Objectif

Régression logistique multinomiale.

La régression logistique est très répandue pour les problèmes de prédiction ou d'explication d'une variable dépendante binaire (malade oui/non, défaillance oui/non, client potentiel oui/non, etc.) à partir d'une série de variables explicatives continues, binaires ou binarisées (dummy variables). On parle dans de cas de régression logistique binaire.

Lorsque la variable dépendante possède plusieurs catégories non ordonnées (K > 2), on parle de régression logistique multinomiale. Elle est peu (ou moins) connue, pourtant cette configuration est finalement assez courante. De plus, elle est directement traitée par les autres méthodes d'apprentissage telles que l'analyse discriminante prédictive, les arbres de décision, etc.

Grosso modo, la régression logistique multinomiale[1] consiste à désigner une catégorie de référence, la dernière ($K^{ème}$) par exemple pour fixer les idées, et à exprimer chaque logit (ou log-odds) des (K-1) modalités par rapport à cette référence à l'aide d'une combinaison linéaire des variables prédictives[2].

Dans ce didacticiel, nous montrons la mise en œuvre de la régression logistique multinomiale dans TANAGRA.

Données

Nous voulons expliquer, pour une série de produits de même catégorie, la marque (3 marques possibles) choisie par 735 consommateurs à partir de leur age et de leur sexe. Les données sont accessibles sur notre serveur[3]. Ces données ont déjà été traitées à l'aide du logiciel R. La description des données, et des résultats associés, est disponible en ligne (http://www.ats.ucla.edu/STAT/R/dae/mlogit.htm). Le lecteur pourra comparer les sorties des logiciels.

Régression logistique multinomiale avec TANAGRA

Importer les données et créer un diagramme

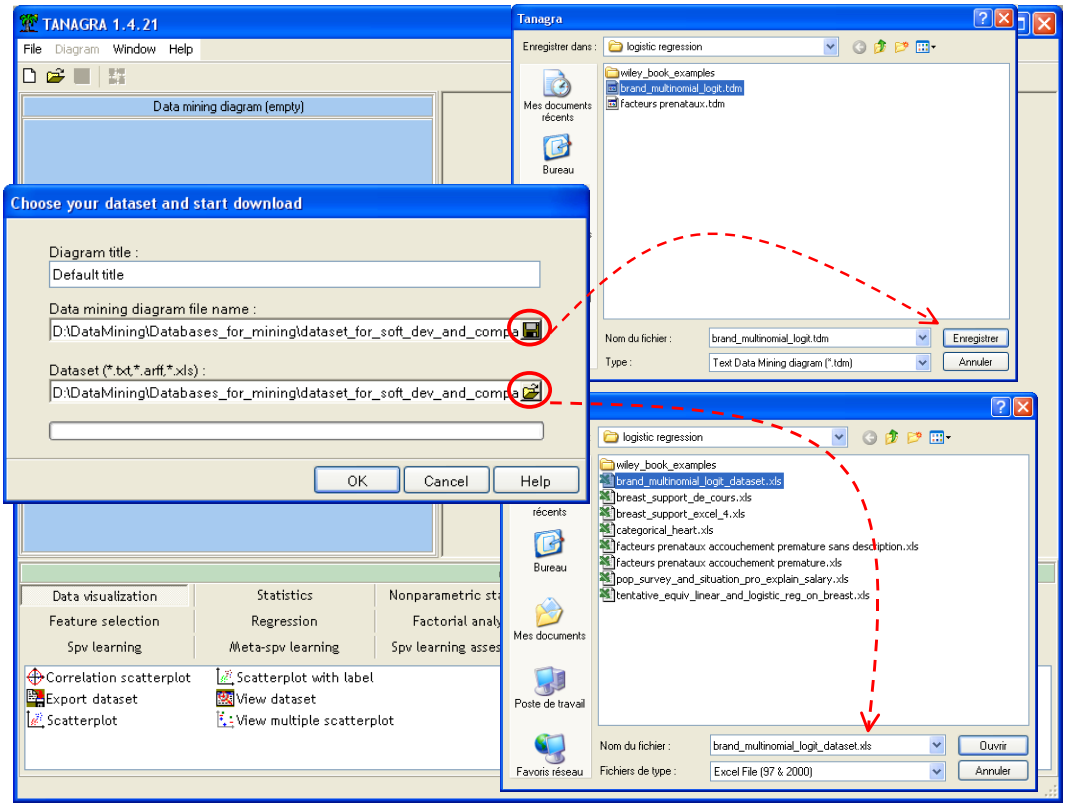
Après avoir lancé TANAGRA, nous créons un nouveau diagramme en activant le menu FILE/NEW. Une boîte de dialogue nous invite à choisir le fichier de données, BRAND_MULTINOMIAL_DATASET.XLS, et à définir le nom du fichier diagramme.

Pour les fichiers XLS, l'importation fonctionnera correctement si le classeur n'est pas en cours d'édition par ailleurs, et que les données sont situées dans la première feuille.

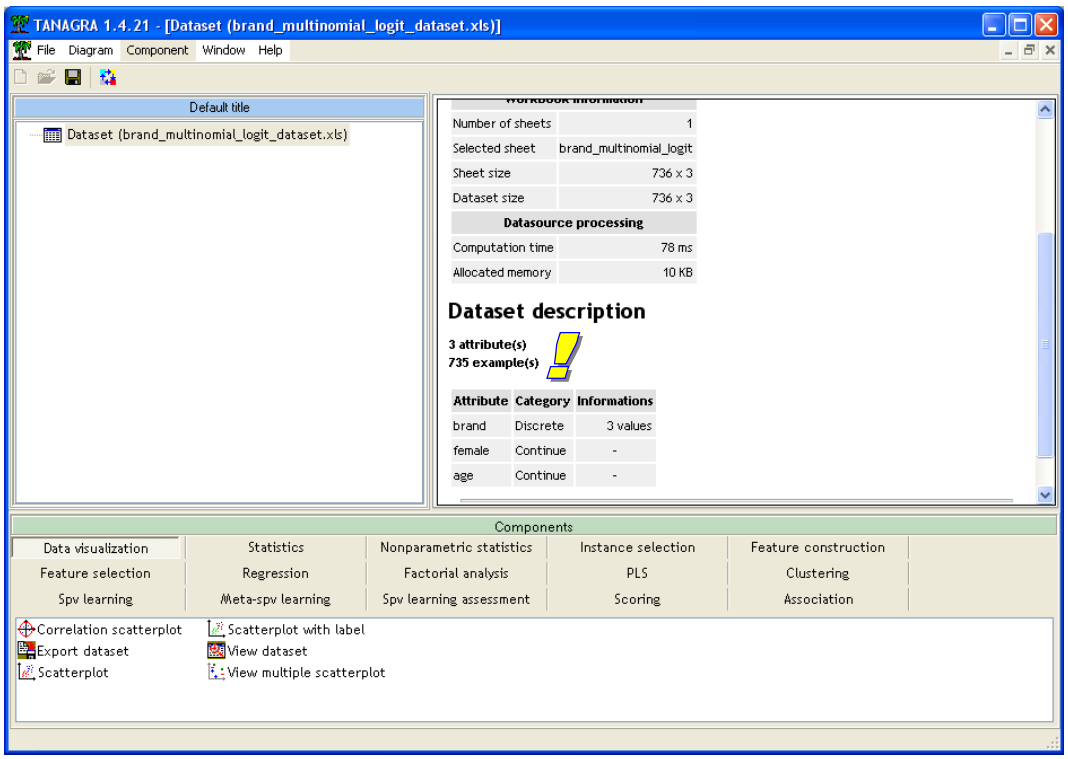
[1] On parle aussi de Régression logistique polytomique à variable dépendante nominale.

[2] Voir http://www.stat.psu.edu/~jglenn/stat504/08_multilog/01_multilog_intro.htm pour plus de précisions.

[3] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/brand_multinomial_dataset.xls



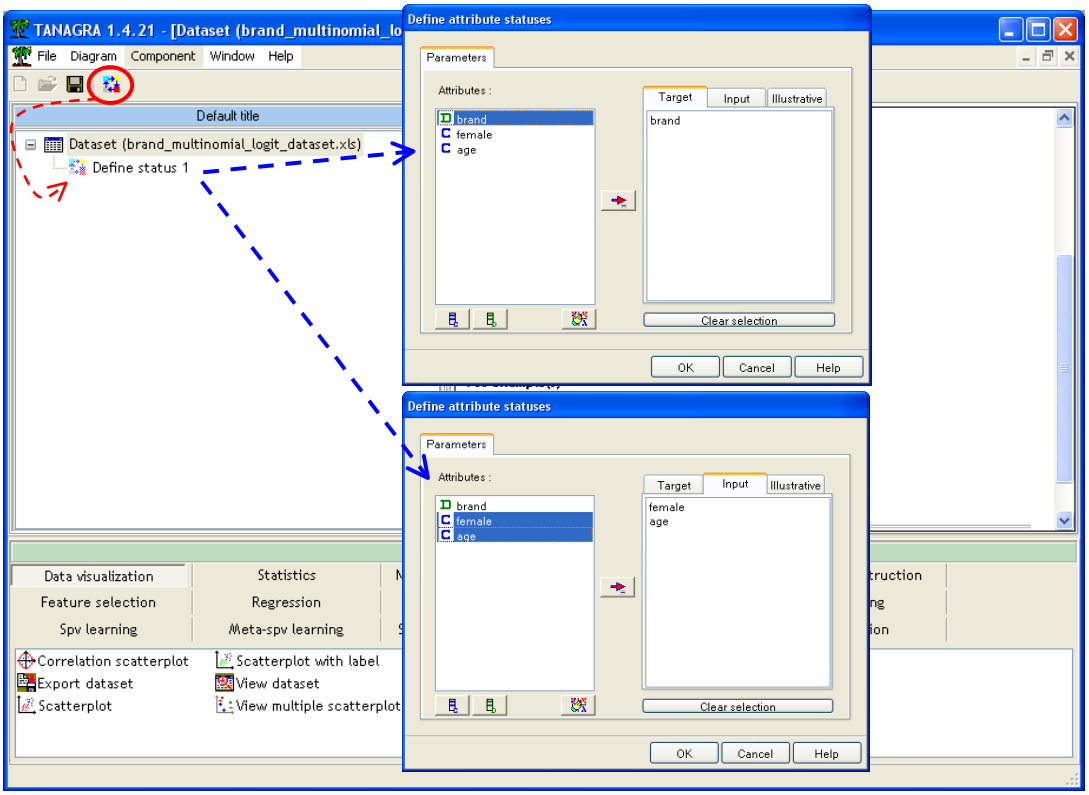
Les données sont chargées. Nous vérifions que 3 variables et 735 observations ont bien été importées.



Définir le problème

Etape suivante, nous devons définir le rôle de chaque variable. Pour cela, nous insérons le composant DEFINE STATUS dans le diagramme en utilisant le raccourci de la barre d'outils.

La variable BRAND est la variable cible (TARGET), les variables FEMALE et AGE sont les variables prédictives (INPUT).



Régression logistique multinomiale

Nous insérons maintenant la régression logistique multinomiale dans le diagramme. Le composant MULTINOMIAL LOGISTIC REGRESSION est accessible dans l'onglet SPV LEARNING.

TANAGRA utilise par défaut la dernière modalité de la variable dépendante comme modalité de référence. Si ce choix ne convient pas, une stratégie simple consiste à trier les données selon la variable dépendante de manière à ce que les observations correspondant à la modalité de référence soient en dernière position dans le fichier. Dans notre exemple, c'est la modalité « _3 » qui joue le rôle de modalité de référence.

Nous activons le menu VIEW pour accéder aux résultats. La fenêtre comporte plusieurs parties.

Matrice de confusion

Elle confronte la vraie valeur de la variable à prédire avec les prédictions du modèle.

Classifier performances

Error rate	0.4476

Values prediction

Value	Recall	1-Precision
_1	0.2802	0.3256
_2	0.7752	0.4989
_3	0.4977	0.3678

Confusion matrix

	_1	_2	_3	Sum
_1	58	136	13	207
_2	18	238	51	307
_3	10	101	110	221
Sum	86	475	174	735

Le taux d'erreur est de 44.76%. On peut en déduire le taux de bon classement qui est égal à (1 – 0.4476) = 55.27%. Dans certaines références, on l'assimile à un **pseudo-R²**, on parle de « Count R-square » (cf. http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm).

On ne sait pas si cette valeur est intéressante ou non dans l'absolu. Seul un expert du domaine peut nous dire si ce résultat est satisfaisant. En revanche, nous avons la possibilité de confronter notre modèle avec le modèle trivial composé de la seule constante. Dans ce cas, la règle d'affectation serait d'attribuer à tous les individus la modalité de Y la plus fréquente dans le fichier de données. Dans notre cas, la modalité de BRAND la plus fréquente est « _ 2 », avec un effectif de 307. On peut proposer un indicateur ajusté, « Adjusted Count R-square », il s'écrit :

$$R^2_{AC} = \frac{\#correct - \max_k(n_k)}{n - \max_k(n_k)} = \frac{(58+238+110)-307}{735-307} = \frac{99}{428} = 0.231$$

Si le modèle évalué ne fait pas mieux que le modèle trivial, nous avons un pseudo-R² égal à 0 ; si nous avons un modèle parfait, classant correctement tous les individus, nous avons un R² = 1.

Dans notre exemple, nous pouvons considérer que l'age et le sexe apportent de l'information dans la connaissance du choix de la marque.

Qualité de l'ajustement

La seconde partie de la fenêtre, retraçant la qualité de l'ajustement au sens de la régression logistique va nous confirmer cette impression. Ici aussi, nous confrontons notre modèle avec le modèle trivial, mais en utilisant principalement le rapport de vraisemblance.

Adjustement quality

Predicted attribute	brand	
Ref. value	_3	
Number of examples	735	
Model Fit Statistics		
Criterion	Intercept	Model
AIC	1595.792	1417.941
SC	1604.991	1445.541
-2LL	1591.792	1405.941
Model Chi² test (LR)		
Chi-2	185.8502	
d.f.	4	
P(>Chi-2)	0.0000	
R²-like		
McFadden's R²	0.1168	
Cox and Snell's R²	0.2234	
Nagelkerke's R²	0.2524	

Nous avons un nouveau lot de pseudo-R², définis différemment[4]. Nous disposons du test du rapport de vraisemblance (CHI-2 test). Nous disposons enfin des critères AIC et SC (ou BIC) qui mettent en balance la qualité de l'ajustement (-2LL) et la complexité du modèle.

[4] http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

Le critère SC est le plus exigeant. Il nous indique que le modèle est effectivement pertinent (SC = 1445.541), meilleur en tous les cas que le modèle trivial (SC = 1604.991).

Coefficients des logit

Attributes in the equation

Class.Value	_1				_2			
Pred.Att.	Coef.	Std.Err	Wald	p-value	Coef.	Std.Err	Wald	p-value
constant	22.721397	-	-	-	10.946741	-	-	-
female	-0.465941	0.2261	4.247	0.0393	0.057873	0.1964	0.08681	0.7683
age	-0.685908	0.06263	120	0.0000	-0.317702	0.04401	52.12	0.0000

La modalité « _ 3 » est la référence. Nous avons 2 (c.-à-d. K – 1) équations, elles s'écrivent

- $\ln\left[\frac{P(Y=\_1/X)}{P(Y=\_3/X)}\right] = 22.721 - 0.466 \times female - 0.686 \times age$
- $\ln\left[\frac{P(Y=\_2/X)}{P(Y=\_3/X)}\right] = 10.947 + 0.058 \times female - 0.318 \times age$

Le test de Wald permet de vérifier la significativité de chaque coefficient dans chaque régression.

A titre de comparaison, voici les coefficients proposés par le package VGAM de R.

```
library(VGAM)
mlogit<- vglm(brand~female+age, family=multinomial(), na.action=na.pass)
summary(mlogit)

Call:
vglm(formula = brand ~ female + age, family = multinomial(),
    na.action = na.pass)

Pearson Residuals:
                      Min       1Q   Median       3Q    Max
log(mu[,1]/mu[,3]) -5.5632 -0.44331 -0.32370 0.55468 7.7720
log(mu[,2]/mu[,3]) -4.7219 -0.68004 -0.44685 0.97285 1.7861

Coefficients:
                  Value Std. Error   t value
(Intercept):1 22.721396   2.058016  11.04043
(Intercept):2 10.946741   1.493160   7.33126
female:1      -0.465941   0.226089  -2.06087
female:2       0.057873   0.196427   0.29463
age:1         -0.685908   0.062626 -10.95243
age:2         -0.317702   0.044007  -7.21939

Number of linear predictors:  2

Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Dispersion Parameter for multinomial family:   1

Residual Deviance: 1405.941 on 1464 degrees of freedom

Log-likelihood: -702.9707 on 1464 degrees of freedom

Number of Iterations: 5
```

Evaluation globale des coefficients

Les variables interviennent dans chaque équation. Nous avons la possibilité de tester globalement leur significativité en mettant en place le test d'hypothèses H0 : le coefficient est nul dans toutes les équations vs. H1 : sur une des équations au moins, il est différent de zéro.

Overall Effect

Attribute	d.f.	Chi-2 Wald	p-value
female	2	7.670	0.0216
age	2	123.388	0.0000

La procédure s'appuie toujours sur une statistique de Wald. Dans notre exemple, nous constatons que nos variables sont globalement significatives à 5%.

Conclusion

Ce didacticiel avait pour but de montrer, succinctement, la mise en œuvre de la régression logistique multinomiale dans TANAGRA.

Pour plus de détails sur la méthode et les calculs sous-jacents, nous conseillons l'excellente référence en ligne : http://www.stat.psu.edu/~jglenn/stat504/08_multilog/01_multilog_intro.htm