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 miners¹. 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 (<u>http://www.sas.com/</u>). Il est présent sur le marché des logiciels de statistique depuis un grand nombre d'années maintenant². 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 Microsoft³. Je l'ai découvert très récemment sur les machines des salles informatiques de notre département (Département Informatique et Statistique – Université Lyon 2 – <u>http://dis.univ-lyon2.fr/</u>). 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. « <u>Data</u> <u>Mining/Analytic Tools Used</u> », 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 <u>http://support.sas.com/documentation/onlinedoc/addin/index.html</u>. 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 « <u>scoring dataset.xls</u> »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.

| 0. | scoring dataset.xls [Mode de compatibilité] - Microsoft Excel | | | | | | | | | | | | | | |
|------|---|-------|------------------------|------------------|------------------------|------------|-----------------------|---------------------|---------|-------------------|------------------|------------|-------------|--------|----------|
| | Ac | ┛ | Insertion | Mise | en page | For | mules | Données | Révis | ion / | Affichage D | éveloppeur | Compléments | SAS 🔘 | - = x |
| Donr | nées (S | Tâche | s Rapports | Favoris SAS + | S Actualiser | N M | lodifier ropriétés | Gérer le contenu | Outils | 2 Aide | | | | | |
| | | | Analy <u>s</u> e des o | données o | le survie 🕨 | ection | n | | Outils | | | | | | |
| | В | | Analyse <u>m</u> ulti | ivariée | • | | | | | | | | | | × |
| | A | | <u>A</u> NOVA | | | <i>й</i> ч | Analyse | de variance | à une d | imensi <u>o</u> n | | | - I | J | K |
| 1 0 | bject | | Ca <u>p</u> abilité | | • | 盃 | <u>A</u> nalyse | de variance | non-pa | ramétriqu | ie à une dimensi | ion | p06rcy | p07rcy | p08rcy 🖵 |
| 2 p | ositi | | <u>C</u> artes de cor | ntrôle | • | * | Modèle | s <u>l</u> inéaires | | | | | 0.99 | 0.99 | |
| 3 p | ositiv | | Data M <u>i</u> ning | | • | * | Modèles | s mixtes | | | | (| 0 0 | 0 | |
| 4 p | ositiv | | Description | | • | H | Test t | | | | | (| 0 0 | 0 | |
| 5 p | ositiv | | <u>D</u> onnées | | • | F | | | • | 202 | v | | 0 0 | 0.99 | |
| 6 p | ositiv | | Graphique | | • | - | 0 | | 0 | 4574 | 0.03 | 0.43 | L 0 | 0 | |
| 7 n | negati | | Pareto | | • | Ŀ | 0.99 | | 0 | 4093 | 0 | (| 0 0 | 0 | |
| 8 n | negati | | Régression | | | - | 0 | | 0 | 123 | 0 | (| 0 0 | 0 | |
| 9 p | osití | | Cária chronol | logique | | ⊢ | 1 | | 0 | 2422 | 0 | (| 0 0 | 1 | |
| 10 n | iegati | | Serie chronol | iogique | | - | 1 | | 0 | 357 | 0 | (| 0 | 0 | |
| 11 n | egati | | Modeles de t | aches | - | J | 1 | | - | 543 | 0 | (| | 0 | |
| H 4 | ► H | data | set 🖉 | | | | | | | | I ◀ <u>Ⅲ</u> | | | | |
| Prêt | | | | | | _ | | | | | | | 100 % (- |) | .:: |

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.

<u>Remarque</u>: 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

| Sélectionner des données | - | - | - | - | X |
|-------------------------------------|----------------|----------------|------------------|---------------------|------------|
| Données d'entrée | | | | | |
| Onnées Excel : | | | | | |
| A1:GS2159 | | | | | |
| Données SAS externes : | | | | | |
| | | | | • Parcou | rir |
| ✓ Détails | | | | Filtrer & 1 | Trier |
| Emplacement des résultats | | | | | |
| Nouvelle feuille de calcul : | Analyse de v | ariance non-pa | aramét | | |
| Feuille de calcul existante : | | | | | |
| Nouveau classeur | | | | | |
| Pourquoi est-il impossible d'indiqu | er un emplacer | ment dans Exc | el en cliquant c | lans une feuille de | e calcul ? |
| | | | ОК | Annuler | Aide |

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.

| Données Analyse p-values exactes Résultats | Données | comparis |
|---|--|----------|
| Titres | on\logistic regression\sas add-in\scoring dataset.xls!dataset | |
| Propriétés | Filtre de tache : Neant | |
| | Variables à attribuer : Fonctions de la tâche : | |
| | Nom Nom | |
| | ▲ objective Image: Constraint of the second se | |
| | | • |
| | Attribue la variable sélectionnée à la fonction que vous avez choisie dans la liste déroulante. | Â. |
| | | |

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

| se exactes Analyse | |
|---|--|
| tats Scores de test Wilcoxon Médiane Savage Van der Waerden Ansari-Bradley Klotz Mood Siegel-Tukey Données brutes | Calculer la loi empirique de la fonction de répartition (EDF lnclure les valeurs manquantes comme niveau de classe Supprimer la correction de continuité REMARQUE : les scores de test doivent être sélectionné afin d'activer les p-values exactes de la page "P-values exactes" et les statistiques de la page "Résultats". |
| Attribue la variable sélectionnée | à la fonction que vous avez choisie dans la liste déroulante. |

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.

| 🖄 Analyse de variar | 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 | âche | | 8 |
| Insérer du code | | | |
| | | | * |
| PROC NPAR | WAY DATA=WORK.TMP0TempTableInput WILCOXON MEDIAN SAVAGE VW AB KLOTZ | MOOD ST EDF | |
| ; VAR to | talspend: | | |
| CLASS | objective; | | |
| /* | | | н |
| Fin du | code de la tâche. */ | | |
| RUN; QUIT; | · · · · · · · · · · · · · · · · · · · | | - |
| • | III | | • |
| | Attribue la variable sélectionnée à la fonction que vou avez choisie dans la liste démulante | | |
| | | | |
| Aperçu du code | Exécuter 🔻 Annuler Aide | | |
| | | | |

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 RATTLE⁵). 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

<u>Remarque</u>: 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, « <u>Comparaison de populations – Tests non paramétriques</u> », 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.

| Scores de Wilcoxon (Sommes du rang) pour la variable totalspend | | | | | | | | |
|---|------------------|---------------|----------------|-------------------|-----------|----------------|--------------|--|
| | Ĩ | Classés pa | r variable ol | bjective | | | | |
| | | So | omme des | | Attendue | Ecart-type | Score | |
| objective | | N | scores | | sous H0 | sous H0 | moyen | |
| positive | 10 | 79 | 1308241 | | 1164780.5 | 14472.9379 | 1212.4569 | |
| negative | 107 | 79 | 1021320 | | 1164780.5 | 14472.9379 | 946.5431 | |
| | Les sco | ores moyens (| ont été utilis | és pour les lie | ens. | | | |
| Test à deux échantill | lons de Wilcoxon | ٦ | | | | | | |
| Statistique | 130824 | 41 | | | | | | |
| | | - | | | | _ | | |
| Approximation normale | | | | | | | | |
| Z | 9.912 | 23 | | | | | | |
| Unilatéral Pr > Z | <.000 | 01 | | 242 | | | | |
| Bilatéral Pr > Z | <.000 | 01 | | 070 | | \ | | |
| | | | | | | | | |
| Approximation t | | | | | | | | |
| Unilatéral Pr > Z | <.000 | 01 | | | | | | |
| Bilatéral Pr > Z | <.000 | 01 | | | | | | |
| Z inclut une correction | on de continuité | | | | | | | |
| de 0. | 5. | | | | | | | |
| | | | | | | | | |
| | | | Results | | | | | |
| | Value | Examples | Average | Rank sum | Rank mean | Mann-Whitney l | J 438660.00 | |
| | positive | 1079 | 1763.1909 | 1308241.0 | 1212.4569 | E(U) | 582120.50 | |
| totalspend objective | negative | 1079 | 992.8267 | 1021320.0 | 946.5431 | V(U) | 209465931.84 | |
| TANAGRA | All | 2158 | 1378.0088 | 2329561.0 | 1079.5000 | IZI | 9.91 | |
| | | | | $\mathbf{\nabla}$ | | P(> Z) | 0.00 | |

Analyse de variance non-paramétrique à une dimension

⁶ Voir aussi <u>http://fr.wikipedia.org/wiki/Test_%28statistique%29</u> 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.



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...

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

| 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 | | | | | | |

... et TANAGRA proposent les deux résultats.



| Attribute_Y | Attribute_X | | | Statistical test | | | | |
|-------------|-------------|----------|---------------|------------------|------------|----------|---------|-----------|
| | | Value | Examples | Average | Scores | Scores | Two-San | nple Test |
| | | | | | sum | mean | S | 440.66667 |
| | objective | 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 | | | | | | | p-value | 0.00000 |
| | | | | | | | One-way | Analysis |
| | | IA | NAGR | | Chi-Square | 72.47882 | | |
| | | « [\ | Nedian | | d.f. | 1 | | |
| | | | | | | | p-value | 0.00000 |

3.1.4 Test de Van der Waerden

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

| SAS | 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 | | |
| | | | | | | | |

| | | | | | | | Results | | | |
|---------------------------------|---------|-------------|-------------|----------------------------|----------|-----------|------------|-----------------|---------|------------|
| Test à deux échantillons de Van | | Attribute_Y | Attribute_X | Description Statistical te | | | | | | cal test |
| der waerden | | | | Volue Scores Scores | | | | Two-Sample Test | | |
| 7 | 0.6706 | | | value | Examples | Average | sum | mean | s | -223.95545 |
| 2 | 9.6706 | | | positive | 1079 | 1763.1909 | 223.9555 | 0.2076 | E(S) | 0.00000 |
| Unilatéral Pr > Z | <.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 | | | | | | 141 | 9.07001 |
| Analyse à une dir | nension | | | | | | | | p-value | 0.00000 |
| de Van der Wa | erden | | | | | | | | One-way | Analysis |
| Khi-2 | 93.5207 | | | TANAGRA | | | Chi-Square | 93.52068 | | |
| DLL | 1 | | | d.f | | | | d.f. | 1 | |
| Pr > Khi-2 | <.0001 | | | | | | | | p-value | 0.00000 |

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.

| Scores selon la formule de Savage (Exponentiel) pour la variable totalspend Classés par variable objective | | | | | | |
|---|-----------|----------------------|-----------------------|----------|-----------|--|
| Somme des Attendue Ecart-type Scor | | | | | | |
| objective | Ν | 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 score | s movens ont été uti | lisés pour les liens. | | | |

Analyse de variance non-paramétrique à une dimension

| 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 | | | | | | |
| Somme des Attendue Ecart-type Sco | | | | | | | |
| 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 | | |
| | | | | | | | |

| s scores moyens ont été utilisés pour les li |
|--|
|--|

| Test à deux échantillons de Siegel-Tukey | | | | | |
|--|--------|--|--|--|--|
| Statistique 1140335.60 | | | | | |
| Z | -1.689 | | | | |
| Unilatéral Pr < Z | 0.0456 | | | | |
| Bilatéral Pr > Z 0.09 | | | | | |
| 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

| Scores Ansari-Bradley pour la variable totalspend | | | | | | | |
|---|--------------------------------|-----------|----------------------|----------------------|------------|-----------|--|
| | Classés par variable objective | | | | | | |
| | SAS | | Somme des | Attendue | Ecart-type | Score | |
| 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 | |
| | | Les score | s movens ont été uti | lisés nour les liens | | | |

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

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

| | | | | | Results | | | | |
|----------------------|-------------|---------|-------------|-------------|------------|---------|-----------------|-------------|--|
| Attribute_Y | Attribute_X | | Description | | | | | stical test | |
| | | Value | Examples | Average | Scores sum | Scores | Two-Sample Test | | |
| totalspend objective | | | | | mean | S | 570436.6666 | 63 | |
| | positive | 1079 | 1763.1909 | 570436.6666 | 528.6716 | E(S) | 582659.9999 | 94 | |
| | negative | 1079 | 992.8267 | 594883.3333 | 551.3284 | V(S) | 52366128.4720 | .01 | |
| | All | 2158 | 1378.0088 | 1165320.0 | 540.0000 | Z | 1.689 | 14 | |
| | | | | | | p-value | 0.0911 | 19 | |
| | | | | | | | One-way | Analysis | |
| | | TANAGRA | | | | | Chi-Square | 2.85318 | |
| | | | | | | d.f. | 1 | | |
| | | | | | | | p-value | 0.09119 | |
| | | | | | | | | | |

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

3.1.8 Test de Klotz

Ce test est également présent dans les deux logiciels.

| | Scores Klotz Scores pour la variable totalspend | | | | | | | | |
|--------------------------------|---|-----------|----------------------|----------------------|------------|----------|--|--|--|
| Classés par variable objective | | | | | | | | | |
| | | | Somme des | Attendue | Ecart-type | Score | | | |
| 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 score | s movens ont été uti | lisás nour los lions | | | | | |

| es scores moyens d | ont été utilisés | pour les liens. |
|--------------------|------------------|-----------------|
|--------------------|------------------|-----------------|

| | | | | | | | Results | | | | |
|------------------------|--------------|-------------|---------------|----------------|----------|-------------|-----------|--------|------------------|------------|---------|
| lest a deux echantillo | ons de Klotz | Attribute_Y | Attribute_X | | | Description | ı | | Statisti | cal test | |
| Statistique | 1131.6905 | | | | | | Scores | Scores | Two-San | nnle Test | |
| Z | 1.8492 | | | Value | Examples | Average | sum | mean | 5 | 1012.90160 | |
| Unilatéral Pr > Z | 0.0322 | | | positive | 1079 | 1763.1909 | 1131.6905 | 1.0488 | E(S) | 1072.29604 | |
| Bilatéral Pr > Z | 0.0644 | | | negative | 1079 | 992.8267 | 1012.9016 | 0.9387 | V(5) | 1031.62442 | |
| | | | | All | 2158 | 1378.0088 | 2144.6 | 0.9938 | Z | 1.84920 | |
| | | totalspend | end objective | pend objective | | | | | | p-value | 0.06443 |
| Analyse à une din | nension | | | TANAGRA | | | | | One-way Analysis | | |
| de Klotz | | | | | | | | | Chi-Square | 3.41956 | |
| Khi-2 | 3.4196 | | | | | | | | d.f. | 1 | |
| DLL | 1 | | | | | | | | p-value | 0.06443 | |
| Pr > Khi-2 | 0.0644 | | | | | | | | | | |

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).

| | Sco | res Mood pour | la variable t | otalspend | 1 | | | | | | |
|---------------------|---------------|----------------|----------------|-------------|----------|-----------|--------------|----------------|------------|------------------|------------|
| | | Classés par v | ariable objec | tive | | | | | | | |
| S | AS | Som | ne des | | Attend | ue Eca | art-type | Score | | | |
| objective | | N | scores | | sous | но | sous H0 | moyen | | | |
| positive | 107 | 9 4317 | 67608 | | 4187385 | 90 8064 | 157.45 40 | 0155.337 | | | |
| negative | 107 | 9 4057 | 09571 | | 4187385 | 90 8064 | 157.45 37 | 6005.163 | | | |
| | Les sco | res moyens ont | été utilisés p | oour les li | ens. | | | | | | |
| | | | | | | | | | | | |
| Test à deux échanti | llons de Mood | | | | | | Results | | | | |
| Statistique | 431767608 | .4 Attribute | Y Attribute_X | (| | Descrip | tion | | 5 | Statistical test | |
| Z | 1.615 | 7 | | Value | Examples | Average | Scores sum | Scores | Τv | vo-Sample Te | st |
| Unilatéral Pr > 7 | 0.053 | 1 | | | | - | | mean | S | 405709 | 9571.57231 |
| | 0.055 | | | positive | 1079 | 1763.1909 | 431767608.48 | 63 400155.3369 | E(S) | 418738 | 3590.02930 |
| Bilateral Pr > [2] | 0.106 | 52 | | negative | 1079 | 992.8267 | 405709571.57 | 23 376005.1636 | V(5) | 65030635439 | 095.49220 |
| | | _ | | All | 2158 | 1378.0088 | 837477180 | .1 388080.2503 | Z | | 1.61567 |
| Analyse à une | dimension | totalspend | objective | | | | | | p-value | | 0.10617 |
| de Mo | bd | | | | | | | | One-wa | y Analysis | |
| Khi-2 | 2.610 | 14 | | | | | | | Chi-Square | 2.61039 | |
| DLL | | 1 | | | T | ANA | GRA | | d.f. | 1 | |
| Pr > Khi-2 | 0.106 | 2 | | | | | | | 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.**

| 6 | | 17 | · (° •) = | scoring d | ataset avec | : results. | xls [N | 1ode de con | npatibilit | té] - Mic | rosoft Excel | | | x |
|-----|---------------|-------|---------------------|--------------------|-------------|--------------------|---------------|-----------------------|------------|-----------|---------------|-----------|--------|-------|
| | Acc | ueil: | Insertion | Mise en page | Formules | Donn | ées I | Révision At | fichage | Dévelo | ppeur Complé | ments SAS |) - 🗖 | X |
| D | onnées SAS | Tâch | es Rapport | s Favoris SAS + | (tualiser | 🔊 Modil 🗄 Propr | fier iétés | Gérer le contenu | Outils | Aide | | | | |
| | | | Analy <u>s</u> e de | es données de | survie 🕨 | ection | | | Outils | | J | | | |
| | P | | Analyse <u>m</u> | ultivariée | • | | | | | | | | | × |
| | 4 | | <u>A</u> NOVA | | • | D | | E | | F | G | Н | - I | |
| 1 | object | | Ca <u>p</u> abilité | | • | 3rcy 👘 | р | 04rcy | totals | spend | p05spend | p05trans | p06rcy | |
| 2 | positiv | | <u>C</u> artes de | contrôle | • | | 1 | | 0 | 4012 | 0 | 0 | | 0 |
| 3 | positi | | Data M <u>i</u> nii | ng | • | | 0 | | 0 | 13 | 0 | 0 | | |
| 4 | positi | | De <u>s</u> criptio | n | • | 0. | .95 | | 0 | 2628 | 0 | 0 | | |
| 5 | positi | | Données | | • | | 0 | | 0 | 962 | 0 | 0 | | |
| 6 | positi | | - Graphique | | • | | 0 | | 0 | 4574 | 0.03 | 0.41 | | - 11 |
| 7 | negati | | Pareto | | | 0. | .99 | | 0 | 4093 | 0 | 0 | | - 11 |
| 8 | negati | | Págrassio | | - | | 0 | P - 1 - 1 1 | 0 | 123 | 0 | 0 | | _ |
| 9 | positiv | | Regression | | | M. M | odele | lineaire ger | eralise | 122 | 0 | 0 | | _ |
| 10 | negati | | Serie chro | nologique | • | <u><u></u>Ré</u> | égress | ion linéaire. | | 357 | 0 | 0 | | |
| 11 | negati | | <u>M</u> odèles d | le tâches | • | ii Ré | égress | ion logi <u>s</u> tiq | ue | | 0 | 0 | | - |
| I | < → >I [| dat | aset 🖉 A | nalyse de vari | ance non- | 🖉 Ré | égr 🎧 | SAS Add- | in 4.3 for | r Microso | ft Office | | • | I I |
| Pré | êt 🔝 | | | | | _ | | Appuyez s | ur F1 po | ur obten | ir de l'aide. | | | Ð .:: |

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.

| onnées odèle | Données | | | |
|--|--|--|--|--|
| Réponse Effets Sélection Options aphes | Source de données : D:\D add- Filtre de tâche : Néar | DataMining\Databa in\scoring datase | ases_for_mining\dataset_for_soft_dev_and_c at avec results.xls!dataset | omparison\logistic regression\sas |
| édictions | Variables à attribuer : | | Fonctions de la tâche : | |
| ues opriétés | Nom / liabslunivnd liabsluni | | Variable dépendante (Linite : 1) Variables quantitatives Variables | I gender 3 ^e unités de cha Unités Ecart-type Saisissez un ou plusieurs nombres (unités de change) séparés par des espaces. Pa exemple: 0.2 0.3 0.5 0.7 |
| Aperçu du cod | le | | Exécuter 🔻 | Annuler Aide |

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

| odèle | Modèle > Réponse | |
|---|---|--|
| Effets Sélection | Type de réponse : | Binaire |
| Options Graphes Prédictions Fitres Propriétés | Type de modèle : | logit probit log-log complémentaire glogit |
| | Niveaux de réponse pour objective : | negative positive |
| | Ajuster le modèle au nivear : Indiquer le type de réponse. votre variable de réponse. Si | Les types de réponses disponibles dépendent du nombre de niveaux dans cette demière ne contient que deux niveaux, le type de réponse est alors |
| | binaire. Si elle en contient plu | us, vous pouvez choisir une réponse de type classée ou non classée. |

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).

| Régression logist | ue pour D:\DataMining\Databases_for_mining\dataset_for_soft_dev_and_comparison\logist |
|--|---|
| Sélection Options Graphes Prédictions Titres Propriétés | |
| Aperçu du code | Exécuter 🔻 Annuler Aide |
| | |

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 $\alpha = 1\%$.

| lodèle | Modèle > Sélection | |
|-----------|--|---|
| Fffets | 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 |
| raphes | Niveaux de significativité | "sélectionnés" et seront transférés dans cette liste. |
| itres | Pour entrer dans le modèle : 0.01 | éléments "sélectionnés" en les |
| ropriétés | Pour rester dans le modèle : 0.05 | sélectionnant et en utilisant les boutons fléchés vers le haut et vers le bas. |
| | | p01rcy |
| | | p02rcy |
| | | p04rcy |
| | | totalspend |
| | | p05spend |
| | | p05trans 👻 |
| | Indiquez le niveau de significativité à utiliser pou | ir saisir une variable quantitative (explicative) dans le modèle. |

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.

| Donnees Modèle | Modèle > Options | |
|--|---|---|
| Réponse Effets Sélection Options Graphes | Détails sur les estimations Matrice de corrélation des paramètres estimés Matrice de covariance des estimations | Table de classification Image: Afficher la 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. Par exemple : 0.2 0.3 0.5 0.7 |
| | Intervalle de confiance Paramètres Mald Vraisemblance du profil | Rapports de cotes conditionnels |
| | Niveau de confiance : 952 Calcule les intervalles de confiance pour le rapport de co | ▼ tes. |

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 | | Fréquence | | |
| ordonnée | objective | 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.

11 brlanglic

12 p12rcy

| | SAS | | | | | | |
|-------|----------------|---------------|----------------|-------------------|------------|--|--|
| | Récapit | ulatif sur la | sélection e | n 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

12

8.692

7.420

0.0032

0.0064

TANAGRA

| N | Current Reg. | Moved | Sol.1 |
|----|------------------|--------------------------------------|-----------------|
| | AIC: 2993.62 | gender3 | gender3 |
| 1 | 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 | | |
| | AIC: 2576.00 | productcount | productcount |
| 2 | 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 |
| 3 | CHI-2:574.63 | Chi-2:54.575 | Chi-2 : 54.575 |
| 5 | d.f. : 2 | p:0.0000 | p:0.0000 |
| | p-value : 0.0000 | | |
| | AIC: 2361.99 | tf37 | tf37 |
| | CHI-2:637.63 | Chi-2:48.638 | Chi-2:48.638 |
| 4 | d.f.: 3 | p:0.0000 | p:0.0000 |
| | p-value : 0.0000 | | |
| | AIC: 2313.22 | p05trans | p05trans |
| 5 | CHI-2:688.40 | Chi-2:18.716 | Chi-2:18.716 |
| 5 | d.f. : 4 | p:0.0000 | p:0.0000 |
| | p-value : 0.0000 | | |
| | AIC: 2293.07 | ahh6ppers | ahh6ppers |
| 6 | 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 | | |
| | AIC: 2280.93 | tf68 | tf68 |
| 7 | 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 | | |
| | AIC: 2268.53 | amtfrench | amtfrench |
| 8 | 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 | | |
| | AIC: 2260.39 | p09tenure | p09tenure |
| 9 | 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 | | |
| | AIC: 2250.76 | | tf128 |
| 10 | CHI-2 : 760.86 | Chi-Z : 9.480 | Ch1-Z : 9.480 |
| | d.t. : 9 | p:0.0021 | p:0.0021 |
| | p-value : 0.0000 | halaa alka | halaa alka |
| | AIC: 2243.02 | | |
| 11 | CHI-2:770.00 | CIII-2 : 8.693 | 0.0022 |
| | | p : 0.0032 | p : 0.0032 |
| - | AIC · 2236 /0 | n12rcy | n12rcy |
| | CHI-2 · 779 13 | $\frac{P_1 L Cy}{Chi_2 \cdot 7 421}$ | Chi-2 · 7 421 |
| 12 | d f · 11 | -10,0064 | n : 0.0064 |
| | n-value : 0.0000 | P . 0.0004 | P . 0.0001 |
| | AIC : 2230 97 | | p02rcv |
| | 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 |

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

| | Tanagra |
|-----------------------|---------|
| Attributes in the equ | uation |

| 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 |
| oknfren | -8.0473 | 1.4203 | 32.1021 | 0.0000 |
| orlanglic | 2.2944 | 0.7998 | 8.2292 | 0.0041 |
| gender3 | -1.9310 | 0.1188 | 264.3180 | 0.0000 |
| o05trans | -4.5013 | 1.2440 | 13.0927 | 0.0003 |
| o09tenure | 26.8725 | 14.3488 | 3.5074 | 0.0611 |
| o12rcy | 0.5115 | 0.1886 | 7.3549 | 0.0067 |
| productcount | 0.1970 | 0.0202 | 95.1812 | 0.0000 |
| f 128 | 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 | | | | | | | |
| | 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 | | | |

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

| SAS |
|-----|
|-----|

TANAGRA

| 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 | |

| Hosmer Lemeshow Goodness-of-Fit Test | | | | | | |
|--------------------------------------|-------|-------------------|---------|----------|----------|-------|
| | | Positive | | 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 | | |

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

| TANAGRA 1.4.43 - [Hosm | er Lemeshow Test 1] | _ | | | | | 8 |
|----------------------------|---------------------------|--------------|--------------------------|--------------------|--------------------|--------------|----|
| Tile Diagram Compo | nent Window Help | | | | | - 5 | '× |
| 🗅 📽 🔚 | | | | | | | |
| | 10 | 1.000 | 199.0 | 00 196.856 | | | |
| ⊡-∰ Dataset (scoring da | Hosmer Lemeshow Statistic | | | | | | |
| 🖻 🙀 Define status 2 | | | | Chi-Square | d.f. | Significance | |
| Forward-logit 1 | | | Goodness-of- test | fit 6.4875 | 8 | 0.5928 | |
| I Hosmer Lemeshow Test 1 | | | • | III | | 4 | Ŧ |
| | (| Components | | | | | |
| Data visualization | Statistics | Nonparame | Nonparametric statistics | | Instance selection | | |
| Feature construction | Feature selection | Regression | | Factorial analysis | | ysis | |
| PLS | Clustering | Spv learning | | Meta-spv learning | | ning | |
| Spv learning assessment | Scoring | Asso | ciation | | | | |
| Binary logistic regression | | | 🎇 C-PLS | | | 🕂 C-RT | |
| | | | | | | .:: | |

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.