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. <u>Many people use a spreadsheet</u> in their data preparation phase. Recently, I have presented an add-in for the <u>connection between R and Excel</u>. In this document, I describe a similar tool for the SAS software.

SAS (<u>http://www.sas.com/</u>) 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).

C		12	(°ı •) ≠		scori	ing datas	et.xls [Mod	e de con	npatibilit	é] - Microsoft E	xcel			
	A		Insertion I	Mise en page	For	mules	Données	Révis	ion /	Affichage D	éveloppeur	Compléments	SAS 🔞	- 🗖 X
Do	années SAS	Tâche	Rapports Favo	ris Actualiser	N E Pr	lodifier ropriétés	Gérer le contenu	Outils	2 Aide				-	
			Analy <u>s</u> e des donne	ées de survie 🔸	ection	n		Outils						
	В		Analyse <u>m</u> ultivarié	e 🕨 🕨										×
	A		<u>A</u> NOVA	•	<u>й</u> ч	Analyse	de variance	à une d	imensi <u>o</u> n			I	J	K
1	object		Ca <u>p</u> abilité	•		Analyse	de variance	non-pa	ramétriqu	ie à une dimens	ion	p06rcy	p07rcy	p08rcy 🗖
2	positiv		<u>C</u> artes de contrôle	•	125	Modèle	s <u>l</u> inéaires				C	0.99	0.99	
3	positiv		Data M <u>i</u> ning	•	*	Modèle	s mixtes				C	0	0	
4	positiv		Description	•	Н	Test t					C	0	0	
5	positiv		Données	•				•	202	~	(0	0.99	
6	positiv		Graphique	•	L	0		0	4574	0.03	0.41	. 0	0	
7	negati		Pareto		L	0.99		0	4093	0	C	0	0	
8	negati		Démonsion		L-	0		0	123	0	C	0	0	
9	positiv		Regression			1		0	2422	0	C	0	1	
10	negati		Serie chronologiq	ue 🕨	_	1		0	357	0	C	0	0	
11	negati		Modèles de tâches	; +		1		0	543	0	C	0	0	
14	() H	data	set 🖉	-		-		-		· ∢ <u>m</u>		-		► I
Prê	t 🞦											100 % 🧲)	- + .::

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.

Sélectionner des données	-			-	X
Données d'entrée					
Onnées Excel :					
A1:GS2159					
Données SAS externes :					
			,	Parcourir	
✓ Détails				Filtrer & Trie	r
Emplacement des résultats					
Nouvelle feuille de calcul :	Analyse de v	rariance non-pa	aramét		
Feuille de calcul existante :					
Nouveau classeur					
Pourquoi est-il impossible d'indique	er un emplace	ment dans Exc	el en cliquant da	Annuler	lcul ? Aide

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.

Analyse p-values exactes Résultats Titres Propriétés	Source de données : D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparie on\logistic regression\sas add-in\scoring dataset xls!dataset Filtre de tâche : Néant
	Variables à attribuer : Nom
	posty posty posty t t t t
	Attribue la variable sélectionnée à la fonction que vous avez choisie dans la liste déroulante.

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

🖄 Analyse de variar	nce non-paramétrique à une dimension pour D:\DataMining\Databases_for_mining\dataset_f						
Données Analyse p-values exactes Résultats Titres Propriétés	Analyse Scores de test Ø Wilcoxon Ø Médiane Ø Savage						
	 Van der Waerden Supprimer la correction de continuité Ansari-Bradley Kotz Mood Siegel-Tukey Données brutes REMARQUE : les scores de test doivent être sélectionnés afin d'activer les p-values exactes de la page "P-values exactes" et les statistiques de la page "Résultats". 						
Aperçu du code	Attribue la variable sélectionnée à la fonction que vous avez choisie dans la liste déroulante. Exécuter Annuler Aide						

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.

🖄 Analyse de varian	ce non-paramétrique à une dimension pour D:\DataMining\Databases_for_mining\dataset_f
Données Analyse	Analyse
p-values exactes Résultats Titres	Scores de test
Aperçu du code de la t	âche
Insérer du code	
PROC NPARI ; VAR to CLASS /* Fin du	WAY DATA=WORK.TMPOTempTableInput WILCOXON MEDIAN SAVAGE VW AB KLOTZ MOOD ST EDF talspend; objective;
RUN; QUIT;	······································
Aperçu du code	Attribue la variable sélectionnée à la fonction que vous avez choisie dans la liste déroulante.

This feature is very similar to the one of the RATTLE package for R¹. 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-in². We set a nonparametric statistic analysis (e.g. <u>http://data-mining-tutorials.blogspot.fr/2008/11/nonparametric-statistics.html</u>). Into the DEFINE STATUS component, we set TOTALSPEND as TARGET and OBJECTIVE as INPUT).

¹ <u>http://data-mining-tutorials.blogspot.fr/2011/08/data-mining-with-r-rattle-package.html</u>

3.1.1 Wilcoxon-Mann-Whitney test

SAS computes the Wilcoxon statistic, Tanagra the Mann-Whitney one³. 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

3.1.2 Kruskal-Wallis test



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 test⁴ 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 (<u>http://en.wikipedia.org/wiki/Median_test</u>). Both SAS and Tanagra provide the results for the two approaches. SAS...

Scores médians (Nbre de points au-dessus de la médiane) pour la variable totalspend										
Classés par variable objective										
Somme des Attendue Ecart-type										
objective	N	scores	sous H0	sous H0	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.										

SAS

Analyse de variance non-paramétrique à une dimension

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						

...TANAGRA.

					Results							
Attribute_Y	Attribute_X			Statistical test								
		Value	Examples	Average	Scores	Scores	Two-San	nple Test				
					sum	mean	5	440.66667				
		positive	1079	1763.1909	638.3333	0.5916	E(S)	539.50000				
		negative	1079	992.8267	440.6667	0.4084	V(S)	134.77079				
						All	2158	1378.0088	1079.0	0.5000	Z	8.51345
totalspend	objective						p-value	0.00000				
		-					One-way	Analysis				
		IA	NAGR		Chi-Square	72.47882						
		« [\	Jedian	d.f.	1							
							p-value	0.00000				

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 test⁵ provides also the two kinds of results (based on the Z-statistic and on the chi-squared statistic).

SAS	Scores de Van der	Scores de Van der Waerden (Normal) pour la variable totalspend Classés par variable objective							
		Somme des	Attendue	Ecart-type	Score				
objective	N	scores	sous H0	sous H0	moyen				
positive	1079	223.95545	0	23.158361	0.207558				
negative	1079	-223.95545	0	23.158361	-0.207558				
Los scoros mouens ont été utilisés neur les liens									

Analyse de variance non-paramétrique à une dimension

							Results			
Test à deux échantillons de Van		Attribute_Y	Attribute_X	Description Statist					Statisti	cal test
der Waerden Statistique 223 9554				Value	Examples	Average	Scores	Scores	Two-San	nple Test
7	9 6706					Bo	sum	mean	5	-223.95545
Linilat (nal Du y 7	10001			positive	1079	1763.1909	223.9555	0.2076	E(S)	0.00000
Unilateral Pr > 2	<.0001			negative	1079	992.8267	-223.9554	-0.2076	V(5)	536,30966
Bilatéral Pr > Z	<.0001			All	2158	1378.0088	0.0	0.0000	171	9.67061
		totalspend	objective						n-value	0.00000
Analyse à une dir	nension								p relide	0100000
de Van der Waerden									One-way	Analysis
Khi-2	93.5207			Т		RΔ			Chi-Square	93.52068
DLL	1						d.f.	1		
Pr > Khi-2	<.0001								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									
Somme des Attendue Ecart-type Sco									
objective	N	scores	sous H0	sous H0	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 nour 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

Siegel and Tukey test 3.1.6

The Siegel and Tukey test⁶ is available into SAS only also. From here, we compare the differences in scale of the distributions (differences in location previously).

Scores Siegel-Tukey pour la variable totalspend								
Classés par variable objective								
Somme des Attendue Ecart-type Scor								
objective	N	scores	sous H0	sous H0	moyen			
positive	1079	1140335.6	1164780.5	14472.8912	1056.84486			
negative	1079	1189225.4	1164780.5	14472.8912	1102.15514			
l es scores movens ont été utilisés nour les liens								

Analyse de variance non-paramétrique à une dimension

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									
SAS Somme des Attendue Ecart-type Scc									
objective		N	scores	sous H0	sous H0	moyen			
positive		1079	570436.667	582660	7236.44447	528.67161			
negative		1079	594883.333	582660	7236.44447	551.32839			
	l es scores moyens ont été utilisés nour les liens								

s scores moyens ont	été utilisés	pour les liens.
---------------------	--------------	-----------------

Test à deux échantillons de Ansari-Bradley		Results									
		Attribute_Y	Attribute_X Description Statistical					stical test			
Statistique	570436.6667							Scores	Two-S	ample Test	
z	-1.6891			Value	Examples	Average	Scores sum	mean	c	570/36 666	63
Unilatéral Pr < Z	0.0456			positive	1079	1763.1909	570436.6666	528.6716	E(S)	582659.999	94
Bilatéral Pr > Z	0.0912			negative	1079	992.8267	594883.3333	551.3284	V(5)	52366128.472	.01
				All	2158	1378.0088	1165320.0	540.0000	Z	1.689	14
		totalspend	objective						p-value	0.091	19
Analyse a une din	nension								One-way	/ Analysis	
de Ansari-Bradley				TANAGRA				Chi-Square	2.85318		
Khi-2	2.8532								d.f.	1	
DLL	1								p-value	0.09119	

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

0.0912

Khi-2 DLL Pr > Khi-2

3.1.8 Klotz test

Scores Klotz Scores pour la variable totalspend									
Classés par variable objective									
Somme des Attendue Ecart-type									
objective	SAS	N	scores	sous H0	sous H0	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.									

The Klotz test is a nonparametric test for scale differences.

3.4196

0.0644

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

	Results								
Attribute_Y	Attribute_X			Statistical test					
	Value	Examples	Average	Scores sum	Scores mean	Two-San	nple Test		
		positive	1079	1763.1909	1131.6905	1.0488	5 E(S)	1072,29604	
	negative	1079	992.8267	1012.9016	0.9387	V(S)	1031.62442		
		All	2158	1378.0088	2144.6	0.9938	Z	1.84920	
totalspend	objective			p-value	0.06443				
				One-way Analysis					
			TA	Chi-Square	3.41956				
				d.f.	1				
							p-value	0.06443	

3.1.9 Mood test

Khi-2

DLL Pr > Khi-2

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).

	Scores Mood pour la variable totalspend											
	Classés par variable objective											
	SAS		Somme des	Attendue	Ecart-type	Score						
objective	0/10	N	scores	sous H0	sous H0	moyen						
positive		1079	431767608	418738590	8064157.45	400155.337						
negative		1079	405709571	418738590	8064157.45	376005.163						
		Les score	s moyens ont été uti	lisés pour les liens.								

Test à deux échantillo		Results									
Statistique	Attribute_Y	Attribute_Y Attribute_X Description Statistical test									
Z	1.6157			Value	Examples	Average	Scores sum	Scores	Tw	o-Sample Te	st
Unilatéral Pr > Z	0.0531			positive	1079	1763.1909	431767608.4863	400155.3369	5 E(S)	405709	571.57231 590.02930
Bilatéral Pr > Z	0.1062			negative	1079	992.8267	405709571.5723	376005.1636	V(S)	65030635439	095.49220
			All	2158	1378.0088	837477180.1	388080.2503	Z		1.61567	
Analyse à une din	totalspend	objective						p-value		0.10617	
de Mood									One-way Analysis		
Khi-2	2.6104			TANAGRA Chi-Square 2.6				Chi-Square	2.61039		
DLL	1							1			
Pr > Khi-2	0.1062			p-value 0						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 Tanagra⁷.

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

TANAGRA 1.4.43 - [Mann	-Whitney Comparison 1]			-						x
Tile Diagram Comp	💇 File Diagram Component Window Help									
🗅 📽 🔛 🎎										
Defau	Default title							-		
□		Results								
📄 🙀 Define status 1		<u> </u>	Value	Examples	Average	Rank sum	Rank mean	Mann-Whitney U	438660.00000	
Mann-Whitn	ey Comparison 1	-	positive	1079	1763.1909	1308241.0	1212.4569	E(U)	582120.50000	
Kruskal-Wall	Kruskal-Wallis 1-way ANOVA 1		negative	1079	992.8267	1021320.0	946.5431	V(U)	209465931.84330	
Median test 1			All	2158	1378.0088	2329561.0	1079.5000	Z	9.91233	-
Van der Wa	erden 1-way ANOVA 1							P(> Z)	0.00000	=
Misari-bradley scale Test 1										
Mood Scale Test 1		Computation time : 47 ms.								
		Created at 13/04/2012 08	8:10:55							T
			Components							
Data visualization	Statistics	Nonparametric statist	tics Instance s	election	Featur	e construct	ion	Feature selection	n	
Regression	Factorial analysis	PLS	Cluste	ring	Sp	ov learning		Meta-spv learning	2	
Spv learning assessment	Scoring	Association								
Correlation scatterplot	Export dataset	🖉 Scatterplot	🖉 Scatt	erplot with la	abel 🔣	View datas	et	👯 View mult	iple scatterplot	
		_		_	_					<i>a</i>

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.

6		127	२ (२ २) ₹	scoring d	ataset avec	results.xls	[Mode de c	ompatibili	ité] - Mic	rosoft Excel			x
	Ac	cueil	Insertion N	/lise en page	Formules	Données	Révision	Affichage	Dévelo	ppeur Complé	ments SAS) - 🗖	×
D	onnées SAS	Tâch	es Rapports	Favoris SAS -	(tualiser	🔊 Modifier 🔝 Propriété	Gérer le content	Outils	2 Aide				
			Analy <u>s</u> e des	données de	survie 🕨	ction		Outils		J			
	F		Analyse <u>m</u> ul	tivariée	•		1				1	1	×
	1		<u>A</u> NOVA		•	D	E		F	G	Н	I.	
1	object	1	Ca <u>p</u> abilité		•	3rcy	p04rcy	total	spend	p05spend	p05trans	p06rcy	
2	positi		<u>C</u> artes de co	ontrôle	•	1		0	4012	0	0		0
3	positi		Data M <u>i</u> ning)	•	0		0	13	0	0		-
4	positi		Description		•	0.95		0	2628	0	0		-
5	positi		<u>D</u> onnées		•	0		0	962	0	0		-1
6	positi		Graphique		•	0		0	4574	0.03	0.41		-1
7	negat		Pareto			0.99		0	4093	0	0		-8
8	negat		Pégression				ta tha ƙatao a	0	123	0	0		-
9	positi		<u>Regression</u>	laninus		Mo <u>d</u> e	ele lineaire g	eneralise	. 122	0	0		-
10) negat		Serie chrono	biogique		<u>Régre</u>	ession linéai	re	357	0	0		-
11	negat		Modèles de	tâches	•	Régre	ession logi <u>s</u> t	ique		0	0		
I4 Pr	4 ▶ ▶ êt 1	dat	aset An	alyse de vari	ance non-	🥼 Régr	SAS Ad Appuye	d-In 4.3 fo z sur F1 p	or Microso our obten	ft Office ir de l'aide.		•) .::

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

lonnées Iodèle	Données		
Réponse Effets Sélection Options	Source de données : D:\D add- Filtre de tâche : Néan	ataMining\Data in\scoring data t	bases_for_mining\dataset_for_soft_dev_and_comparison\logistic regression\sas set avec results.xls!dataset
rédictions ìtres ropriétés	Variables à attribuer : Nom /	•	Fonctions de la tâche : Variable dépendante (Linite : 1) Jacobie dépendante (Linite : 1) Jacobie de cha_
	Bilunivna first_month gender1 gender2 gender3 julpinicome		Variables quantitatives Unités Unités Ecart-type Cart-type Cart-t
	 Instantion Instantion		Chaspend Considered Con
	(2) p03rcy (2) p04rcy (2) p05spend	-	
**			

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

Effets Sélection Options Graphes Prédictions Titres Propriétés	Type de réponse : Type de modèle :	Binaire ▼ © logit © probit © log-log complémentaire © glogit	
Options Graphes Prédictions Titres Propriétés	Type de modèle :	 logit probit log-log complémentaire glogit 	
	objective :	negative positive	
	Ajuster le modèle au niveau : Indiquer le type de réponse. I votre variable de réponse. Si binaire. Si elle en contient plu	positive positive contract de réponses disponibles dépendent du nombre de niveaux dans cette demière ne contient que deux niveaux, le type de réponse est alors us, vous pouvez choisir une réponse de type classée ou non classée.	A

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

Données Modèle Réponse	Modèle > Effets			
Effets Sélection Options Graphes Prédictions Titres Propriétés	Variables de classe et var. qu p p01rcy p p02rcy p p03rcy p p04rcy p totalspend p p05spend p p05rans p p06rcy p p09rcy p p09rcy p p09rcy p p09rcy p p11rcy p p12rcy p p13rcy p p14rcy v	Antitatives : Principal Croiser Imbriquer Factoriel Degrés : 200 Polynomial Degrés : 2	Effets : p01rcy p02rcy p03rcy p04rcy totalspend p05spend p05rens p06rcy p07rcy p08rcy p08rcy p09rcy p09rcy p11rcy p12rcy Sup	orimer les effets
Aperçu du co	de	E	xécuter 🔻 Annuler	Aide

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 $\alpha = 1\%$.

Régression logist	tique pour D:\DataMining\Databases_for_mining\	dataset_for_soft_dev_and_comparison\logis
Données Modèle	Modèle > Sélection	
Réponse Effets	Méthode de sélection du modèle :	Effets à forcer dans le modèle
Sélection	Sélection ascendante	Si les éléments sont cochés dans la liste ci-dessous, ils deviendront
Graphes Prédictions	Niveaux de significativité	"sélectionnés" et seront transférés dans cette liste. Vous pouvez ensuite réorganiser les
Titres	Pour entrer dans le modèle : 0.01	éléments "sélectionnés" en les
Prophetes	Pour rester dans le modèle : 0.05	fléchés vers le haut et vers le bas.
		p01rcy
		p02rcy p03rcy
		p04rcy
		p05spend
		p05trans -
	Indiquez le niveau de significativité à utiliser pour saisir	r une variable quantitative (explicative) dans le modèle. 🔺
		_
[^{222]} A d d -		
Aperçu du code	•	Executer Annuler Aide

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.

odèle Réponse	Modele > Uppons							
Effets Sélection Options aphes	Détails sur les estimations Matrice de corrélation des paramètres estimés Matrice de covariance des estimations	Table de classification Table de classification P-values (points de rupture) :						
Prédictions Titres Propriétés	Evaluation de l'ajustement du modèle Statistiques d'influence Test d'adéquation de Hosmer et Lemeshow Tests d'ajustement de l'écart et de Pearson R ² généralisé	Saisissez un ou plusieurs nombres séparés par des espaces. Pa exemple : 0,2 0,3 0,5 0,7						
	Intervalle de confiance Paramètres Wald Vraisemblance du profil	Rapports de cotas conditionnels Wald Vraisemblance du profil						
	Niveau de confiance : 95% Calcule les intervalles de confiance pour le rapport de co Calcule les intervalles de confiance sur la base des tests	▼ tes. individuels de Wald.						

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.SORTTEMPTABLESORT							
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		Fréquence
ordonnée	objective	totale
1	negative	1079
2	positive	1079

	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			
	AIC: 2993.62	gender3	gender3			
	CHI-2:0.00	Chi-2: 397.887	Chi-2: 397.887			
1	d.f. : 0	p:0.0000	p:0.0000			
	p-value : 0.0000	-	-			
	AIC: 2576.00	productcount	productcount			
	CHI-2 : 419.63	Chi-2: 143.299	Chi-2:143.299			
2	d.f. : 1	p:0.0000	p:0.0000			
	p-value : 0.0000					
	AIC · 2422 99	bknfren	bknfren			
	CHI-2 : 574 63	Chi-2 : 54 575	Chi-2 : 54 575			
3	d f : 2	n : 0.0000	n : 0.0000			
	n.value : 0.0000	p : 0.0000	p : 0.0000			
	AIC + 2361 00	+f37	+f37			
	AIC . 2301.99					
4	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					
	AIC : 2313.22	p05trans	pU5trans			
5	CHI-2: 688.40	Cn1-2: 18.716	Cn1-2: 18./16			
	u.i.: 4	p:0.000	p:0.0000			
	p-value . 0.0000	abbénnarr	abbénnors			
	AIC . 2293.07					
6	df 15	CIII-2 . 13.003	ciii-2 . 13.865			
		p . 0.0002	p. 0.0002			
	AIC + 2280 93	+f68	+f68			
	CHI-2 · 724 69	Chi-2 · 14 344	Chi-2 · 14 344			
7	d.f. : 6	p : 0.0002	p : 0.0002			
	p-value : 0.0000	P	F · · · · · · · ·			
	AIC: 2268.53	amtfrench	amtfrench			
	CHI-2:739.09	Chi-2: 10.014	Chi-2 : 10.014			
8	d.f. : 7	p:0.0016	p:0.0016			
	p-value : 0.0000		·			
	AIC: 2260.39	p09tenure	p09tenure			
	CHI-2:749.24	Chi-2:9.440	Chi-2 : 9.440			
9	d.f. : 8	p:0.0021	p:0.0021			
	p-value : 0.0000					
	AIC: 2250.76	tf128	tf128			
10	CHI-2:760.86	Chi-2:9.480	Chi-2 : 9.480			
10	d.f. : 9	p:0.0021	p:0.0021			
	p-value : 0.0000					
	AIC: 2243.02	brlanglic	brlanglic			
11	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					
	AIC: 2236.49	p12rcy	p12rcy			
12	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					
	AIC: 2230.92	4	p02rcy			
	CHI-2:786.70	-	Chi-2 : 6.506			
13	d.f. : 12	-	p:0.0108			
			"p" higher than			
	p-value : 0.0000	-	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						
Statisti	ques d'ajustement d	u modèle				
Constante Constante						
	uniquement	et				
Critère		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	0.4073
		l'échelle max.	

Test de l'hypothèse nulle globale : BETA=0							
Test Khi-2 DDL Pr > H							
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

Adjuste	ment 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					Tanagra						
Estimations p	ar l'aı	nalyse du i	maximur	n de vrais	emblance		Attributes in the equation				
		Valeur	Erreur	Khi-2		_					
Paramètre	DDL	estimée	type	de Wald	Pr > Khi-2		Attribute	Coef.	Std-dev	Wald	Signif
Intercept	1	-1.9280	0.2419	63.5181	<.0001		constant	-1.9280	0.2419	63.5182	0.0000
ahh6ppers	1	-5.9698	1.9885	9.0125	0.0027		ahh6ppers	-5.9698	1.9885	9.0125	0.0027
amtfrench	1	2.7341	0.7459	13.4352	0.0002		amtfrench	2.7341	0.7459	13.4352	0.0002
bknfren	1	-8.0473	1.4203	32.1021	<.0001		bknfren	-8.0473	1.4203	32.1021	0.0000
brlanglic	1	2.2944	0.7998	8.2292	0.0041		brlanglic	2.2944	0.7998	8.2292	0.0041
gender3	1	-1.9310	0.1188	264.3180	<.0001		gender3	-1.9310	0.1188	264.3180	0.0000
p05trans	1	-4.5013	1.2440	13.0927	0.0003		p05trans	-4.5013	1.2440	13.0927	0.0003
p09tenure	1	26.8724	14.3487	3.5074	0.0611		p09tenure	26.8725	14.3488	3.5074	0.0611
p12rcy	1	0.5115	0.1886	7.3549	0.0067		p12rcy	0.5115	0.1886	7.3549	0.0067
productcount	1	0.1970	0.0202	95.1812	<.0001		productcount	0.1970	0.0202	95.1812	0.0000
tf128	1	17.6755	5.9650	8.7805	0.003		tf128	17.6755	5.9650	8.7805	0.0030
tf37	1	0.0443	0.0073	36.5450	<.0001		tf37	0.0443	0.0073	36.5450	0.0000
tf68	1	0.0003	0.0001	10.3427	0.0013		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										
	Valeur Intervalle de									
Effet	Unité	estimée	confianc	e 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 test	s de Hosmer	et de Lemesh	ow					
Groupe	Total	objective	= positive	objective	= negative					
		Observé	Attendu	Observé	Atten					
1	216	11	12.96	205	203.					
2	216	31	29.38	185	186.					
3	216	45	48.37	171	167.					
4	216	78	78.07	138	137.					
5	216	118	107.12	98	108.					
6	216	129	126.66	87	89.					
7	216	143	142.83	73	73.					
8	216	148	159.51	68	56.					
9	216	177	177.23	39	38.					
10	214	199	196.86	15	17.					

t	s de Hosmer e	et de Lemesho	ow		Hosm	nei
•	= positive	objective =	negative			
	Attendu	Observé	Attendu	Decile	Prob.	0
L	12.96	205	203.04	1	0.103	
L	29.38	185	186.62	2	0.172	
5	48.37	171	167.63	3	0.278	
3	78.07	138	137.93	4	0.441	
3	107.12	98	108.88	5	0.543	
)	126.66	87	89.34	6	0.621	
3	142.83	73	73.17	7	0.701	
3	159.51	68	56.49	8	0.774	
7	177.23	39	38.77	9	0.863	
)	196.86	15	17.14	10	1	

TANAGRA	

Hosmer Lemeshow Goodness-of-Fit Test									
		Pos	itive	Nega					
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			

Test d'adéquation de HosmerKhi-2DDLPr > Khi-26.487580.5928

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.

TANAGRA 1.4.43 - [Hosm	er Lemeshow Test 1]						8	
Tile Diagram Compo	ment Window Help					- 8	×	
D 📽 🔛 🗱								
	10	1.000	199.0	00 196.856				
⊡∰ Dataset (scoring da ⊕‱ Define status 1	Hosmer Lemeshow Statistic							
📄 🚰 Define status 2		Chi-Square	d.f.	Significance				
E	Goodness-of-f test	ït 6.4875	8	0.5928				
Hosm	· · · · · · · · · · · · · · · · · · ·							
Components								
Data visualization	Statistics Nonparame		tric statistics Instance sel		selec	tion		
Feature construction	Feature construction Feature selection		Regression		Factorial analysis			
PLS	Spv learning		Meta-spv learning					
Spv learning assessment	Asso	ciation						
Binary logistic regression			🎊 C-PLS			🚖 C-RT		
					_			

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 Tanagra⁸, 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