1 Objectif

Description de la macro complémentaire (add-in) SAS version 4.3 pour Excel. Comparaison des résultats avec ceux de Tanagra.

J'avais vu à la télé, il y a un moment déjà, la biographie d'un homme politique français célèbre. Au crépuscule de sa vie, il se livrait sur son parcours, ses combats. Il a alors prononcé une phrase qui m'avait beaucoup marqué : « avec l'âge, soit on se redit, soit on se contredit » ; et il avait ajouté non sans malice « je crois que j'appartiens plutôt à la première catégorie ». Bon, je ne suis pas encore à l'heure des bilans, loin s'en faut, mais il n'en reste pas moins que j'ai quelques convictions bien ancrées, et j'aime bien les ressortir de temps en temps au risque de me répéter (de radoter).

Entres autres, je pense que la connexion directe entre un logiciel de data mining et un tableur est une idée forte, parce que le tableur est un acteur incontournable de la manipulation des données pour les data miners[1]. Et… je ne suis pas le seul à le penser (ouf ! c'est toujours rassurant de savoir que d'autres partagent votre avis). Il n'y a pas longtemps j'avais présenté la solution RExcel pour le logiciel R. Dans ce tutoriel, je décris l'add-in SAS 4.3 (la macro complémentaire SAS version 4.3) pour Excel. Si SAS s'y est mis, c'est qu'il y a réellement une attente derrière. Personne ne peut en douter.

Le logiciel SAS est bien connu des statisticiens (http://www.sas.com/). Il est présent sur le marché des logiciels de statistique depuis un grand nombre d'années maintenant[2]. Il jouit d'une excellente réputation. Son principal défaut, outre le fait qu'il n'est pas accessible gratuitement, est qu'il faut connaître les instructions SAS, et de manière plus générale le langage de macro-commandes, pour pouvoir réellement l'exploiter.

SAS propose plusieurs solutions pour dépasser cet écueil. Entres autres, il a développé une macro complémentaire (add-in en anglais) pour la suite Office de Microsoft[3]. Je l'ai découvert très récemment sur les machines des salles informatiques de notre département (Département Informatique et Statistique – Université Lyon 2 – http://dis.univ-lyon2.fr/). Je me suis intéressé en particulier à l'add-in dévolue au tableur Excel. De fait, 3 tâches pas toujours évidentes à mettre en œuvre dans la version standard de SAS sont très largement facilitées : l'importation d'un fichier Excel dans SAS, le paramétrage et le lancement des techniques statistiques, la récupération des résultats dans le tableur aux fins de visualisation ou d'élaboration des rapports.

---

[1] N'en déplaise aux allergiques à Excel, ce dernier est un outil majeur de la pratique du data mining (cf. « Data Mining/Analytic Tools Used », Kdnuggets Polls, 2011 et 2010). Je me suis toujours posé la question d'ailleurs. Est-ce que cette défiance repose sur le rejet de Microsoft, ou sur le rejet des tableurs en général ? Je n'ai jamais compris en vérité. Je pense surtout qu'il s'agit d'un faux débat. Notre rôle consiste à choisir l'outil le plus adapté compte tenu des objectifs de notre étude, des caractéristiques de nos données, et des circonstances. Toute autre considération ne me paraît pas très défendable. Je le dis d'autant plus volontiers que je passe mon temps à défendre R (un autre objet de culte) auprès de ceux qui ne jurent que par Excel.

[2] http://en.wikipedia.org/wiki/SAS_%28software%29

[3] http://support.sas.com/documentation/onlinedoc/addin/index.html. Plusieurs tutoriels PDF décrivent l'installation et la mise en œuvre de la macro complémentaire (ex. statistiques descriptives, régression linéaire, etc.).

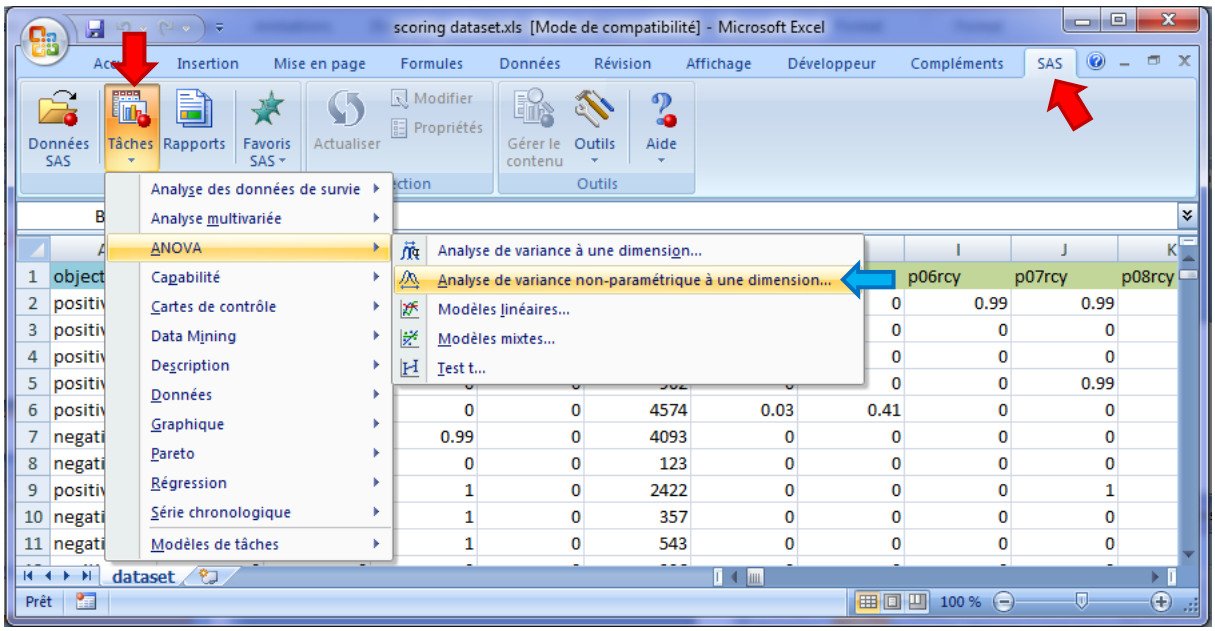
Dans ce tutoriel, nous décrivons le comportement de la macro complémentaire lors de la mise en œuvre des tests non paramétriques de comparaisons de populations et de la régression logistique avec sélection de variables. Nous mettrons en parallèle les résultats obtenus avec le logiciel Tanagra. L'idée est de comparer les calculs et le mode de présentation des résultats.

2 Données

Nous utilisons les données « scoring dataset.xls »[4]. Il comporte 2158 observations et 201 variables. La variable « objective » joue un rôle particulier. Les positifs (objective = positive) correspondent aux individus qui ont répondu positivement à une campagne de mailing direct. Nous chargeons les données dans Excel 2007.

3 Utilisation de l''add-in SAS 4.3

Au démarrage d'Excel 2007, nous disposons d'un onglet supplémentaire SAS dans le ruban supérieur. Les techniques statistiques sont disponibles dans le menu TACHES.



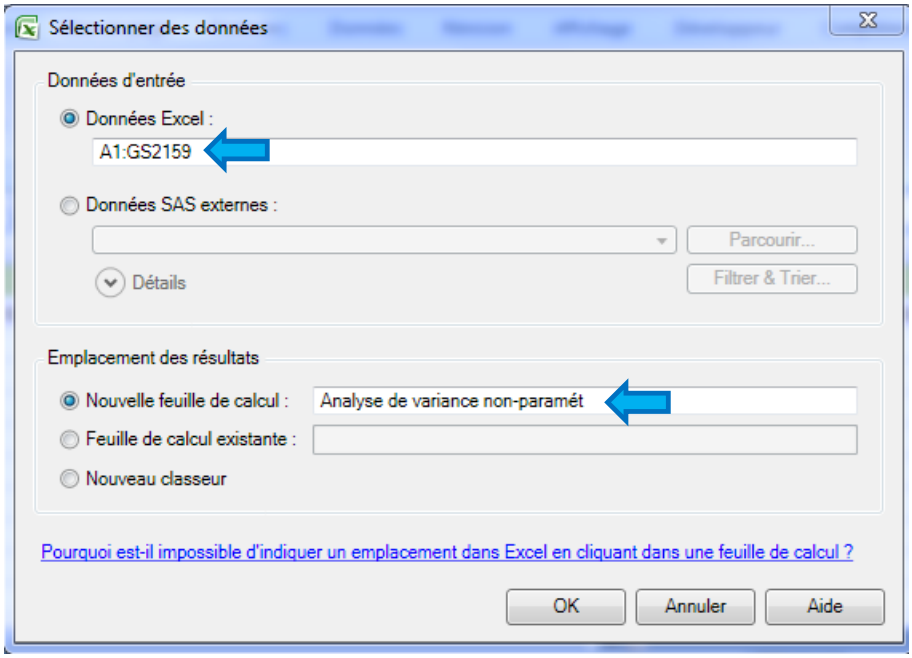
3.1 Tests non paramétriques

Dans cette section, nous cherchons à comparer les dépenses des clients (« total spend ») selon leur réponse à la sollicitation marketing. Après avoir sélectionné une des cellules de la plage de données, nous actionnons le menu TACHES / ANOVA / ANALYSE DE VARIANCE NON PARAMETRIQUE A UNE DIMENSION.

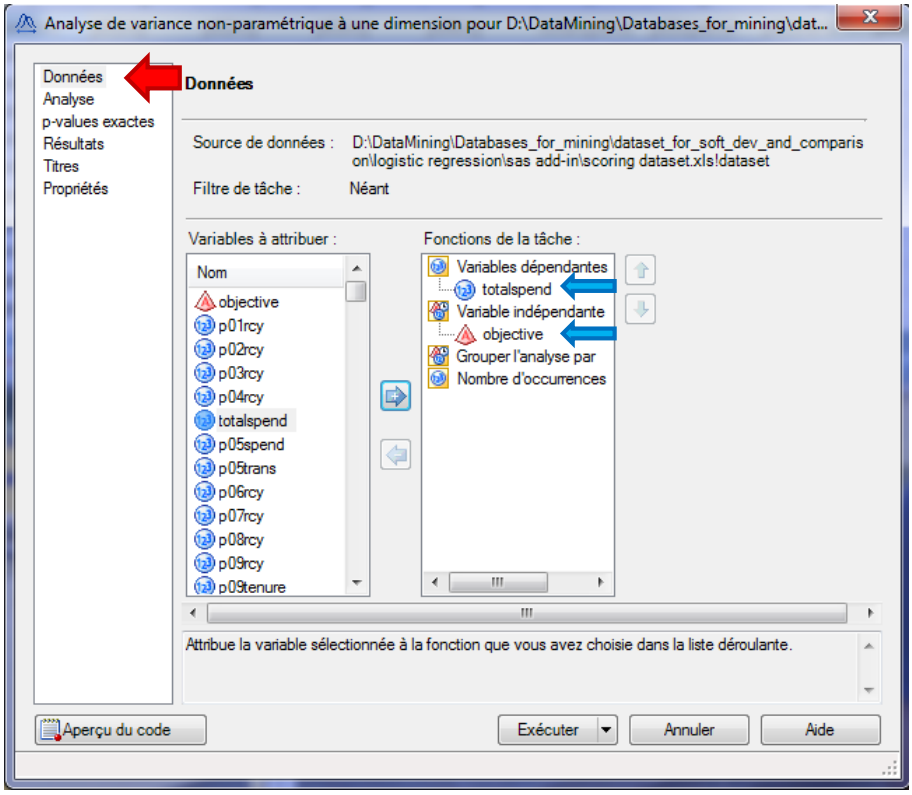
Une boîte de dialogue apparaît. Elle permet de préciser la plage des données (A1 :GS2159) et l'intitulé de la feuille dans laquelle sera affichée les résultats des calculs. Nous validons en cliquant sur le bouton OK.

Remarque : La connexion est un peu longue la première fois. Il faut patienter simplement.

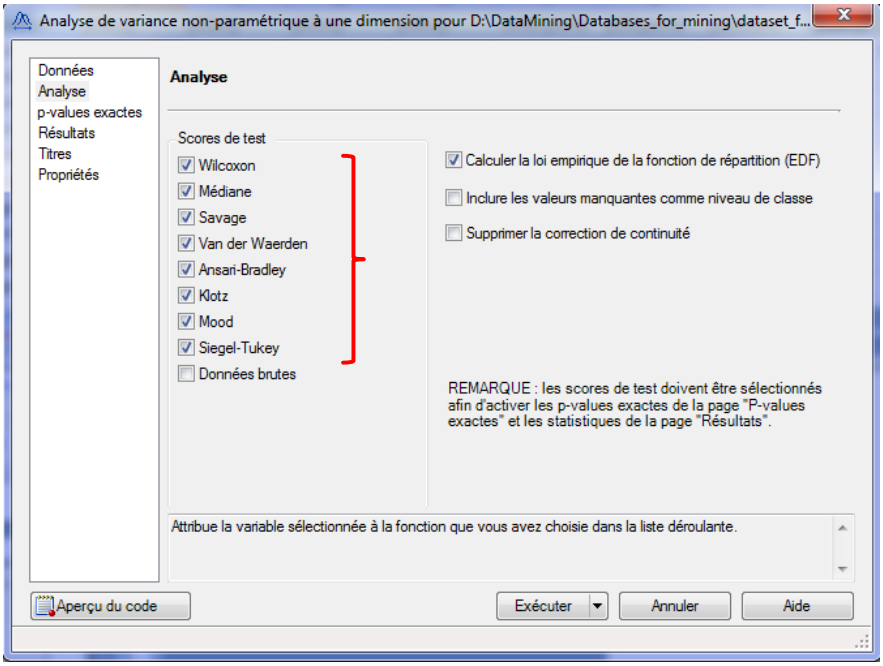
[4] http://www.math.mcmaster.ca/peter/sora/case_studies_00/etudes_de_cas.html



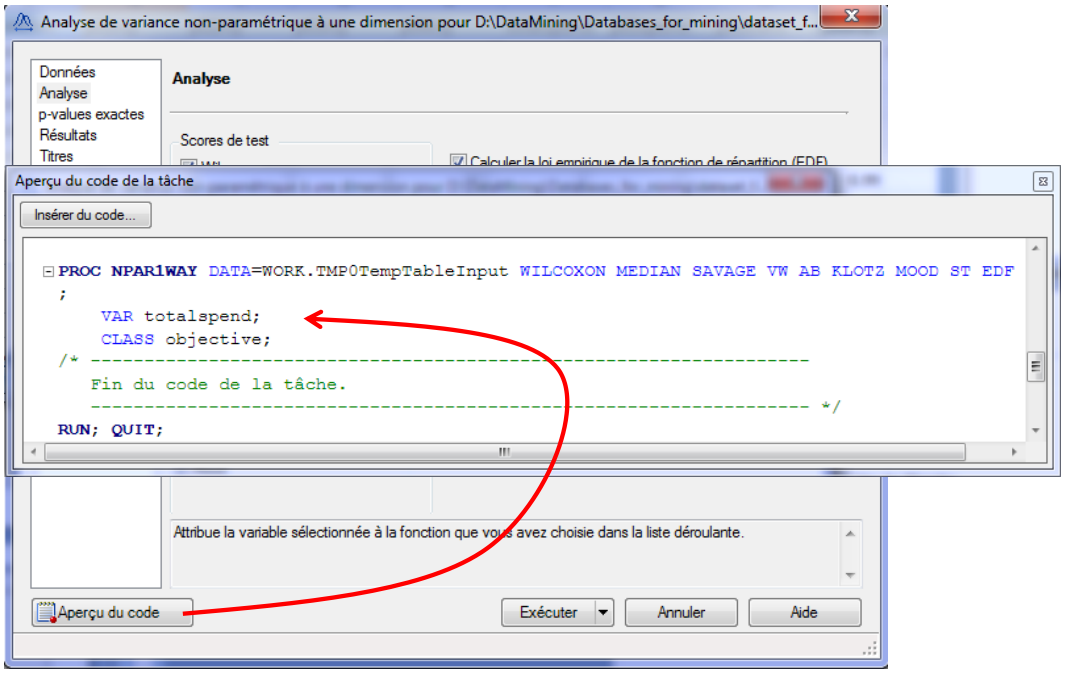
Une seconde boîte de dialogue permet de paramétrer la technique. Dans un premier temps, nous sélectionnons les variables de l'analyse dans l'onglet DONNEES. Nous plaçons OBJECTIVE en variable indépendante, et TOTALSPEND en dépendante.



Dans la page « ANALYSE », nous spécifions les tests à réaliser. Nous les sélectionnons tous à l'exception de « Données brutes ».



Dans « P-VALUES EXACTES », nous avons l'opportunité d'utiliser les lois exactes pour asseoir la décision (rejet ou non de l'hypothèse nulle). Attention, les calculs peuvent être très longs, surtout sur un fichier avec un grand nombre d'observations. Dans notre cas, cette option n'est pas pertinente. Enfin, les autres pages servent à préciser la nature des sorties. Nous les ignorons.



Notons une option très intéressante, il est possible de consulter les instructions SAS générées pour l'analyse en cliquant sur le bouton APERCU DU CODE. Nous retrouvons une fonctionnalité proposée par les packages du logiciel R qui permettent de piloter par menu les analyses (ex. le package RATTLE[5]). Nous pouvons modifier manuellement les instructions pour affiner notre analyse.

[5] http://tutoriels-data-mining.blogspot.fr/2010/06/data-mining-sous-r-le-package-rattle.html

Remarque : Précisons que avons souhaité obtenir des sorties au format HTML dans ce didacticiel (menu OUTILS / OPTIONS) : pour d'une part, bénéficier d'une mise en forme plus attrayante dans la feuille de résultats ; et d'autre part, parce que les sorties standards n'ont pas fonctionné lorsque j'ai utilisé la régression logistique dans la section suivante.

Il ne nous reste plus qu'à actionner le bouton EXECUTER. Les résultats des calculs sont insérés dans une nouvelle feuille du classeur. Voyons-en le détail, et comparons-les avec ceux de Tanagra.

Dans Tanagra, TOTALSPEND est la variable cible TARGET, OBJECTIVE est la variable d'entrée INPUT. Les tests non paramétriques sont regroupés dans l'onglet NONPARAMETRIC STATISTICS. Les méthodes abordées dans cette section sont décrites dans un ouvrage libre accessible en ligne (R. Rakotomalala, « Comparaison de populations – Tests non paramétriques », Université Lyon 2, 2008)[6].

3.1.1 Test de Wilcoxon-Mann-Whitney

Nous comparons les caractéristiques de tendances centrales des distributions conditionnelles. SAS calcule la statistique de Wilcoxon, Tanagra celle de Mann-Whitney. A la sortie, les deux procédures obtiennent la même statistique centrée-réduite |Z| = 9.91233. Au regard de la taille de notre échantillon, la correction de continuité introduite par SAS n'est pas 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

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

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

TANAGRA

[6] Voir aussi http://fr.wikipedia.org/wiki/Test_%28statistique%29 pour le positionnement des différents tests.

3.1.2 Test de Kruskal-Wallis

Le résultat du test de Kruskal-Wallis est fourni dans la foulée par SAS. Dans TANAGRA, nous utilisons un composant dédié.

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

3.1.3 Test de la médiane

Deux approches sont disponibles pour le test de la médiane. La première est basée sur la statistique de rangs. Elle est asymptotiquement normale. La seconde sur un tableau de contingence. Elle suit une loi du KHI-2 sous l'hypothèse nulle. 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

SAS

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

…et TANAGRA proposent les deux résultats.

Results

Attribute_Y	Attribute_X	Description				
		Value	Examples	Average	Scores sum	Scores mean
totalspend	objective	positive	1079	1763.1909	638.3333	0.5916
		negative	1079	992.8267	440.6667	0.4084
		All	2158	1378.0088	1079.0	0.5000

Statistical test	
Two-Sample Test	
S	440.66667
E(S)	539.50000
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 »

3.1.4 Test de Van der Waerden

Le résultat est double également pour le test de Van der Waerden.

SAS

Analyse de variance non-paramétrique à une dimension

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

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

TANAGRA

3.1.5 Test de Savage

Le test de Savage est présent uniquement dans SAS. Voilà un test à rajouter dans la TODO LIST de Tanagra donc. Il s'agit tout simplement de modifier le score utilisé pour le calcul des statistiques.

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

3.1.6 Test de Siegel et Tukey

Idem, le test de Siegel et Tukey est présent uniquement dans SAS pour l'instant. Attention, la finalité des tests est modifiée à partir d'ici : il s'agit de comparer les caractéristiques de dispersion dans les deux sous-populations.

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 Test de Ansari-Bradley

Ce test est présent dans Tanagra. Les résultats sont cohérents avec ceux de SAS bien évidemment.

SAS

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.					

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

TANAGRA

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

3.1.8 Test de Klotz

Ce test est également présent dans les deux logiciels.

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 Test de Mood

Il s'agit du test de comparaison des caractéristiques d'échelles (MOOD SCALE TEST dans Tanagra), à ne pas confondre avec le test des séquences (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

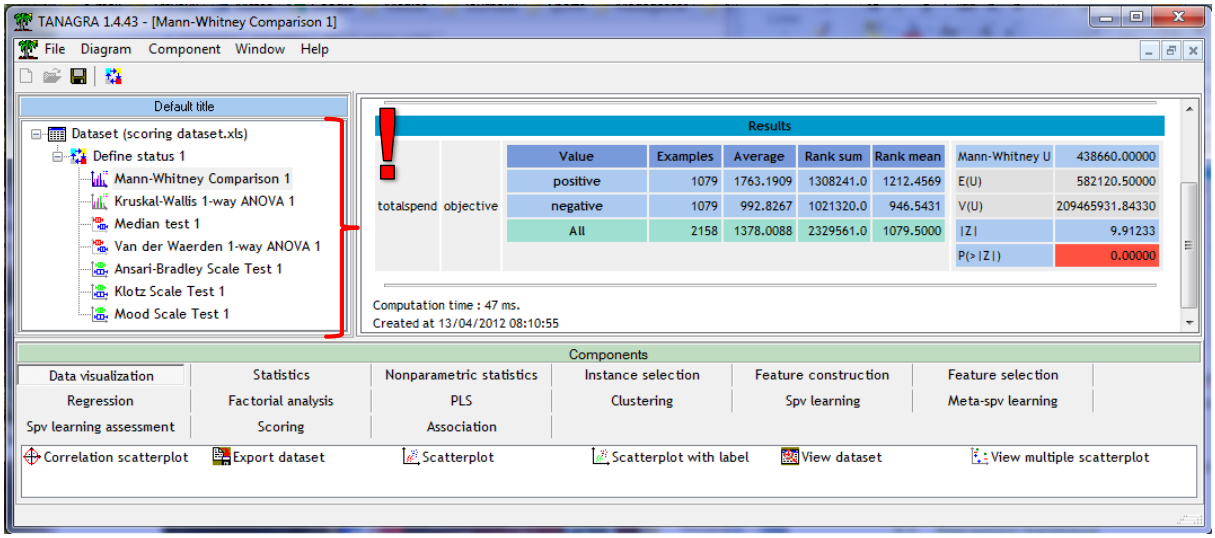
Results

Attribute_Y	Attribute_X	Description					Statistical test	
		Value	Examples	Average	Scores sum	Scores mean	Two-Sample Test	
totalspend	objective	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
							|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 fournit en plus les tests de Klomogorov-Smirnov et de Cramer-von Mises qui ne sont pas présents encore dans Tanagra, mais que nous avons décrit dans notre support de cours.

3.1.10 Diagramme Tanagra

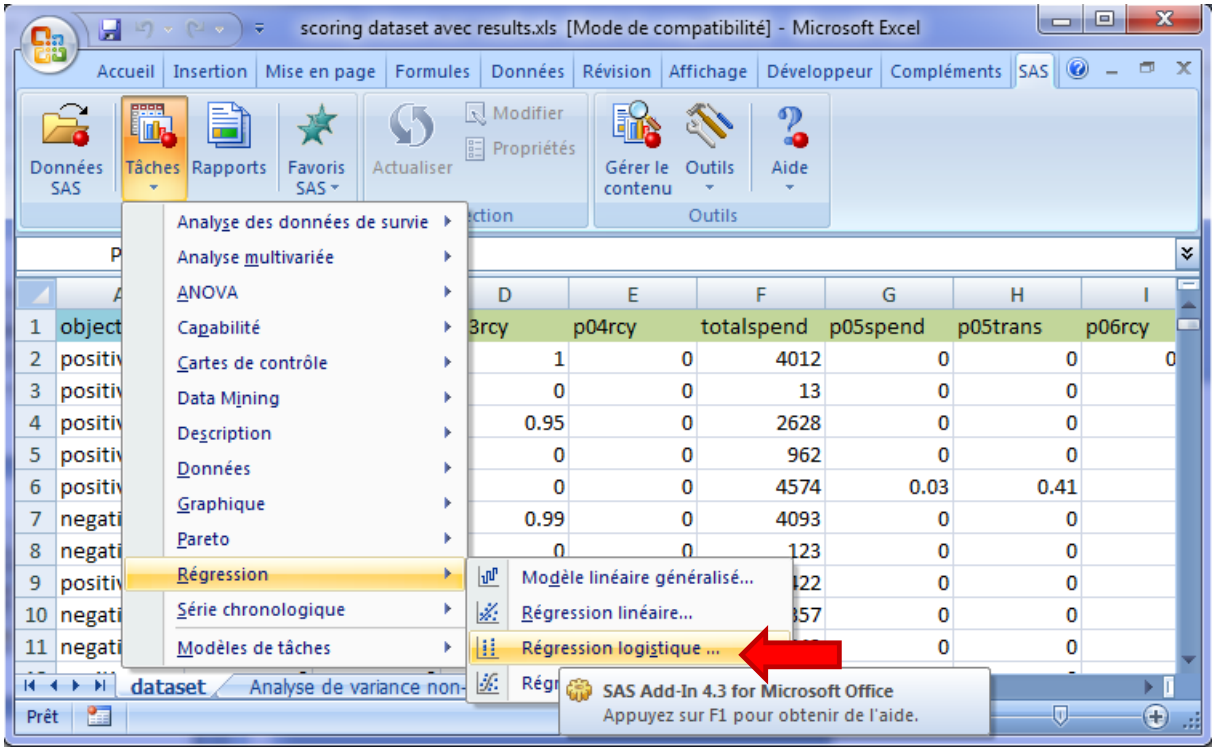
Pour réaliser ces analyses, nous avons élaboré le diagramme de traitements suivant sous Tanagra.



3.2 Régression logistique

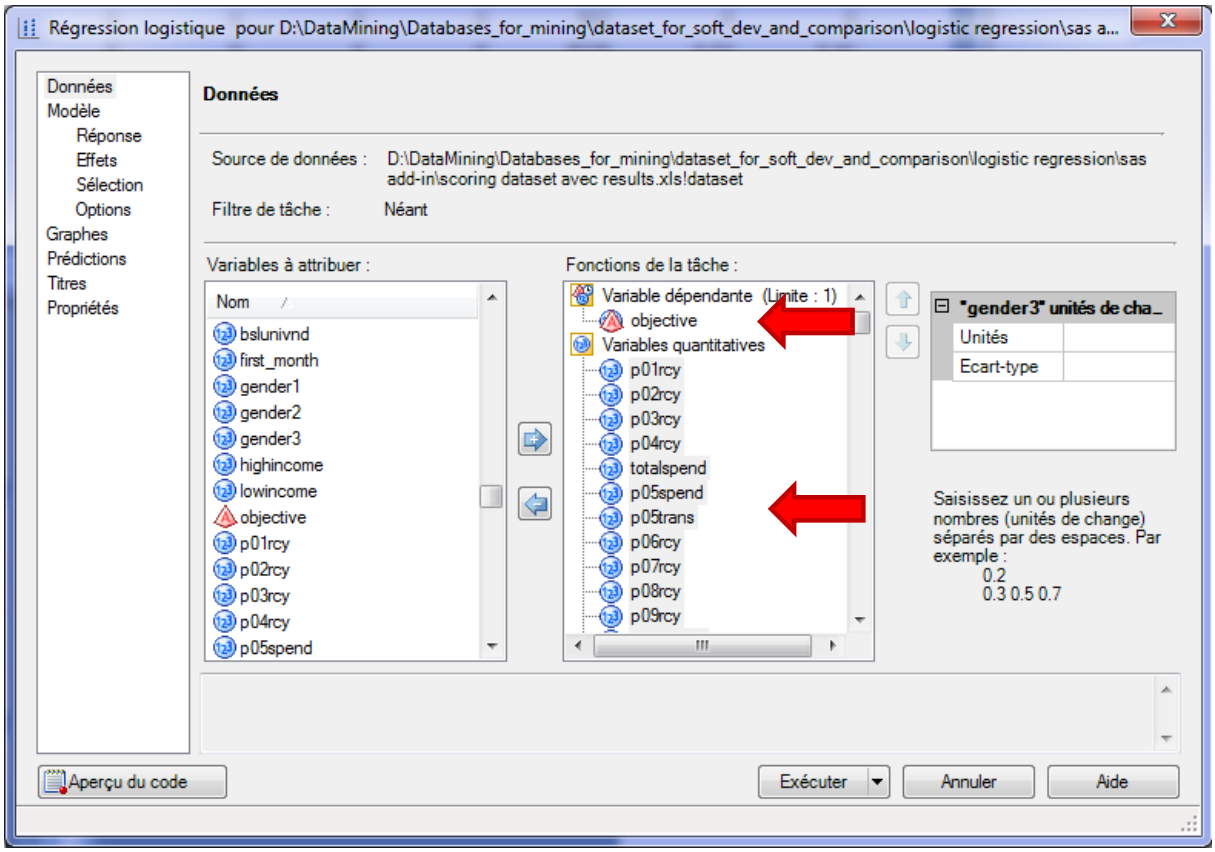
Dans cette section, nous cherchons à expliquer (prédire) le mieux possible la variable cible OBJECTIVE à partir des autres variables en utilisant la régression logistique (pour le détail de la méthode, voir R. Rakotomalala, « Pratique de la régression logistique – Régression logistique binaire et polytomique », 2011). Le problème de l'estimation des paramètres du modèle est couplé avec un processus de sélection de variables. L'affaire n'est pas triviale. En effet, il y a un nombre assez important de variables candidates (200), plusieurs d'entres elles sont certainement non pertinentes ou redondantes. Nous ne devrions retenir qu'un nombre réduit de variables prédictives à la sortie.

Revenons dans la feuille « dataset » dans notre classeur Excel. Toujours en veillant à ce qu'une des cellules de la plage de données soit activée, nous actionnons le menu SAS / TACHES / REGRESSION / REGRESSION LOGISTIQUE.

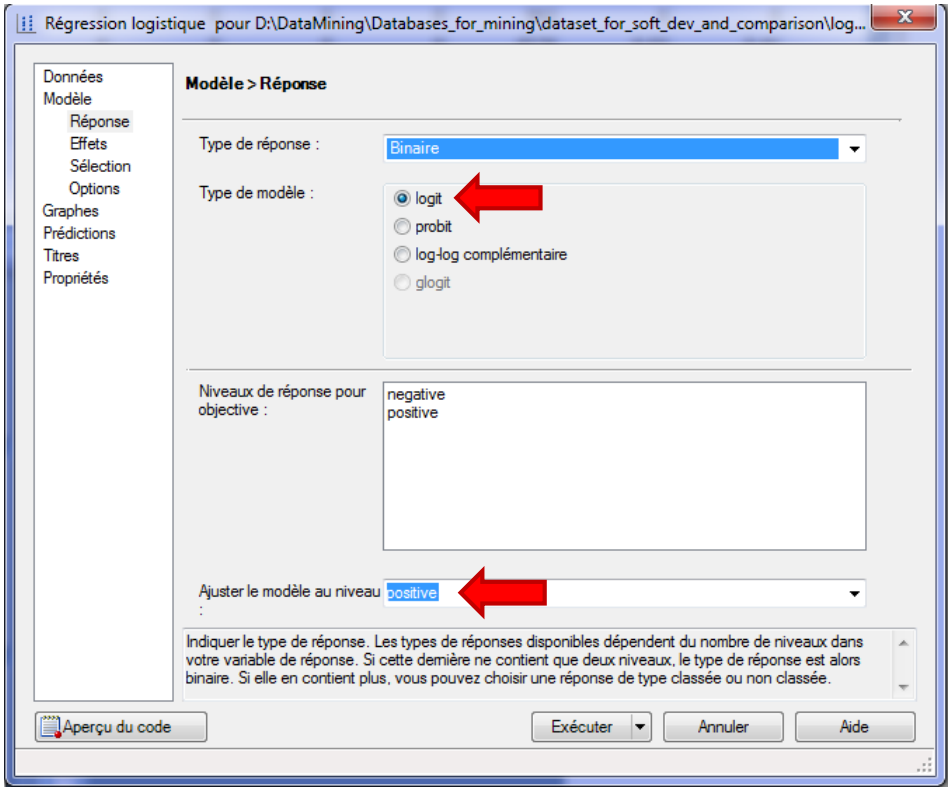


Comme précédemment, une boîte de dialogue permet de préciser l'ensemble de données utilisé et l'emplacement des résultats. Nous validons. La boîte de paramétrage apparaît.

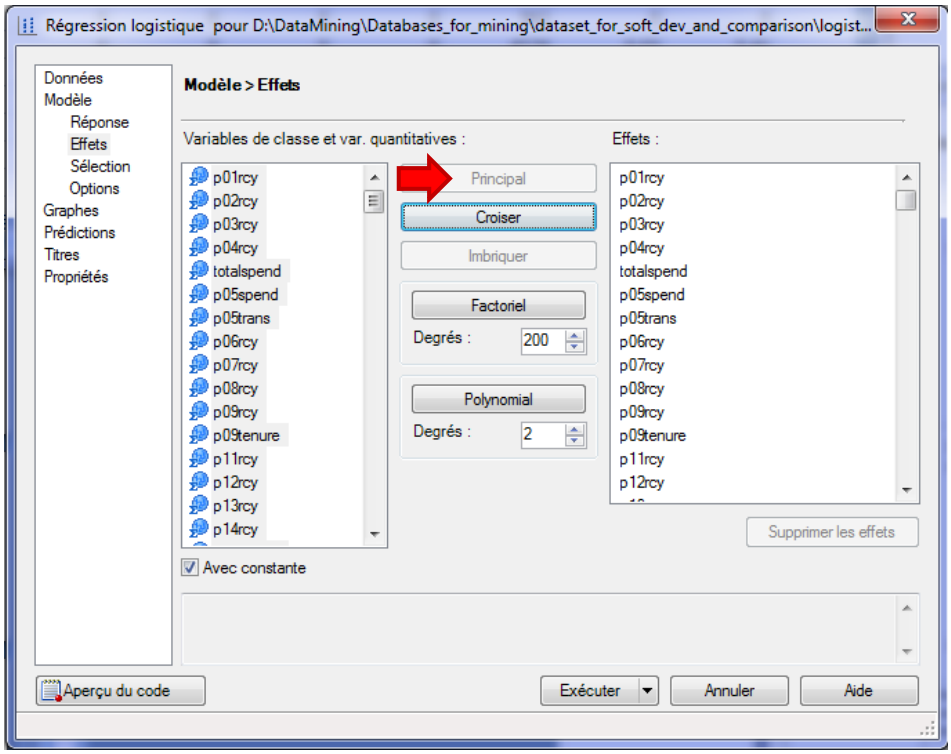
Dans DONNEES, nous définissons le rôle des variables. OBJECTIVE est la variable dépendante, les autres correspondent aux variables quantitatives.



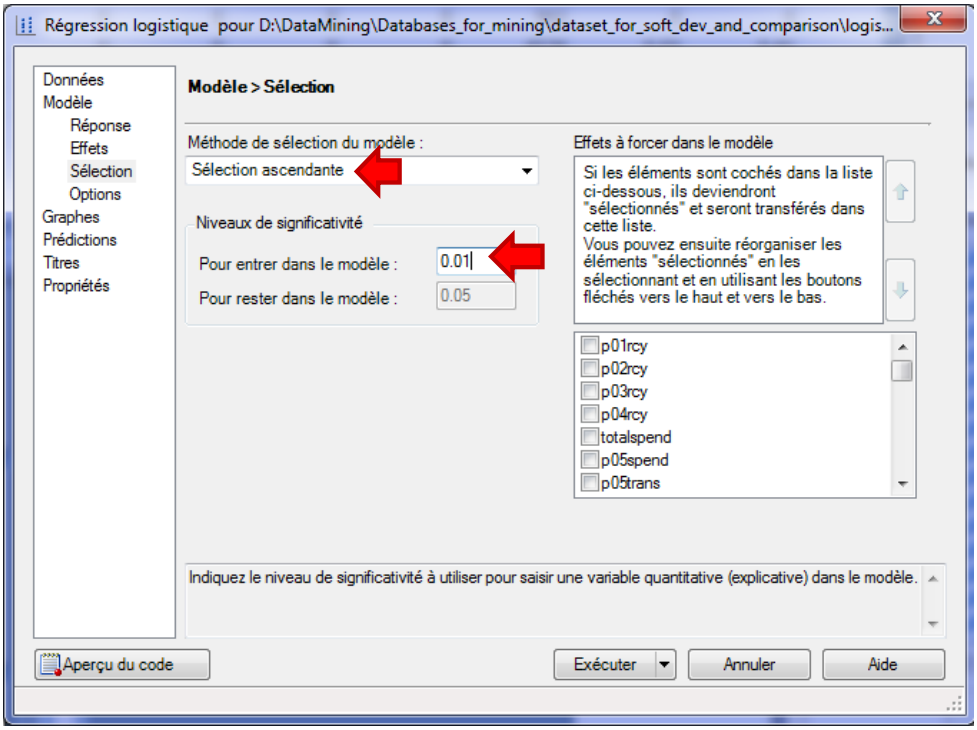
Dans l'onglet MODELE / REPONSE, nous spécifions le type de modèle LOGIT, nous indiquons également la modalité positive de la variable cible.



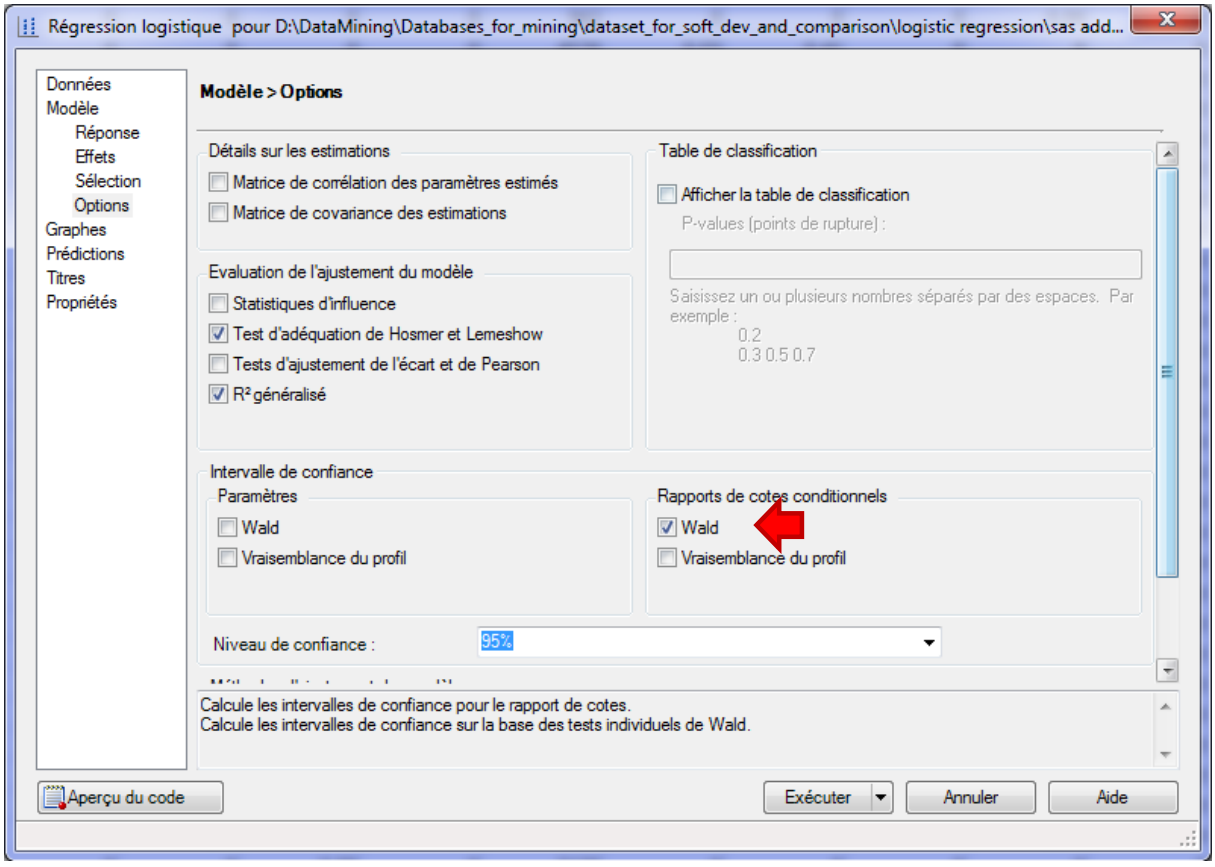
Dans MODELE / EFFETS, les variables explicatives quantitatives correspondent à l'effet PRINCIPAL. Notons qu'il est possible d'implémenter des expressions plus sophistiquées des variables (croisement, passage à la puissance, imbrication).



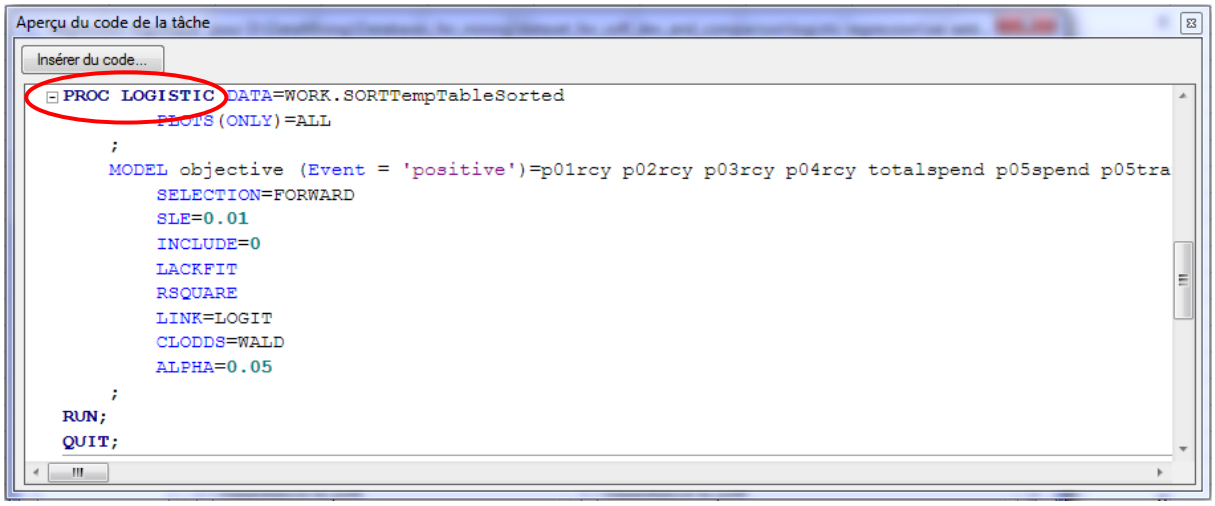
Dans MODELE / SELECTION, nous indiquons la technique de sélection de variables : une sélection ascendante (FORWARD), basée sur le test des scores dans SAS, avec un risque critique α = 1%.



Enfin, dans l'onglet MODELE / OPTIONS, nous spécifions les options supplémentaires pour compléter les sorties de l'analyse. Nous demandons, entres autres, les intervalles de confiance des odds-ratio.



Ici également, en cliquant sur le bouton APERCU DU CODE, nous pouvons visualiser les instructions en langage SAS.



Il ne nous reste plus qu'à lancer l'analyse en actionnant le bouton EXECUTER.

Une nouvelle feuille est insérée dans notre classeur. Voyons en le détail.

Un **résumé** indique les principales caractéristiques de l'étude. Nous constatons ainsi que notre échantillon est équilibré (50% de positifs et 50% de négatifs).

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

Ensuite, nous avons le **détail de la sélection ascendante**. Nous ne montrons que le résumé de la procédure dans ce tutoriel. Les valeurs de la statistique de test [test des scores] à chaque étape du processus sont strictement identiques à celles fournies par le composant FORWARD-LOGIT (onglet FEATURE SELECTION) de Tanagra.

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

12 variables prédictives sont sélectionnées en définitive.

SAS fournit **les indicateurs de qualité globale du modèle** : critère AIC (Akaike), BIC, test du rapport de vraisemblance, etc. Les sorties de SAS sont particulièrement exhaustives.

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

Nous disposons ensuite des **coefficients du modèle**. SAS les énumère dans l'ordre des variables initiales, Tanagra dans l'ordre de la sélection. Nous les avons triés selon le nom des variables pour les comparer. Les caractéristiques obtenues (coefficient estimé, écart-type, statistique de Wald, probabilité critique) sont bien les mêmes.

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

SAS produit également les odds-ratio et leur intervalle de confiance à 95%.

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

Enfin, **le test d'adéquation de Hosmer-Lemeshow** teste la compatibilité du modèle avec les données.

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

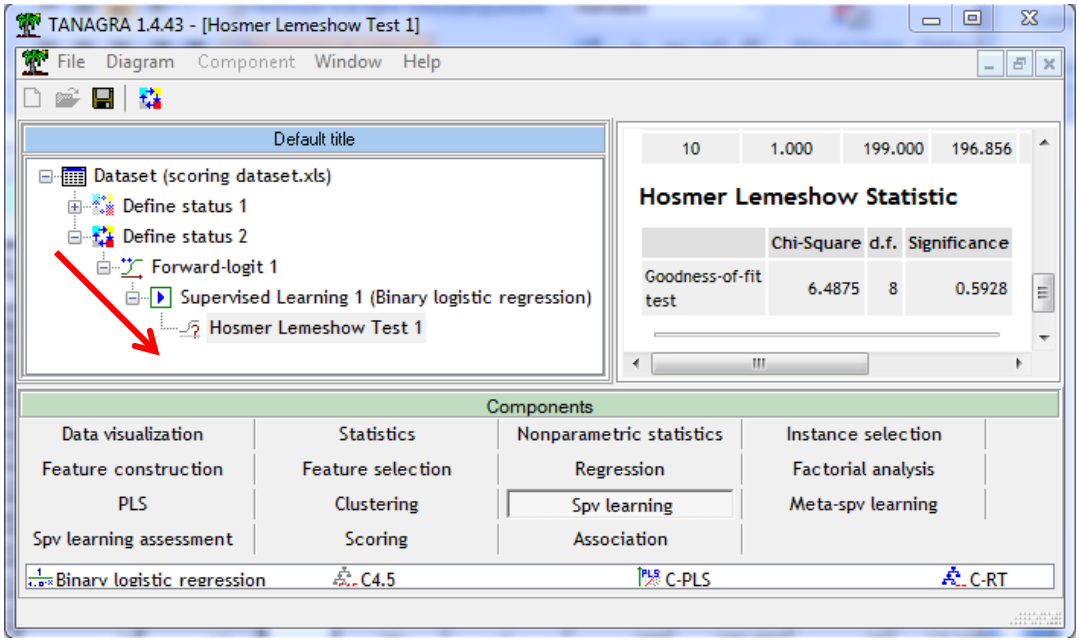
TANAGRA

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

Pour obtenir ces résultats, nous avons construit le diagramme de traitements suivant dans Tanagra.



4 Conclusion

Incorporer des fonctionnalités statistiques avancées dans Excel est un créneau que plusieurs éditeurs de logiciels ont investi depuis longtemps (XLSTAT, XLMINER, etc.). L'idée est suffisamment bonne pour que SAS vienne se positionner sur le créneau. Il apporte ses propres spécificités : une bibliothèque de calculs très riche (avec R, on disposerait d'autant, sinon plus, de méthodes statistiques) ; sa notoriété (est-ce vraiment important, nous avons montré qu'avec des logiciels tels que Tanagra - ou d'autres, R encore une fois, OpenStat, PSPP, etc. - nous obtenons les mêmes résultats) ; son aptitude à traiter les grandes bases (c'est son véritable atout, mais dans ce cas il ne paraît pas très judicieux de manipuler ses données dans Excel). Bref, l'add-in apparaît surtout comme une fonctionnalité bonus pour ceux qui ont déjà investi dans le logiciel. L'acquisition de SAS spécifiquement pour cet outil paraît moins pertinente en revanche.