL	1
Pr > Khi-2	0.1062

TANAGRA

Results

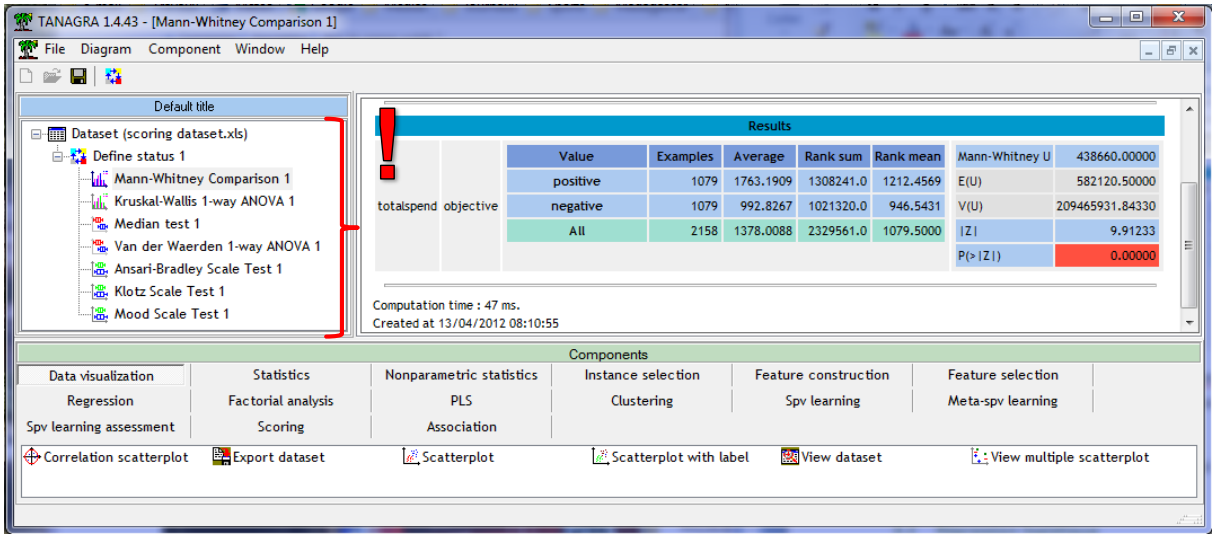
Attribute_Y	Attribute_X	Description					Statistical test	
		Value	Examples	Average	Scores sum	Scores mean	Two-Sample Test	
		positive	1079	1763.1909	431767608.4863	400155.3369	S	405709571.57231
		negative	1079	992.8267	405709571.5723	376005.1636	E(S)	418738590.02930
		All	2158	1378.0088	837477180.1	388080.2503	V(S)	65030635439095.49220
totalspend	objective						|Z|	1.61567
							p-value	0.10617
							One-way Analysis	
							Chi-Square	2.61039
							d.f.	1
							p-value	0.10617

SAS add-in 4.3 provides also the Kolmogorov-Smirnov and Cramer-von Mises nonparametric tests.

3.1.10 Tanagra diagram

To perform these analyses, we defined the following diagram with Tanagra[7].

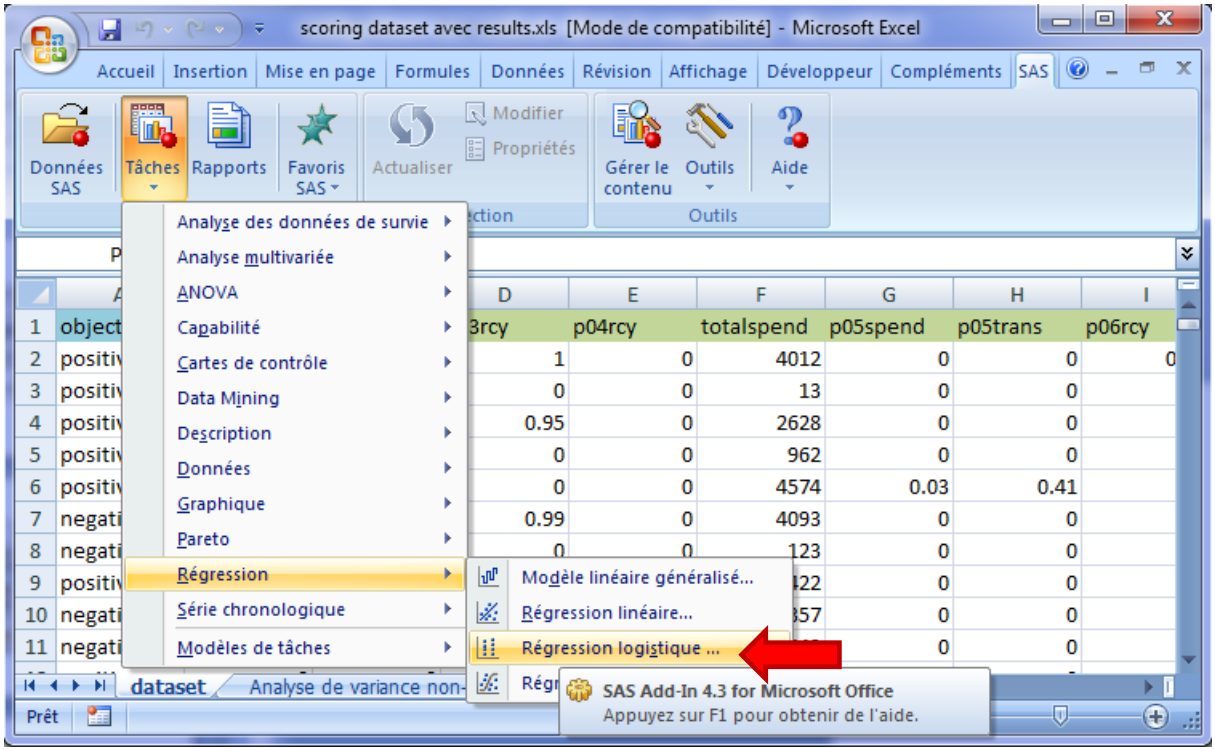
[7] See also « Tests for differences in scale » - http://data-mining-tutorials.blogspot.fr/2009/12/parametric-and-non-parametric-tests-for.html



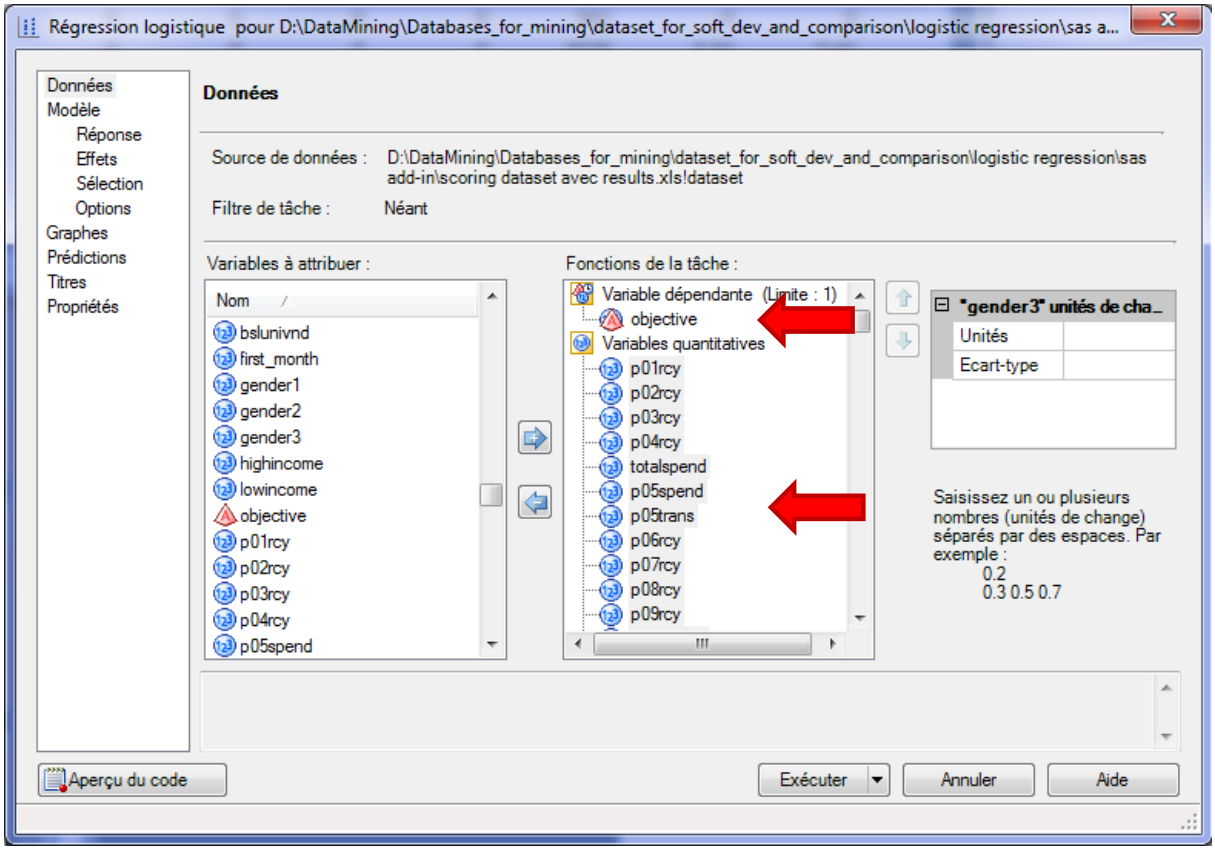
3.2 Logistic regression

In this section, we want to predict the values of OBJECTIVE based on the other available variables using the logistic regression. Because we have a large number of candidate variables (200), we must perform a variable selection in order to obtain the most parsimonious model.

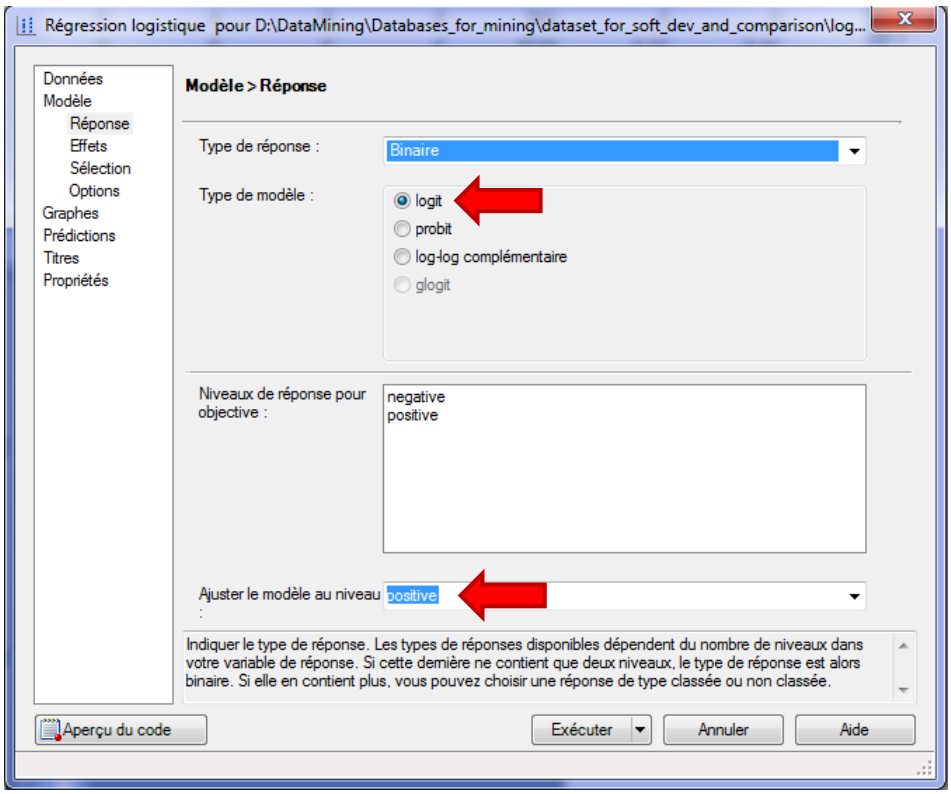
We select the “dataset” sheet into Excel. We click on the SAS / TACHES / REGRESSION / REGRESSION LOGISTIQUE menu.



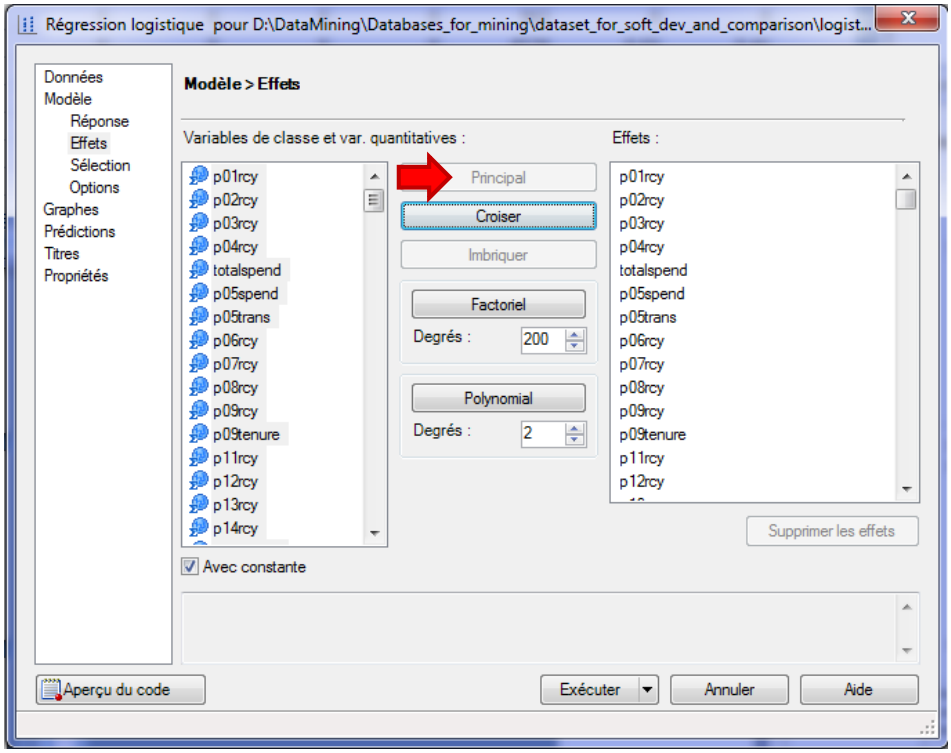
As previously, a dialog box enables to set the dependent variable and the independent variables, we can select also the location of the results.



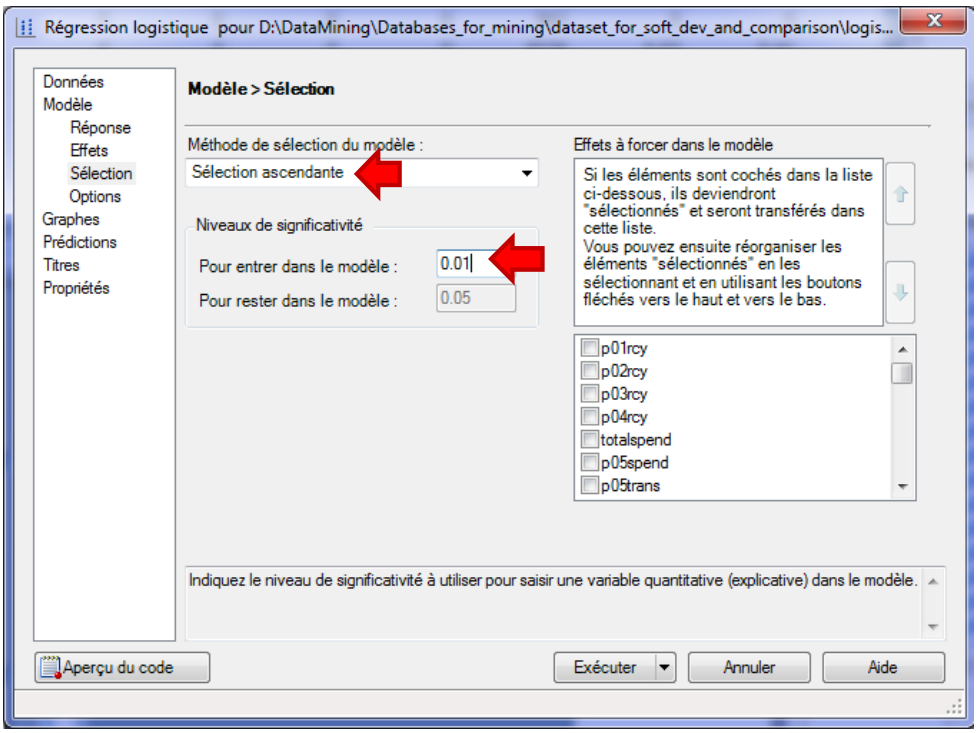
Into the MODELE/REPONSE tab, we set the LOGIT model. We specify the positive value of the target attribute.



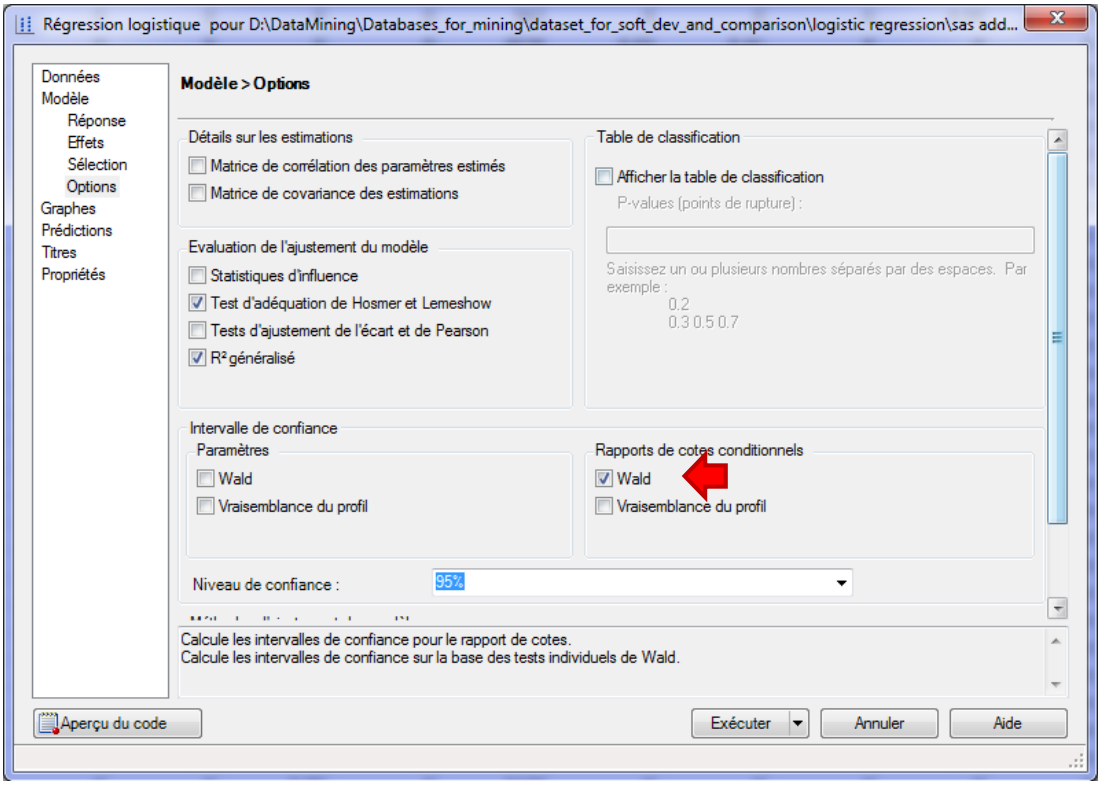
Into the MODELE / EFFETS tab, we set all the independent variables as PRINCIPAL effect. We note that we can set more sophisticated expressions.



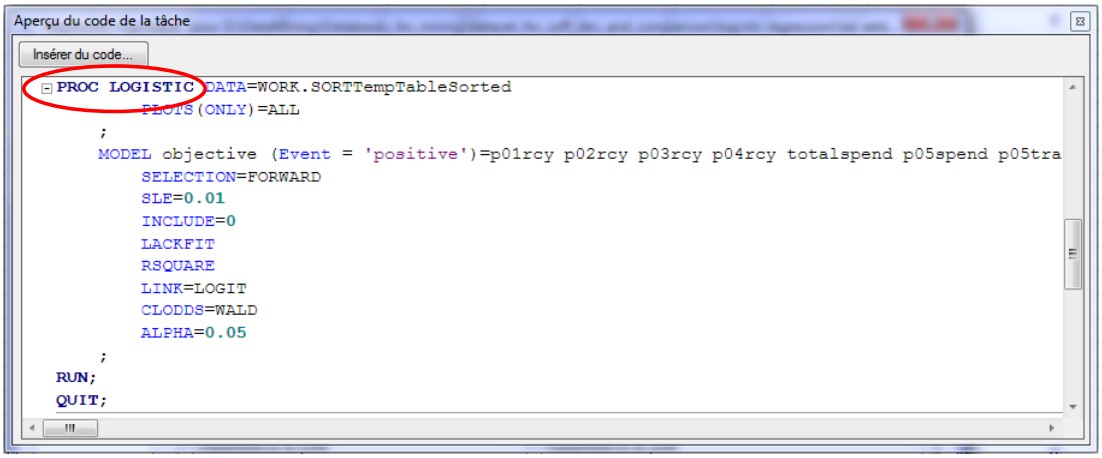
Into the MODEL / SELECTION tab, we set the attribute selection strategy. We select the FORWARD approach which is based on the score test. The significance level is α = 1%.



Last, into the MODELE / OPTIONS, we set the additional options to complete the output of the analysis. We ask, among others, the confidence interval for the odds ratio.



Here also, we can observe the SAS source code for our analysis.



We launch the analysis by clicking on the EXECUTE button. A sheet is added to the Excel workbook.

Informations sur le modèle	
Table	WORK.SORTTEMPTABLESORTED
Variable de réponse	objective
Nombre de niveaux de réponse	2
Modèle	logit binaire
Technique d'optimisation	Score de Fisher

Nombre d'observations lues	2158
Nombre d'observations utili	2158

Profil de réponse		
Valeur ordonnée	**objective**	**Fréquence totale**
1	negative	1079
2	positive	1079

First, a summary details **the characteristics of the analysis**. For instance, we note that we treat a balanced dataset. Then, we have **a detailed description of the variable selection process**.

SAS

Récapitulatif sur la sélection en avant

Etape	Effet saisi	DDL	Nombre dans	Khi-2 du score	Pr > Khi-2
1	gender3	1	1	397.8863	<.0001
2	productcount	1	2	143.2981	<.0001
3	bknfren	1	3	54.5739	<.0001
4	tf37	1	4	48.6375	<.0001
5	p05trans	1	5	18.715	<.0001
6	ahh6ppers	1	6	13.8786	0.0002
7	tf68	1	7	14.3437	0.0002
8	amtfrench	1	8	10.0118	0.0016
9	p09tenure	1	9	9.4223	0.0021
10	tf128	1	10	9.4496	0.0021
11	brlanglic	1	11	8.6923	0.0032
12	p12rcy	1	12	7.4206	0.0064

TANAGRA

N	Current Reg.	Moved	Sol.1
1	AIC : 2993.62	gender3	gender3
	CHI-2 : 0.00	Chi-2 : 397.887	Chi-2 : 397.887
	d.f. : 0	p : 0.0000	p : 0.0000
	p-value : 0.0000		
2	AIC : 2576.00	productcount	productcount
	CHI-2 : 419.63	Chi-2 : 143.299	Chi-2 : 143.299
	d.f. : 1	p : 0.0000	p : 0.0000
	p-value : 0.0000		
3	AIC : 2422.99	bknfren	bknfren
	CHI-2 : 574.63	Chi-2 : 54.575	Chi-2 : 54.575
	d.f. : 2	p : 0.0000	p : 0.0000
	p-value : 0.0000		
4	AIC : 2361.99	tf37	tf37
	CHI-2 : 637.63	Chi-2 : 48.638	Chi-2 : 48.638
	d.f. : 3	p : 0.0000	p : 0.0000
	p-value : 0.0000		
5	AIC : 2313.22	p05trans	p05trans
	CHI-2 : 688.40	Chi-2 : 18.716	Chi-2 : 18.716
	d.f. : 4	p : 0.0000	p : 0.0000
	p-value : 0.0000		
6	AIC : 2293.07	ahh6ppers	ahh6ppers
	CHI-2 : 710.56	Chi-2 : 13.883	Chi-2 : 13.883
	d.f. : 5	p : 0.0002	p : 0.0002
	p-value : 0.0000		
7	AIC : 2280.93	tf68	tf68
	CHI-2 : 724.69	Chi-2 : 14.344	Chi-2 : 14.344
	d.f. : 6	p : 0.0002	p : 0.0002
	p-value : 0.0000		
8	AIC : 2268.53	amtfrench	amtfrench
	CHI-2 : 739.09	Chi-2 : 10.014	Chi-2 : 10.014
	d.f. : 7	p : 0.0016	p : 0.0016
	p-value : 0.0000		
9	AIC : 2260.39	p09tenure	p09tenure
	CHI-2 : 749.24	Chi-2 : 9.440	Chi-2 : 9.440
	d.f. : 8	p : 0.0021	p : 0.0021
	p-value : 0.0000		
10	AIC : 2250.76	tf128	tf128
	CHI-2 : 760.86	Chi-2 : 9.480	Chi-2 : 9.480
	d.f. : 9	p : 0.0021	p : 0.0021
	p-value : 0.0000		
11	AIC : 2243.02	brlanglic	brlanglic
	CHI-2 : 770.60	Chi-2 : 8.693	Chi-2 : 8.693
	d.f. : 10	p : 0.0032	p : 0.0032
	p-value : 0.0000		
12	AIC : 2236.49	p12rcy	p12rcy
	CHI-2 : 779.13	Chi-2 : 7.421	Chi-2 : 7.421
	d.f. : 11	p : 0.0064	p : 0.0064
	p-value : 0.0000		
13	AIC : 2230.92		p02rcy
	CHI-2 : 786.70		Chi-2 : 6.506
	d.f. : 12		p : 0.0108
	p-value : 0.0000	-	"p" higher than 1%, not selected

The chi-squared statistics computed during the process are strictly identical to those of Tanagra. Ultimately, 12 independent variables are selected.

SAS provides many indicators **to evaluate globally the quality of the final model** (AIC, BIC, etc.).

SAS

Statistiques d'ajustement du modèle		
Critère	Constante uniquement	Constante et covariables
AIC	2993.623	2230.92
SC	2999.3	2304.72
-2 Log L	2991.623	2204.92

R carré	0.3055	R carré remis à l'échelle max.	0.4073

Test de l'hypothèse nulle globale : BETA=0			
Test	Khi-2	DDL	Pr > Khi-2
Rapport de vrais	786.703	12	<.0001
Score	659.1976	12	<.0001
Wald	474.7472	12	<.0001

Test du Khi-2 résiduel		
Khi-2	DDL	Pr > Khi-2
227.1726	187	0.0239

TANAGRA

Adjustement quality		
Model Fit Statistics		
Criterion	Intercept	Model
AIC	2993.623	2230.92
SC	2999.3	2304.72
-2LL	2991.623	2204.92
Model Chi test (LR)		
Chi-2		786.703
d.f.		12
P(>Chi-2)		0
R-like		
McFadden's R		0.263
Cox and Snell's R		0.3055
Nagelkerke's R		0.4073

We have the coefficients of the model. SAS enumerates them according to their location into the initial dataset, Tanagra according to their introduction during the variable selection process. But the coefficients, the standard error, the chi-squared Wald statistic and the p-value are the same.

SAS

Estimations par l'analyse du maximum de vraisemblance					
Paramètre	DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept	1	-1.9280	0.2419	63.5181	<.0001
ahh6ppers	1	-5.9698	1.9885	9.0125	0.0027
amtfrench	1	2.7341	0.7459	13.4352	0.0002
bknfren	1	-8.0473	1.4203	32.1021	<.0001
brlanglic	1	2.2944	0.7998	8.2292	0.0041
gender3	1	-1.9310	0.1188	264.3180	<.0001
p05trans	1	-4.5013	1.2440	13.0927	0.0003
p09tenure	1	26.8724	14.3487	3.5074	0.0611
p12rcy	1	0.5115	0.1886	7.3549	0.0067
productcount	1	0.1970	0.0202	95.1812	<.0001
tf128	1	17.6755	5.9650	8.7805	0.003
tf37	1	0.0443	0.0073	36.5450	<.0001
tf68	1	0.0003	0.0001	10.3427	0.0013

Tanagra

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-1.9280	0.2419	63.5182	0.0000
ahh6ppers	-5.9698	1.9885	9.0125	0.0027
amtfrench	2.7341	0.7459	13.4352	0.0002
bknfren	-8.0473	1.4203	32.1021	0.0000
brlanglic	2.2944	0.7998	8.2292	0.0041
gender3	-1.9310	0.1188	264.3180	0.0000
p05trans	-4.5013	1.2440	13.0927	0.0003
p09tenure	26.8725	14.3488	3.5074	0.0611
p12rcy	0.5115	0.1886	7.3549	0.0067
productcount	0.1970	0.0202	95.1812	0.0000
tf128	17.6755	5.9650	8.7805	0.0030
tf37	0.0443	0.0073	36.5450	0.0000
tf68	0.0003	0.0001	10.3427	0.0013

Both SAS and TANAGRA can provide **the estimated odds-ratio and their confidence intervals** (at 95% confidence level).

SAS

confiance de Wald				
Effet	Unité	Valeur estimée	Intervalle de confiance à 95 %	
ahh6ppers	1	0.003	<0.001	0.126
amtfrench	1	15.396	3.568	66.429
bknfren	1	<0.001	<0.001	0.005
brlanglic	1	9.918	2.068	47.56
gender3	1	0.145	0.115	0.183
p05trans	1	0.011	<0.001	0.127
p09tenure	1	>999.999	0.286	>999.999
p12rcy	1	1.668	1.152	2.414
productcoun	1	1.218	1.171	1.267
tf128	1	>999.999	397.123	>999.999
tf37	1	1.045	1.03	1.06
tf68	1	1	1	1

TANAGRA

Odds ratios and 95% confidence intervals			
Attribute	Coef.	Low	High
ahh6ppers	0.003	0.000	0.126
amtfrench	15.396	3.569	66.429
bknfren	0.000	0.000	0.005
brlanglic	9.918	2.068	47.560
gender3	0.145	0.115	0.183
p05trans	0.011	0.001	0.127
p09tenure	4.684E+11	0.286	7.662E+23
p12rcy	1.668	1.152	2.414
productcount	1.218	1.171	1.267
tf128	4.746E+07	397.124	5.673E+12
tf37	1.045	1.030	1.060
tf68	1.000	1.000	1.000

Last, the **Hosmer-Lemeshow test** enables to check the adequacy of the model to the dataset.

SAS

Partition pour les tests de Hosmer et de Lemeshow					
Groupe	Total	objective = positive		objective = negative	
		Observé	Attendu	Observé	Attendu
1	216	11	12.96	205	203.04
2	216	31	29.38	185	186.62
3	216	45	48.37	171	167.63
4	216	78	78.07	138	137.93
5	216	118	107.12	98	108.88
6	216	129	126.66	87	89.34
7	216	143	142.83	73	73.17
8	216	148	159.51	68	56.49
9	216	177	177.23	39	38.77
10	214	199	196.86	15	17.14

Test d'adéquation de Hosmer		
Khi-2	DDL	Pr > Khi-2
6.4875	8	0.5928

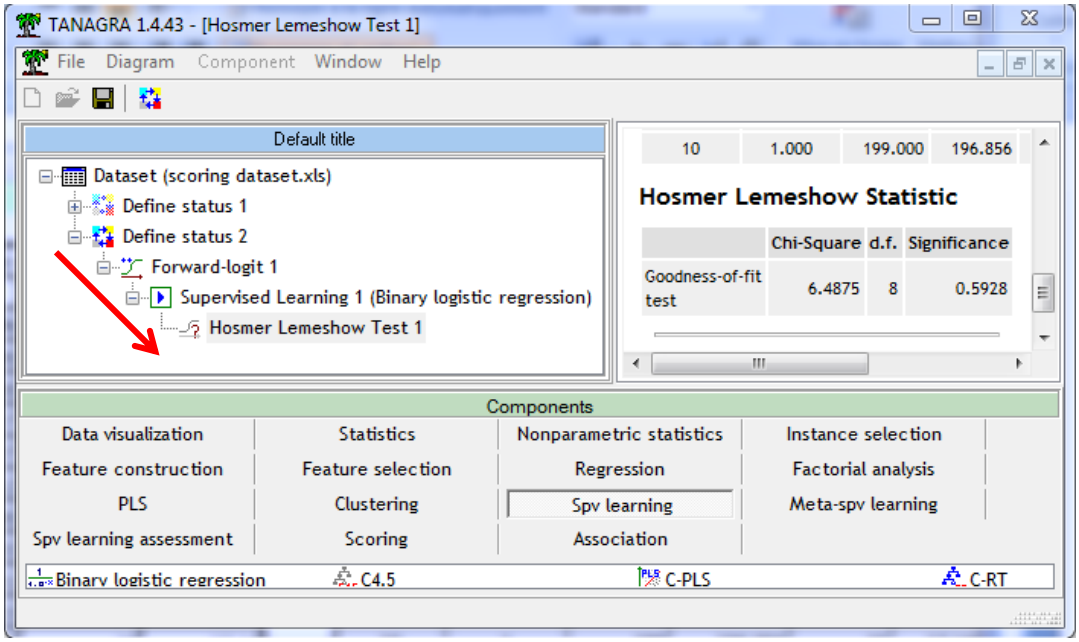
TANAGRA

Hosmer Lemeshow Goodness-of-Fit Test						
		Positive		Negative		
Decile	Prob.	Observed	Expected	Observed	Expected	Total
1	0.103	11	12.962	205	203.038	216
2	0.172	31	29.383	185	186.617	216
3	0.278	45	48.373	171	167.627	216
4	0.441	78	78.067	138	137.933	216
5	0.543	118	107.122	98	108.878	216
6	0.621	129	126.664	87	89.336	216
7	0.701	143	142.834	73	73.166	216
8	0.774	148	159.511	68	56.489	216
9	0.863	177	177.228	39	38.772	216
10	1	199	196.856	15	17.144	214

Hosmer Lemeshow Statistic

	Chi-Square	d.f.	Significance
Goodness-of-fit test	6.4875	8	0.5928

To obtain these results, we set the following diagram into Tanagra.



4 Conclusion

Incorporating advanced data mining techniques into a spreadsheet application is a valuable feature. It is available for Tanagra, for R (using RExcel). We describe in this tutorial the solution developed by SAS. But, unlike Tanagra[8], it seems that SAS has not planned a solution for the open source tools such as Open Office Calc.

[8] http://data-mining-tutorials.blogspot.fr/2011/07/tanagra-add-on-for-openoffice-calc-33.html