Subject

In this tutorial, we show how to implement a multinomial logistic regression with TANAGRA.

Logistic regression is a technique for making predictions when the dependent variable is a dichotomy, and the independent variables are continuous and/or discrete. The technique can be modified to handle dependent variable with several (K > 2) levels.

When the responses categories are unordered, we have the multinomial logistic regression. Roughly speaking, we compute the logit function for each (K-1) categories related to a reference group [http://www.stat.psu.edu/~jglenn/stat504/08_multilog/01_multilog_intro.htm].

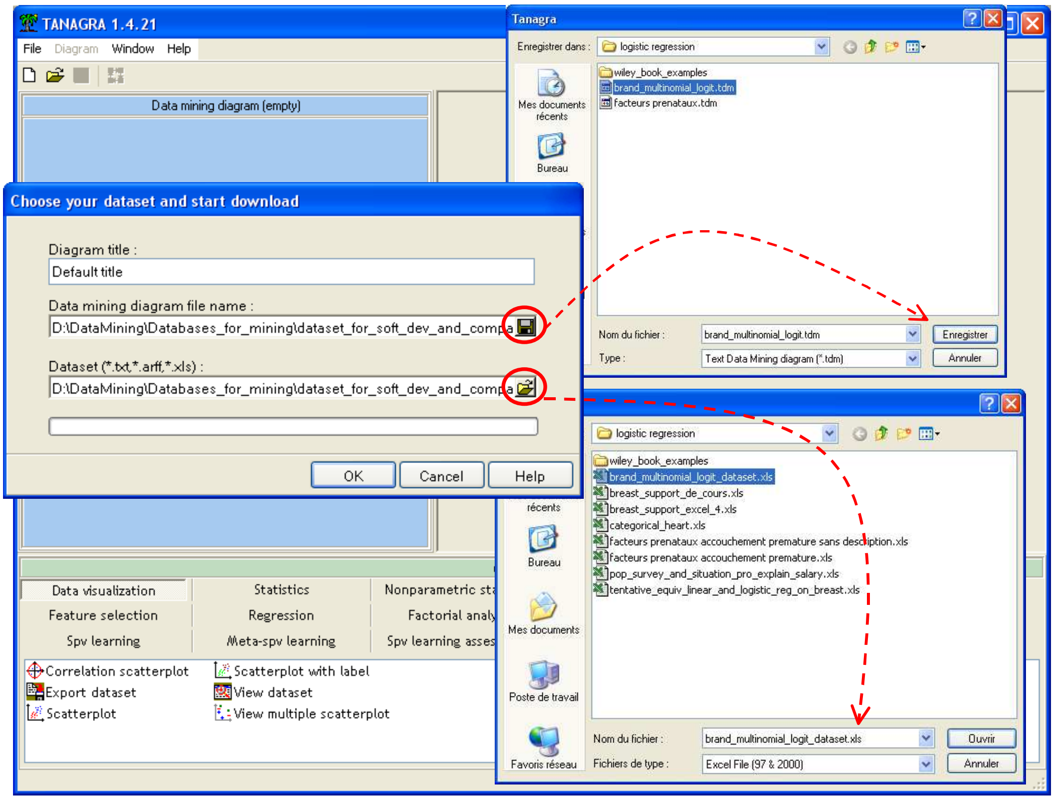
Dataset

We want to explain the brand, for some commodity, chosen by consumers starting from their age and their sex. The dataset is available on line[1]. We can see the results obtained with other software such as R on the same dataset [http://www.ats.ucla.edu/STAT/R/dae/mlogit.htm].

Multinomial logistic regression with TANAGRA

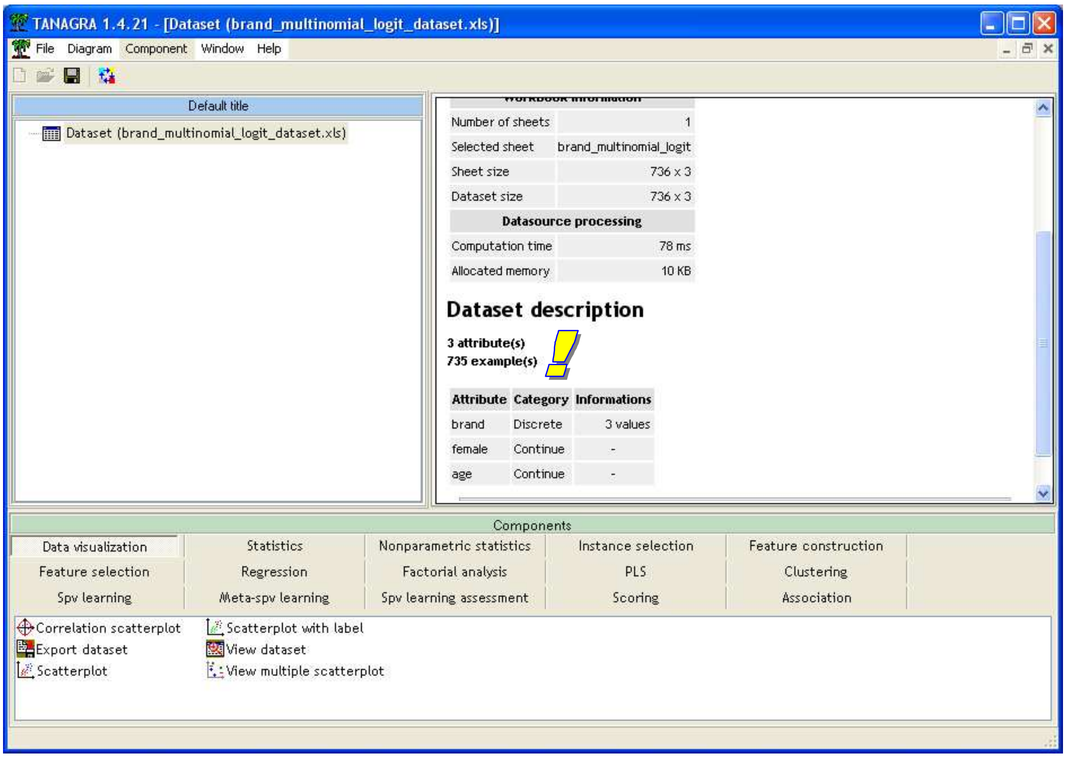
Accessing the data and creating a new diagram

After starting TANAGRA, we create a new diagram by activating the FILE/NEW menu. In the dialog box, we choose the data file BRAND_MULTINOMIAL_DATASET.XLS and then we specify the name of the diagram. For XLS files, the importation functions properly if the folder is not being edited further, and that the data are located in the first sheet.



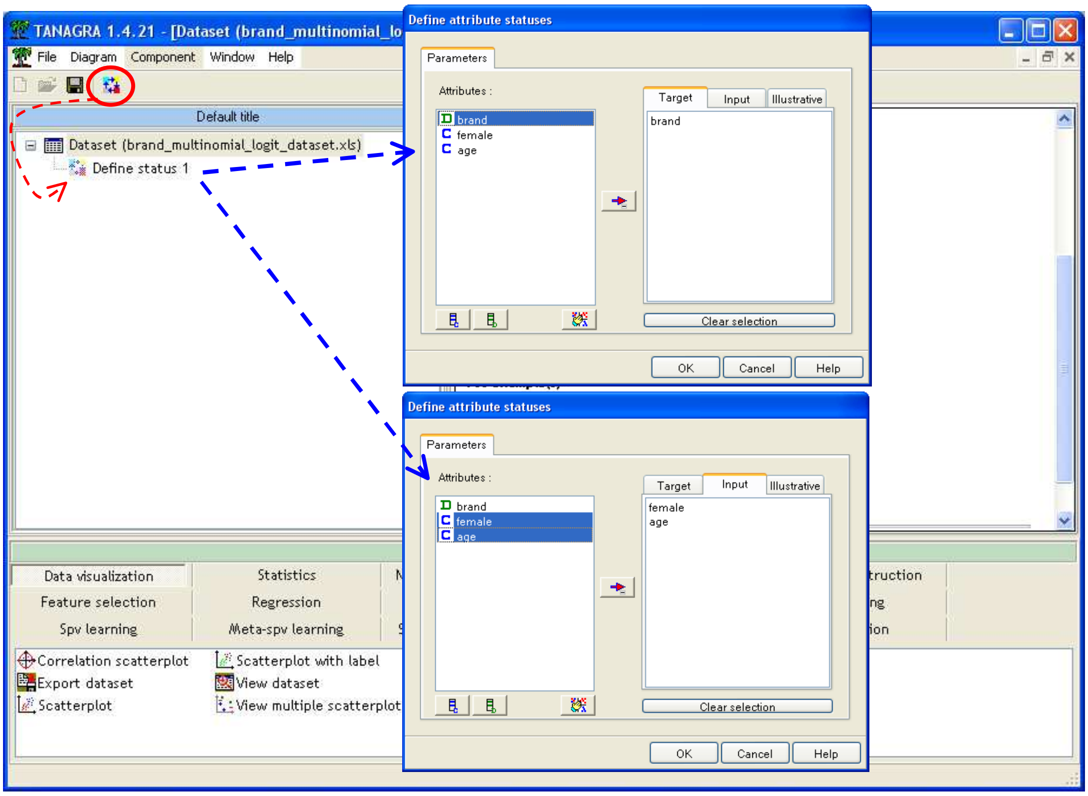
[1] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/brand_multinomial_dataset.xls

The data is loaded. We check that 3 variables and 735 records have been imported.



Defining the role of the variables

In the next step, we define the role of the variables. BRAND is the TARGET attribute; FEMALE and AGE are the INPUT ones.



Multinomial logistic regression

We add the MULTINOMIAL LOGISTIC REGRESSION component (SPV LEARNING tab) into the diagram.

By default, TANAGRA uses the last encountered value of the dependent variable as the reference group. If you want to modify the choice, the simplest way is to sort adequately the dataset.

We obtain the following results (VIEW menu).

Confusion matrix (classification matrix)

The confusion matrix compares the observed value and the predicted value of the dependent variable.

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

Some ratio can be computed e.g. error rate or accuracy rate. An interesting ratio is the adjusted count pseudo r-square which corrects the accuracy rate with the most frequent value of the dependent variable (cf. http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm).

For our example, we obtain the following "adjusted count r-square"

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

If our classifier is no more competitive as the default classifier (predict with the most frequent value of the dependent variable), we obtain 0; for a perfect prediction, we obtain 1.

Adjustment quality

The next section compares the initial model, predict with the constant only, and our model, using the likelihood ratio principle. Other pseudo R-square indicators are available. Other indicators such as AIC or SC (BIC) statistics make a trade-off between the deviance and the complexity (number of parameters) of the model. SC is the most rigorous indicator. It shows that our model seems really relevant (SC of the initial model = 1604.991; SC of the model = 1445.541).

The likelihood ratio test (LR) reaches to the same conclusion. The whole model is significant.

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	

Logit coefficients

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

The « _ 3 » value is the reference group. We have 2 (i.e. K – 1) equations:

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

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

The Wald test is used to test the significance of each coefficient, for each equation. The Wald statistic is the square of the ratio between the coefficient and its standard error. It follows a CHI-SQUARE distribution with 1 degree of freedom.

We show below the results obtained with the VGAM package for the R software.

```
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
```

Global evaluation of variables

In the previous step, we can evaluate the relevance of each variable into each equation. Now, we try to evaluate the global relevance of each variable i.e. the coefficient of the variable is it equal to 0 into all the equations?

Overall Effect

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

This test relies also on a Wald statistic. We see here that all variables are relevant for a 5% significance level.

Conclusion

In this tutorial, we show how to implement and read the results of the multinomial logistic regression with TANAGRA.

For more details about the method and the underlying computations, we recommend the following reference http://www.stat.psu.edu/~jglenn/stat504/08_multilog/01_multilog_intro.htm