1 Objectif

Sélection des variables explicatives catégorielles en régression logistique.

La régression logistique vise à construire un modèle permettant de prédire une variable cible binaire à partir d'un ensemble de variables explicatives (descripteurs, prédicteurs, variables indépendantes) numériques et/ou catégorielles. Elles sont traitées telles quelles lorsque ces dernières sont numériques. Elles doivent être recodées lorsqu'elles sont catégorielles. Le codage en indicatrices 0/1 (dummy coding) est certainement la méthode la plus utilisée.

Prenons un exemple simple pour préciser les idées. Mettons que la variable X prend 3 modalités {A, B, C}. Nous créerons deux indicatrices : X_A qui prend la valeur 1 lorsque X = A, 0 sinon ; et X_B qui est égale à 1 lorsque X = B. On identifie la configuration X = C lorsque X_A et X_B prennent simultanément la valeur 0. « C » est dite modalité de référence. Le choix de la modalité de référence joue sur l'interprétation des coefficients de la régression. En revanche il ne pèse en rien sur sa qualité prédictive c.-à-d. quelle que soit la modalité de référence choisie pour chaque variable explicative catégorielle intervenant dans le modèle, le taux d'erreur en généralisation, si l'on s'en tient à se critère, ne sera pas affecté.

La situation se complique lorsque l'on procède à une sélection de variables. L'idée est de déterminer les prédicteurs qui contribuent significativement à l'explication de la variable cible. Il n'y a aucun problème quand nous considérons une variable numérique, elle est soit exclue soit conservée dans le modèle. Mais comment procéder lorsqu'on manipule une explicative catégorielle ? Devons-nous traiter les indicatrices associées à une variable comme un bloc indissociable ? Ou bien pouvons-nous les dissocier, en ne conservant que certaines d'entres elles ? Est-ce que cette stratégie est légitime ? Comment lire les coefficients dans ce cas.

Dans ce tutoriel, nous étudions les solutions proposées par les logiciels **R 3.1.2**, **SAS 9.3**, **Tanagra 1.4.50** et **SPAD 8.0**. Nous verrons que les algorithmes de sélection de variables s'appuient sur des critères spécifiques selon les logiciels. Nous constaterons surtout qu'ils proposent des approches différentes lorsque nous sommes en présence des explicatives catégorielles. Cela n'est pas sans conséquence sur la qualité prédictive des modèles.

2 Données

Nous utilisons les données « heart-disease » du serveur UCI[1]. Nous avons transformé la variable cible NUM en DISEASE avec deux valeurs possibles {absence, présence}. Les prédictives (AGE…THAL) sont constituées d'un mix de variables numériques et catégorielles. Voici leur description :

Variable	Type
age	numeric
sex	{male, female}
cp	{asympt, atyp_angina, non_anginal, typ_angina}
trestbps	numeric
chol	numeric
fbs	{f, t}
restecg	{left_vent_hyper, normal, st_t_wave_abnormality}
thalach	numeric
exang	{no, yes}
oldpeak	numeric
slope	{down, flat, up}
ca	numeric
thal	{fixed_defect, normal, reversable_defect}

Figure 1 - Liste des variables prédictives candidates

Les prédicteurs catégoriels sont mis en évidence, leurs modalités énumérées.

3 Régression logistique et sélection de variables sous R

Nous importons le fichier « heart-c.xlsx » en utilisant le package « xlsx ».

```
#vider la mémoire
rm(list=ls())

#modifier le répertoire par défaut
setwd("…")


#charger le fichier au format xlsx
library(xlsx)
heart <- read.xlsx(file="heart-c.xlsx",sheetIndex=1,header=T)
print(summary(heart))
```

Nous obtenons la description suivante, « disease » est la variable cible :

[1] https://archive.ics.uci.edu/ml/datasets/Heart+Disease

```
> print(summary(heart))
      age            sex                cp          trestbps          chol          fbs
 Min.   :29.00   female: 96   asympt      :143   Min.   : 94.0   Min.   :130.0   f:258
 1st Qu.:47.50   male  :207   atyp_angina: 50   1st Qu.:120.0   1st Qu.:210.0   t: 45
 Median :55.00                non_anginal: 87   Median :130.0   Median :240.0
 Mean   :54.37                typ_angina : 23   Mean   :132.1   Mean   :246.7
 3rd Qu.:61.00                                  3rd Qu.:140.0   3rd Qu.:275.0
 Max.   :77.00                                  Max.   :200.0   Max.   :560.0
                 restecg        thalach       exang       oldpeak        slope           ca
 left_vent_hyper     :147   Min.   : 71.0   no :204   Min.   :0.00   down: 21   Min.   :0.0000
 normal              :152   1st Qu.:130.0   yes: 99   1st Qu.:0.00   flat:140   1st Qu.:0.0000
 st_t_wave_abnormality:  4   Median :150.0             Median :0.80   up  :142   Median :0.0000
                             Mean   :149.8             Mean   :1.04              Mean   :0.6744
                             3rd Qu.:170.0             3rd Qu.:1.60              3rd Qu.:1.0000
                             Max.   :200.0             Max.   :6.20              Max.   :3.0000
                thal        disease
 fixed_defect     : 18   absence :165
 normal           :168   presence:138
 reversable_defect:117
```

3.1 Régression sur la totalité des variables explicatives candidates

Nous réalisons l'analyse sur la totalité des variables disponibles. Nous constatons que la procédure **glm()** du package « stats », chargé automatiquement au démarrage de R, accepte les prédictives numériques et catégorielles (de type interne « factor ») pour la régression.

```
#régression sur l'ensemble des variables
lr.all <- glm(disease ~ ., data = heart, family = binomial)
#affichage des résultats
print(summary(lr.all))
```

Nous obtenons le modèle suivant avec un Akaike AIC = 230.42.

```
Call:
glm(formula = disease ~ ., family = binomial, data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7327  -0.4903  -0.1488   0.3081   2.7584

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -2.569400   2.823676  -0.910 0.362850
age                         -0.013652   0.024729  -0.552 0.580898
sexmale                      1.658410   0.539754   3.073 0.002123 **
cpatyp_angina               -0.922484   0.561798  -1.642 0.100586
cpnon_anginal               -1.886190   0.496577  -3.798 0.000146 ***
cptyp_angina                -2.055757   0.660289  -3.113 0.001849 **
trestbps                     0.018961   0.011214   1.691 0.090857 .
chol                         0.004058   0.004056   1.000 0.317117
fbst                        -0.348979   0.581774  -0.600 0.548604
restecgnormal               -0.455955   0.382846  -1.191 0.233668
restecgst_t_wave_abnormality 0.521230   2.507208   0.208 0.835313
thalach                     -0.017445   0.010980  -1.589 0.112099
exangyes                     0.821227   0.439628   1.868 0.061762 .
oldpeak                      0.408720   0.230869   1.770 0.076669 .
slopeflat                    0.705005   0.849516   0.830 0.406601
slopeup                     -0.465609   0.925715  -0.503 0.614983
ca                           1.288356   0.276890   4.653 3.27e-06 ***
thalnormal                   0.135676   0.781686   0.174 0.862204
thalreversable_defect        1.453132   0.770360   1.886 0.059254 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 192.42  on 284  degrees of freedom
AIC: 230.42

Number of Fisher Scoring iterations: 6
```

Figure 2 - Régression avec l'ensemble des variables et indicatrices sous R

Nous constatons que les variables explicatives catégorielles ont été automatiquement codées. R choisit la première valeur, dans l'ordre alphabétique, comme modalité de référence, à savoir pour chaque variable (cf. la liste des explicatives catégorielles et de leur modalités dans la Figure 1) : SEX = FEMALE, CP = ASYMPT, FBS = F, RESTECG = LEFT_VENT_HYPER, EXANG = NO, SLOPE = DOWN, THAL = FIXED_DEFECT.

Ce choix est important pour la lecture des résultats. Pour la variable SEX par exemple, nous constatons qu'elle est significative au risque 1% (le coefficient associée à l'indicatrice est significativement différent de 0) et que son signe est positif c.-à-d. par rapport aux femmes, les hommes ont plus de chances de contracter la maladie (PRESENCE est la modalité positive de la variable cible de la variable cible DISEASE, le logiciel R choisit automatiquement la seconde valeur dans l'ordre alphabétique).

A contrario, si nous avions défini SEX = MALE comme modalité de référence, nous aurions obtenu le même coefficient en valeur absolue mais de signe opposé. La lecture est cohérente (elle doit l'être) avec la précédente. Par rapport aux hommes, les femmes ont moins de risque de contracter la maladie.

Le choix de modalité de référence ne joue pas sur la qualité de la prédiction. Quelle que soit la modalité de référence adoptée pour chaque variable explicative catégorielle, nous obtiendrons exactement la même prédiction lorsque le modèle est appliqué sur un individu de la population.

3.2 Sélection Backward - Critère AIC

Nous désirons effectuer une sélection « backward » c.-à-d. nous partons de la régression incluant l'ensemble des variables, que nous éliminons au fur et à mesure au regard d'un critère statistique. La fonction stepAIC() de R (package MASS) procède en optimisant le critère AIC (Akaike). Ce dernier traduit l'arbitrage entre la qualité de l'ajustement (quantifiée par la déviance du modèle) et sa complexité (nombre de paramètres à estimer).

```
#sélection backward - critère AIC
library(MASS)
lr.back <- stepAIC(lr.all, direction = "backward")
print(summary(lr.back))
```

Nous obtenons la régression suivante :

```
Call:
glm(formula = disease ~ sex + cp + trestbps + thalach + exang +
    oldpeak + slope + ca + thal, family = binomial, data = heart)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.7652  -0.4907  -0.1557   0.3275   2.7477

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          -3.02506    2.22560  -1.359  0.17408
sexmale               1.54105    0.50238   3.068  0.00216 **
cpatyp_angina        -0.93653    0.55707  -1.681  0.09273 .
cpnon_anginal        -1.95838    0.48481  -4.039 5.36e-05 ***
cptyp_angina         -2.10448    0.65130  -3.231  0.00123 **
trestbps              0.01857    0.01034   1.796  0.07253 .
thalach              -0.01471    0.01004  -1.465  0.14295
exangyes              0.79254    0.43179   1.835  0.06643 .
oldpeak               0.43673    0.22313   1.957  0.05032 .
slopeflat             0.77596    0.83198   0.933  0.35099
slopeup              -0.43071    0.90105  -0.478  0.63265
ca                    1.24240    0.25821   4.812 1.50e-06 ***
thalnormal            0.28579    0.76239   0.375  0.70776
thalreversable_defect 1.59118    0.74694   2.130  0.03315 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 417.64  on 302  degrees of freedom
Residual deviance: 195.90  on 289  degrees of freedom
AIC: 223.9

Number of Fisher Scoring iterations: 6
```

Figure 3 - Régression après sélection de variables - stepAIC sous R

Par rapport à la régression incluant toutes les variables avec un AIC = 230.42 (Figure 2), celle-ci est meilleure avec un AIC plus faible égal à 223.9 (Figure 3). Nous constatons surtout, c'est ce qui nous intéresse au plus haut point dans ce tutoriel, que les indicatrices des explicatives catégorielles à plus de 2 modalités sont traitées en bloc : soit elles sont toutes exclues (CP, RESTEG), soit conservées (CP, SLOPE, THAL).

```
Start:  AIC=230.42
disease ~ age + sex + cp + trestbps + chol + fbs + restecg +
    thalach + exang + oldpeak + slope + ca + thal

           Df Deviance    AIC
- restecg   2   193.94 227.94
- age       1   192.72 228.72
- fbs       1   192.78 228.78
- chol      1   193.40 229.40
<none>          192.42 230.42
- thalach   1   195.04 231.04
- trestbps  1   195.34 231.34
- oldpeak   1   195.70 231.70
- exang     1   195.87 231.87
- slope     2   199.10 233.10
- thal      2   204.02 238.02
- sex       1   202.79 238.79
- cp        3   212.11 244.11
- ca        1   220.22 256.22

Step:  AIC=227.94
disease ~ age + sex + cp + trestbps + chol + fbs + thalach +
    exang + oldpeak + slope + ca + thal
```

Figure 4 - Première étape de StepAIC sous R

Ce traitement apparaît très clairement lorsque l'on inspecte la trace des calculs. A la première étape (Figure 4), on cherche à exclure la pire variable. Il s'agit de RESTECG qui minimise le critère AIC à 227.94. Ses deux indicatrices sont exclues simultanément comme l'atteste le degré de liberté (la colonne DF – « degree of freedom », DF = 2 pour RESTECG) dans le tableau retraçant les résultats intermédiaires.

Ce comportement paraît cohérent par rapport au cahier des charges initial. On cherche à déterminer le meilleur sous-ensemble de variables pour la régression. Lorsqu'une variable s'exprime à travers un ensemble d'indicatrices, nous ne pouvons pas les dissocier. Nous verrons plus loin que ce raisonnement peut être remis en cause car, au-delà de la sélection de variables, nous cherchons aussi à produire le modèle le plus pertinent. Pouvoir traiter de manière individuelle les indicatrices d'une variable catégorielle introduit une liberté supplémentaire dans la recherche des solutions, et permet d'aboutir à des modèles peut-être plus performants, en tous les cas avec un comportement prédictif différent. Le tout alors est de pouvoir interpréter correctement la solution mise en évidence.

4 Régression sous SAS

Nous analysons la sélection de variables sous SAS dans cette section. Nous allons à l'essentiel. Pour ce qui est de l'importation des données et de la démarche globale, je conseille la lecture du didacticiel « La proc logistic de SAS 9.3 » (avril 2012)[2].

4.1 Régression avec la totalité des variables

Les données HEART ont été stockées dans la banque DATAREG que nous avons créée. Nous utilisons la PROC LOGISTIC de SAS :

```
proc logistic data = datareg.heart;
class sex cp fbs restecg exang slope thal / param = reference ref = first;
model disease (event = last) = age sex cp trestbps chol fbs restecg thalach
exang oldpeak slope ca thal;
run;
```

Le paramètre CLASS joue un rôle essentiel dans le dispositif. Il indique les variables explicatives catégorielles. PARAM spécifie le type de codage à utiliser et REF la modalité de référence.

[2] http://tutoriels-data-mining.blogspot.fr/2012/04/la-proc-logistic-de-sas-93.html

Dans notre exemple, PARAM = REFERENCE préconise le codage 0/1 tel que nous l'avons décrit dans l'introduction ; avec REF = FIRST, nous indiquons à SAS que la première valeur (dans l'ordre alphabétique) correspondra à la modalité de référence. Nous retrouvons ainsi le paramétrage par défaut de R (section 3.1) et les résultats à l'identique (Figure 5, AIC = 230.418).

Statistiques d'ajustement du modèle		
Critère	Constante uniquement	Constante et covariables
AIC	419.638	230.418
SC	423.352	300.979
-2 Log	417.638	192.418

Estimations par l'analyse du maximum de vraisemblance						
Paramètre		DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept		1	-2.5693	2.8236	0.8280	0.3629
age		1	-0.0137	0.0247	0.3048	0.5809
sex	male	1	1.6583	0.5397	9.4400	0.0021
cp	atyp_angina	1	-0.9225	0.5618	2.6963	0.1006
cp	non_anginal	1	-1.8861	0.4966	14.4269	0.0001
cp	typ_angina	1	-2.0557	0.6603	9.6928	0.0018
trestbps		1	0.0190	0.0112	2.8589	0.0909
chol		1	0.00406	0.00406	1.0007	0.3171
fbs	t	1	-0.3490	0.5818	0.3598	0.5486
restecg	normal	1	-0.4559	0.3828	1.4183	0.2337
restecg	st_t_wave_abnor	1	0.5209	2.5070	0.0432	0.8354
thalach		1	-0.0174	0.0110	2.5242	0.1121
exang	yes	1	0.8212	0.4396	3.4893	0.0618
oldpeak		1	0.4087	0.2309	3.1339	0.0767
slope	flat	1	0.7050	0.8495	0.6887	0.4066
slope	up	1	-0.4656	0.9257	0.2530	0.6150
ca		1	1.2883	0.2769	21.6490	<.0001
thal	normal	1	0.1356	0.7817	0.0301	0.8622
thal	reversable_defect	1	1.4531	0.7703	3.5579	0.0593

Figure 5 - Régression logistique sous SAS

A la différence que SAS n'oublie pas que les indicatrices sont associées à des variables. Il va au-delà du test de Wald de significativité individuelle des indicatrices en proposant dans un second tableau le test de significativité des variables (Figure 6). Voyons ce que cela veut dire.

Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > Khi-2
age	1	0.3048	0.5809
sex	1	9.4400	0.0021
cp	3	17.7134	0.0005
trestbps	1	2.8589	0.0909
chol	1	1.0007	0.3171
fbs	1	0.3598	0.5486
restecg	2	1.5117	0.4696
thalach	1	2.5242	0.1121
exang	1	3.4893	0.0618
oldpeak	1	3.1339	0.0767
slope	2	6.4922	0.0389
ca	1	21.6490	<.0001
thal	2	11.2454	0.0036

Figure 6 - Analyse des effets de type 3 de SAS

Lorsque la variable est numérique ou binaire, les degrés de liberté, la statistique de test et la p-value (probabilité critique) sont identique à ceux du tableau précédent (Figure 5). En revanche, quand elle catégorielle et porte plus de 2 modalités, SAS procède à un test de nullité simultanée des coefficients de toutes les indicatrices associées à la variable.

Prenons la variable THAL pour un niveau de significativité à 1% pour préciser notre propos. Apparemment, ni THAL = NORMAL, ni THAL = REVERSABLE DEFECT, pris individuellement ne semblent induire un écart de risque significatif par rapport à la référence THAL = FIXED DEFECT (Figure 5). Mais lors du test de significativité globale de la variable – est-ce qu'on peut considérer que les deux coefficients sont simultanément nuls à 1% ? – la réponse est autre avec une p-value égale à 0.0036 (Figure 6). De fait, traiter les indicatrices individuellement ou en blocs relatifs aux variables correspondent bien à des schémas différents.

4.2 Sélection de variables sous SAS

La sélection backward repose sur le test de Wald sous SAS. Tout comme R, il traite les indicatrices de variables catégorielles en bloc c.-à-d. il s'appuie sur le tableau d'analyse des effets de type 3 (Figure 6) à chaque étape pour retirer la variable la moins pertinente. Les mécanisme est le suivant : (1) SAS repère la variable la moins significative, celle qui propose la p-value la plus élevée ; (2) il la retire du modèle si cette dernière est supérieure à un seuil

spécifié par le paramètre SLSTAY ; (3) il réapprend le modèle sur les variables restantes, et continue ainsi (1) et (2) jusqu'à ce que toutes les variables soient significatives.

Dans notre exemple, à la lumière du tableau des effets (Figure 6), ce serait l'explicative AGE qui serait éliminée à la première itération avec une p-value égale à 0.5809, supérieure au paramètre SLSTAY = 0.01. Pour ce qui est des suivantes, il faut lancer le processus complet.

Voici le code SAS de sélection backward pour la proc logistic.

```
proc logistic data = datareg.heart;
class sex cp fbs restecg exang slope thal / param = reference ref = first;
model disease (event = last) = age sex cp trestbps chol fbs restecg thalach
exang oldpeak slope ca thal / selection = backward slstay = 0.01;
run;
```

SAS récapitule dans un tableau les variables éliminées au fur et à mesure (Figure 7), 8 variables sont retirées du modèle en définitive.

Récapitulatif sur l'élimination en arrière					
Etape	Effet supprimé	DDL	Nombre dans	Khi-2 de Wald	Pr > Khi-2
1	age	1	12	0.3048	0.5809
2	fbs	1	11	0.3808	0.5372
3	restecg	2	10	1.4642	0.4809
4	chol	1	9	1.3571	0.2440
5	thalach	1	8	2.1458	0.1430
6	trestbps	1	7	2.6135	0.1060
7	oldpeak	1	6	4.8451	0.0277
8	exang	1	5	6.2304	0.0126

Figure 7 - Processus de suppression - Sélection backward sous SAS

Nous obtenons donc un modèle à 5 variables : SEX, CP, SLOPE, CA et THAL.

Analyse des effets Type 3			
Effet	DDL	Khi-2 de Wald	Pr > Khi-2
sex	1	8.2582	0.0041
cp	3	31.4955	<.0001
slope	2	20.3103	<.0001
ca	1	28.1557	<.0001
thal	2	17.7864	0.0001

Figure 8 - Analyse des effets de type 3 après sélection - SAS

Elles sont toutes significatives à 1% dans le tableau des effets de type 3 (Figure 8). Heureusement, le contraire constituerait une erreur dans ce processus de sélection.

En revanche, certaines indicatrices peuvent ne pas être significatives à 1% dans le tableau des coefficients du modèle (ex. les coefficients des indicatrices de SLOPE et THAL) (Figure 9).

Estimations par l'analyse du maximum de vraisemblance						
Paramètre		DDL	Valeur estimée	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept		1	-0.8728	0.9776	0.7972	0.3719
sex	male	1	1.3275	0.4619	8.2582	0.0041
cp	atyp_angina	1	-1.5571	0.5151	9.1399	0.0025
cp	non_anginal	1	-2.2028	0.4442	24.5872	<.0001
cp	typ_angina	1	-2.1126	0.5990	12.4390	0.0004
slope	flat	1	0.3641	0.6632	0.3013	0.5831
slope	up	1	-1.4595	0.6783	4.6289	0.0314
ca		1	1.2943	0.2439	28.1557	<.0001
thal	normal	1	-0.1845	0.7042	0.0687	0.7933
thal	reversable_defect	1	1.4054	0.6935	4.1072	0.0427

Figure 9 - Modèle après sélection backward sous SAS

On s'étonne souvent de cette apparente incohérence. Mais ce n'est pas une erreur. SAS s'appuie sur d'autres informations (le tableau des effets de type 3 en l'occurrence) pour éliminer les variables non pertinentes.

Un petit mot enfin concernant le critère AIC, celui de SAS est plus élevé que celui de R (232.167 vs. 223.9). Il ne faut pas s'en émouvoir, SAS ne cherche pas à minimiser explicitement ce critère contrairement à R. Les comparer n'est pas vraiment judicieux ici.

5 Régression sous Tanagra

Tanagra se distingue des outils de ce tutoriel en ne traitant pas directement les explicatives qualitatives. Nous devons les coder préalablement à l'aide d'un outil dédié (« Codage disjonctif complet », mars 2008)[3]. Il y a plusieurs raisons à cela : il me paraissait pédagogiquement souhaitable que les étudiants réalise de manière explicite le codage pour comprendre la

[3] http://tutoriels-data-mining.blogspot.fr/2008/03/codage-disjonctif-complet.html

nécessité et l'influence de cette étape dans la régression ; l'utilisateur peut approfondir l'analyse en introduisant d'autres types de codage (ce que fait SAS par exemple à l'aide options du paramètre CLASS) ; Tanagra s'appuie sur une stratégie spécifique dans la sélection de variables en traitant les indicatrices des variables de manière indépendante, le modèle optimisé est de nature différente par rapport aux approches précédentes.

5.1 Régression avec l'ensemble des variables

5.1.1 Importation des données

Nous chargeons le fichier « heart-c.xlsx » dans le tableur Excel, nous sélectionnons la plage des données et nous l'envoyons vers Tanagra à l'aide de l'add'in « tanagra.xla »[4].

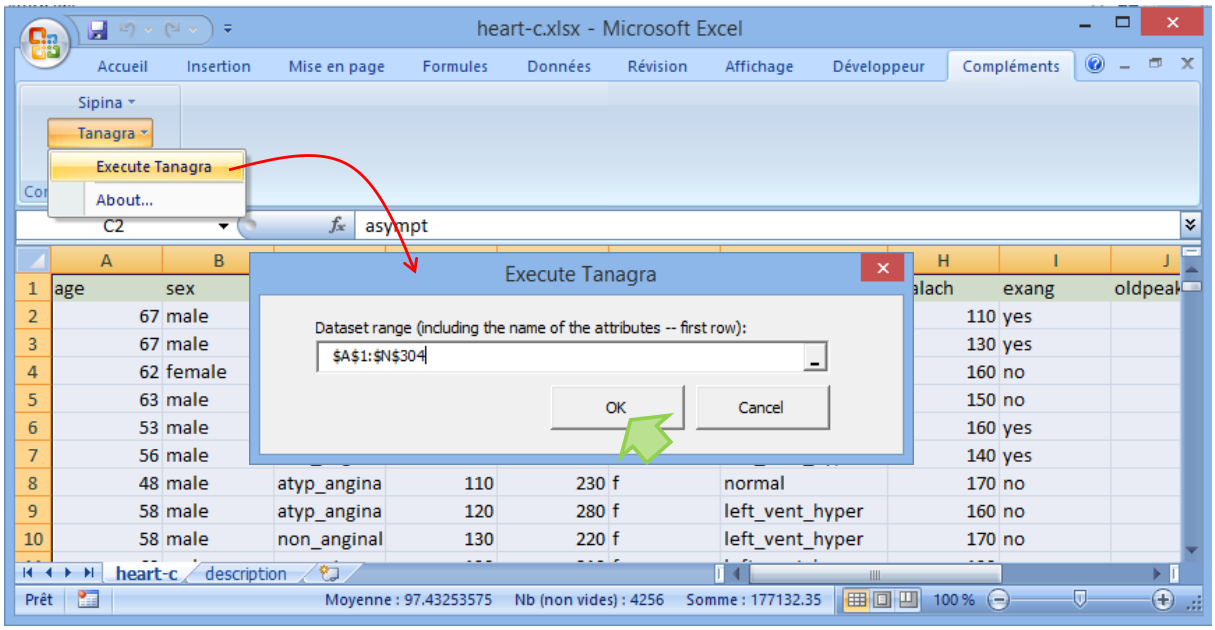
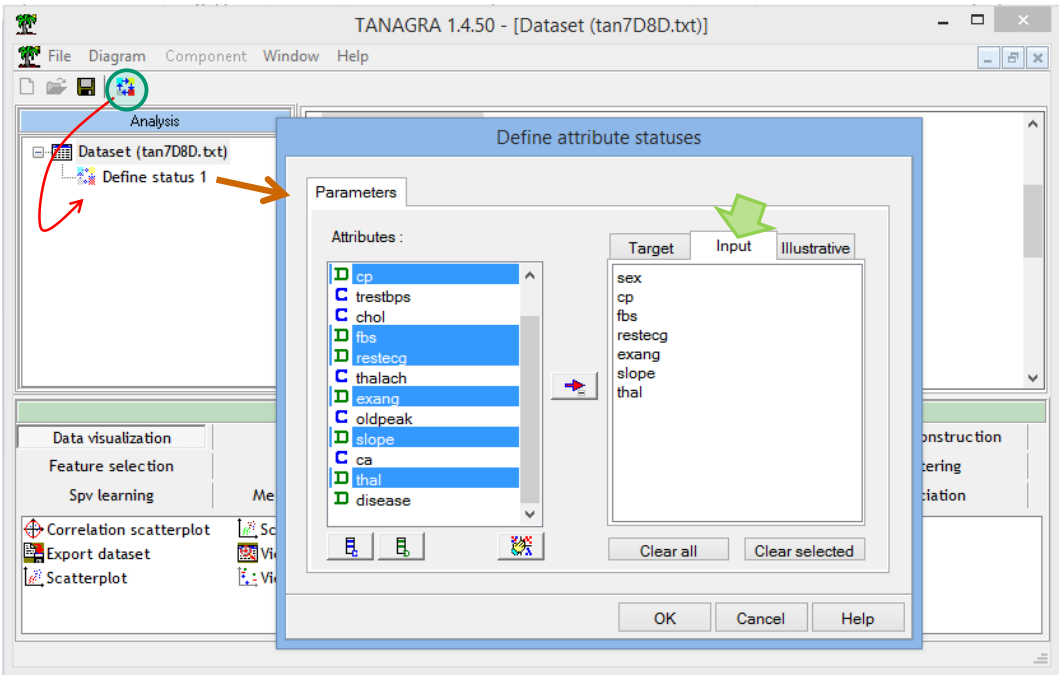


Figure 10 - Envoi des données d'Excel vers Tanagra

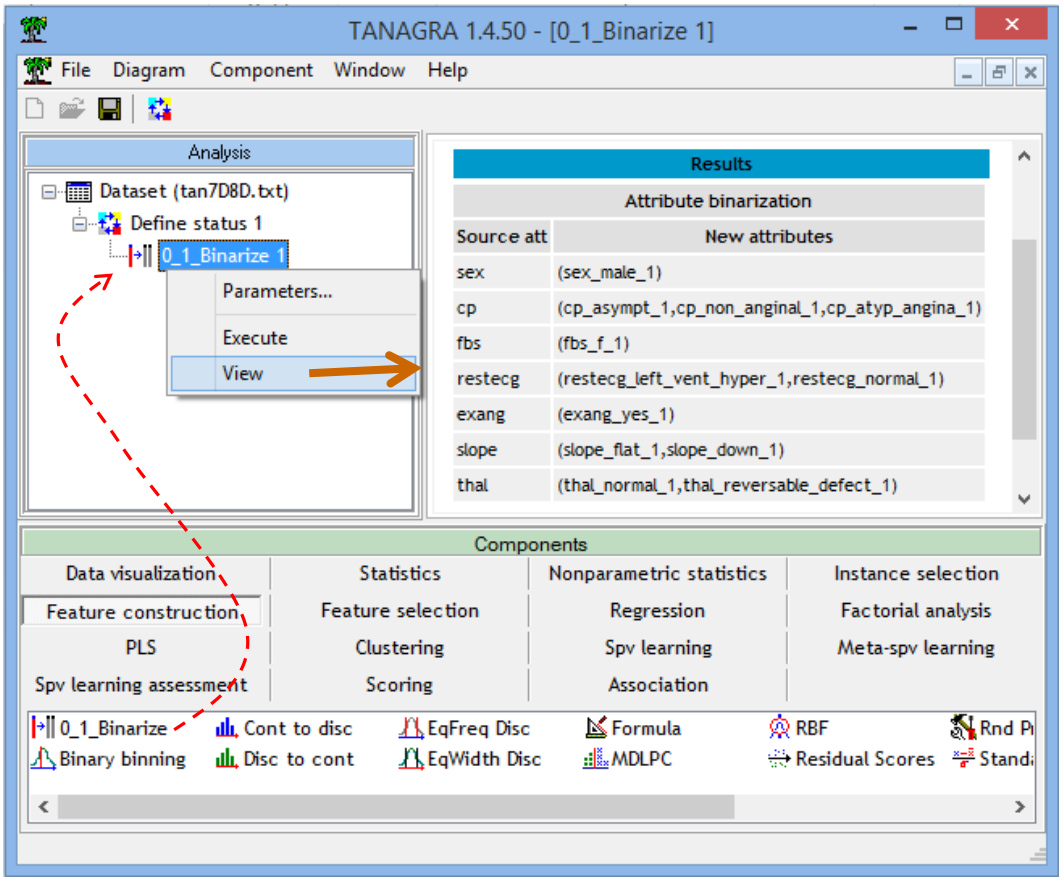
5.1.2 Codage des explicatives catégorielles

Tanagra est automatiquement démarré. Les variables disponibles sont listées. Nous disposons de 14 variables pour 303 individus. Nous nous proposons de coder les explicatives catégorielles en indicatrices. Nous devons tout d'abord spécifier les variables à traiter à l'aide de l'outil DEFINE STATUS accessible dans la barre d'outils.

[4] Tutoriel Tanagra, « L'add-in Tanagra pour Excel 2007 et 2010 », août 2010 ; http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html.



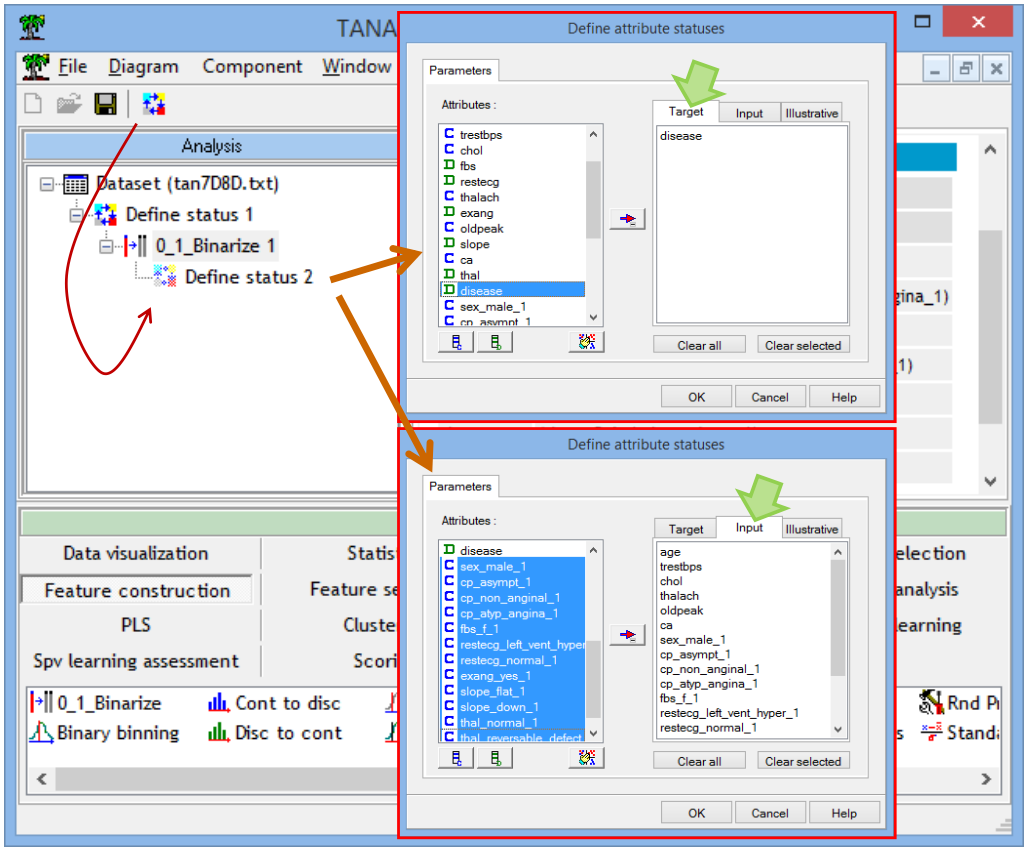
Nous plaçons en INPUT les variables catégorielles sauf DISEASE qui constitue la variable cible. Puis nous insérons le composant 0_1_BINARIZE (onglet FEATURE CONSTRUCTION) qui se charge de créer les indicatrices. Nous cliquons enfin sur le menu contextuel VIEW.



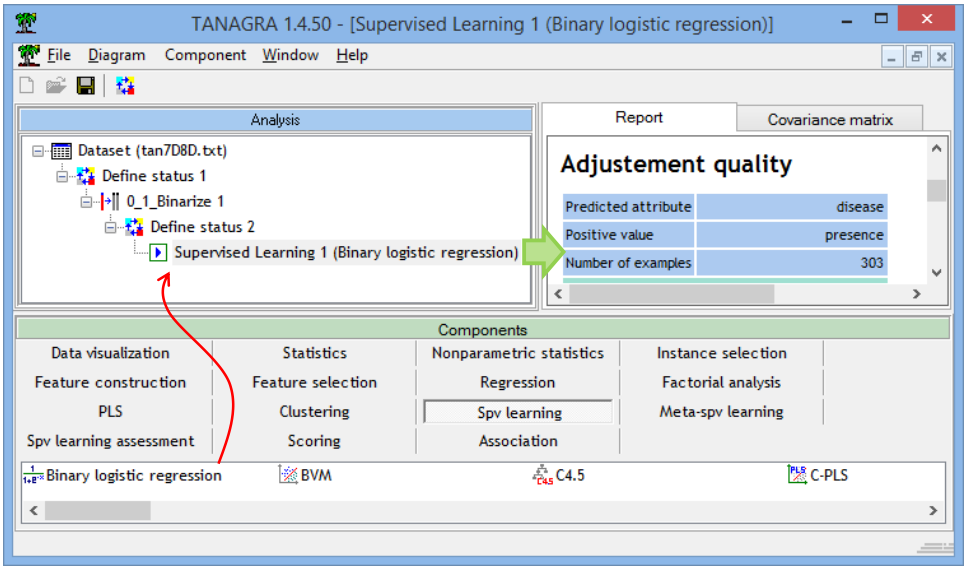
Nous disposons de la liste des indicatrices. La valeur omise constitue la modalité de référence.

5.1.3 Régression logistique

Nous insérons DEFINE STATUS pour spécifier la variable cible (TARGET = DISEASE) et les explicatives (INPUT = les variables numériques + les indicatrices) de la régression logistique.



Nous ajoutons le composant BINARY LOGISTIC REGRESSION (onglet SPV LEARNING). Nous cliquons sur le menu contextuel VIEW pour obtenir les résultats de la régression.



Voyons-en le détail.

Model Fit Statistics		
Criterion	Intercept	Model
AIC	419.638	230.418
SC	423.352	300.979
-2LL	417.638	192.418
Model Chi² test (LR)		
Chi-2		225.2201
d.f.		18
P(>Chi-2)		0.0000
R²-like		
McFadden's R²		0.5393
Cox and Snell's R²		0.5245
Nagelkerke's R²		0.7011

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-4.918515	3.9539	1.5474	0.2135
age	-0.013652	0.0247	0.3048	0.5809
trestbps	0.018961	0.0112	2.8591	0.0909
chol	0.004058	0.0041	1.0008	0.3171
thalach	-0.017445	0.0110	2.5244	0.1121
oldpeak	0.408720	0.2309	3.1341	0.0767
ca	1.288356	0.2769	21.6498	0.0000
sex_male_1	1.658410	0.5398	9.4404	0.0021
cp_asympt_1	2.055757	0.6603	9.6934	0.0018
cp_non_anginal_1	0.169567	0.6563	0.0668	0.7961
cp_atyp_angina_1	1.133273	0.7643	2.1986	0.1381
fbs_f_1	0.348979	0.5818	0.3598	0.5486
restecg_left_vent_hyper_1	-0.521230	2.5072	0.0432	0.8353
restecg_normal_1	-0.977185	2.5054	0.1521	0.6965
exang_yes_1	0.821227	0.4396	3.4894	0.0618
slope_flat_1	1.170614	0.4711	6.1742	0.0130
slope_down_1	0.465609	0.9257	0.2530	0.6150
thal_normal_1	0.135676	0.7817	0.0301	0.8622
thal_reversable_defect_1	1.453132	0.7704	3.5581	0.0593

Figure 11 - Régression sur l'ensemble des variables et indicatrices - Tanagra

Nous avons exactement le même modèle que R et SAS comme en atteste la déviance (-2LL) ou encore le critère AIC c.-à-d. appliqué sur les mêmes individus, ces classifieurs produiront les mêmes scores et, a fortiori, les mêmes prédictions.

Mais, les modalités de référence sont différentes pour certaines variables, les coefficients associés le sont également. Prenons les résultats de R pour fixer les idées (Figure 2) :

- Pour SEX, FEMALE est la référence pour les deux logiciels R et Tanagra, le coefficient de la régression est bien le même 1.658410.
- Pour FS, R a choisit la modalité « F » comme référence, Tanagra a utilisé « T », les coefficients ne sont pas les mêmes, soit respectivement -0.348979 et 0.348979. Ils sont de signe opposé lorsque la variable est binaire.
- Pour SLOPE qui n'est pas binaire, R a choisit « DOWN » tandis que Tanagra a utilisé « UP ». Les coefficients de R sont $a_{FLAT/DOWN}$ = 0.705005 et $a_{UP/DOWN}$ = -0.465609, ceux de Tanagra sont $a_{FLAT/UP}$ = 1.170614 et $a_{DOWN/UP}$ = 0.465609. Nous constatons qu'il est très facile de retrouver les coefficients de R à partir de ceux de Tanagra, en effet :

$$a_{UP/DOWN} = - a_{DOWN/UP}$$

$$a_{FLAT/DOWN} = a_{FLAT/UP} - a_{DOWN/UP}$$

De fait, quelle que soit la modalité de référence utilisée dans le codage des indicatrices 0/1, le comportement prédictif et les interprétations sont préservées. Heureusement, nous serions soumis à l'arbitraire de son choix dans le cas contraire.

5.2 Sélection backward sous Tanagra

Tanagra va réellement se démarquer lors de la sélection de variables. En effet, il traite de manière individuelle les indicatrices sans tenir compte de leur association aux variables.

Nous insérons le composant BACKWARD-LOGIT (onglet FEATURE SELECTION) dans le diagramme. Nous le plaçons à la suite du DEFINE STATUS 2 spécifiant les variables candidates à la régression initiale. Le niveau de signification par défaut est 1%.

Tanagra nous annonce 6 variables et/ou indicatrices sont (Figure 12) : OLDPEAK, CA, SEX = MALE, CP = ASYMPT, SLOPE = FLAT, THAL = REVERSABLE DEFECT. Pour CP, SLOPE et THAL qui sont des variables catégorielles à plus de 2 modalités, les indicatrices ont été dissociées, certaines seulement ont été éliminées.

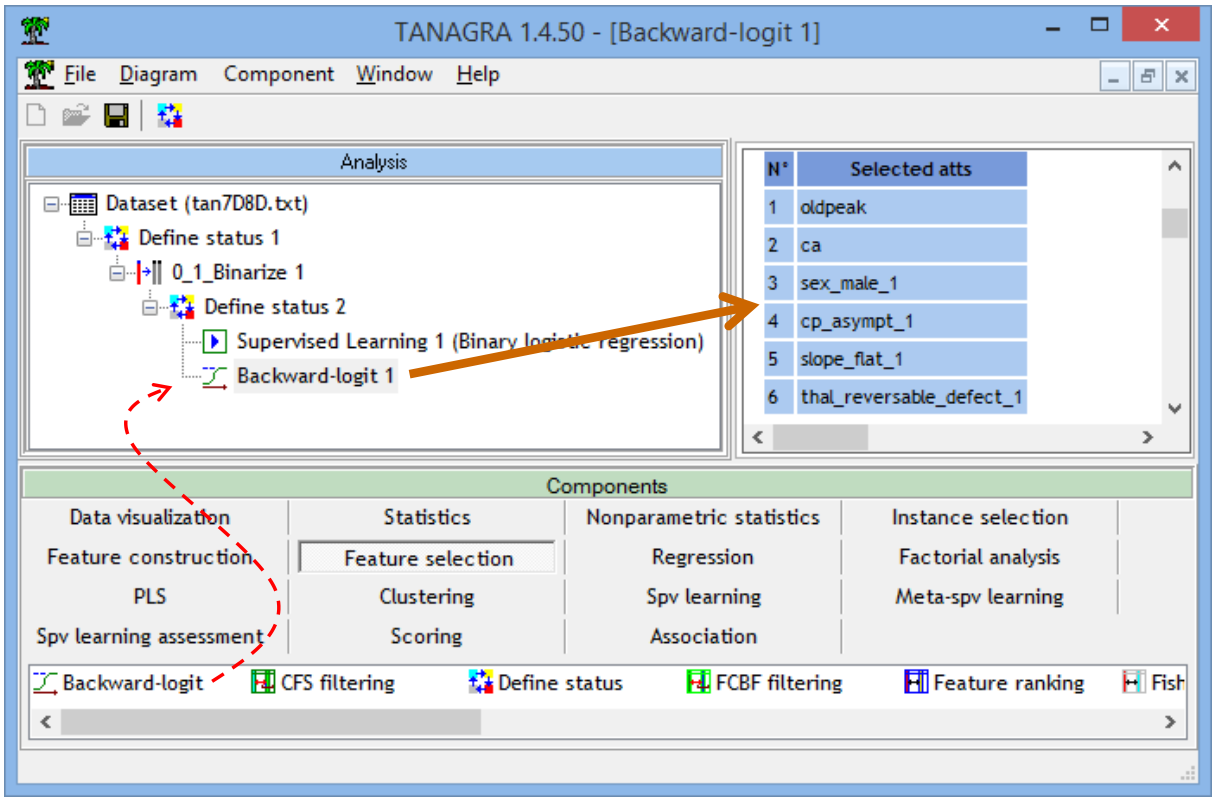


Figure 12 - Composant Bacwkard-Logit pour la sélection de variables - Tanagra

Nous réalisons la régression à l'aide du composant BINARY LOGISTIC REGRESSION.

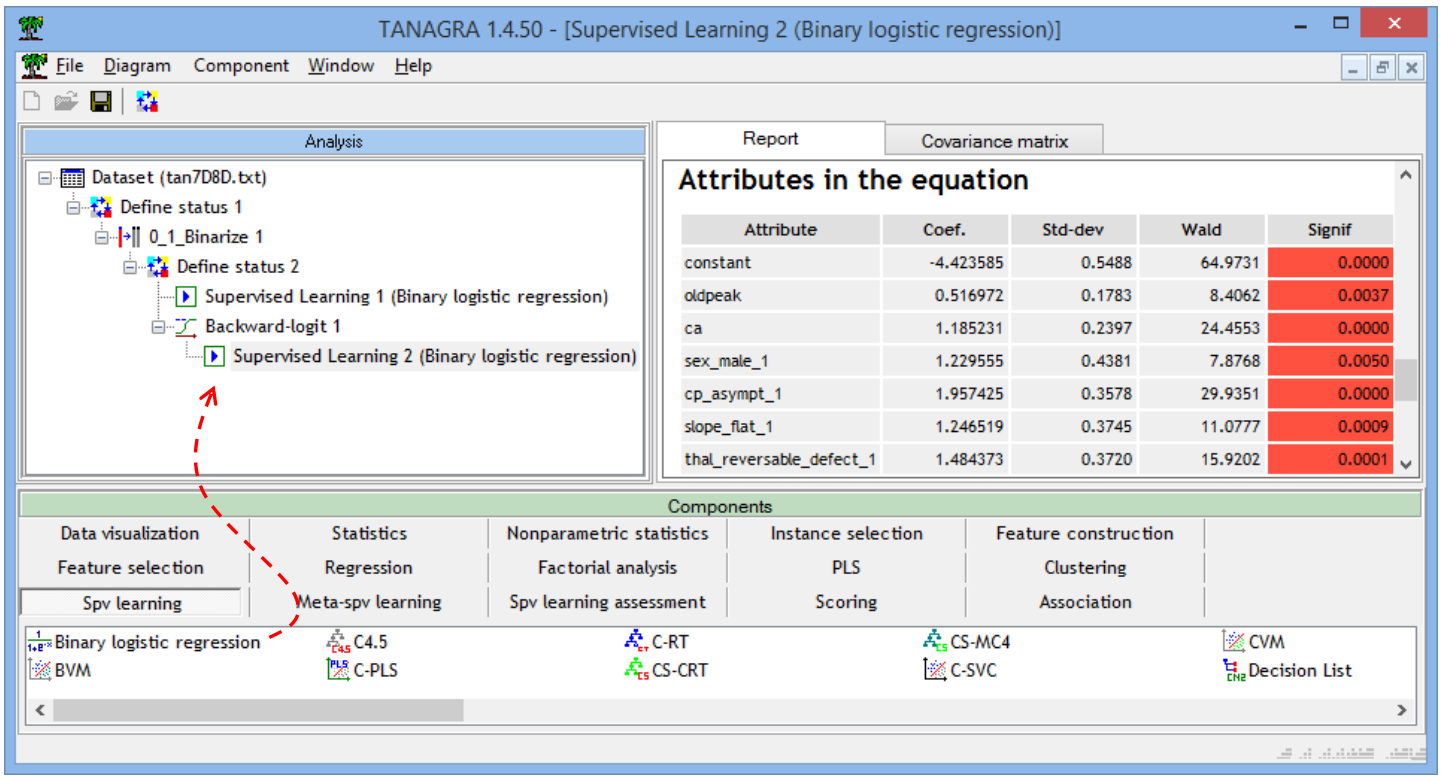


Figure 13 - Régression après sélection backward - Tanagra

Le modèle présente un AIC égal à 222.978 (non visible dans la copie d'écran). Contrairement à SAS, toutes les variables et indicatrices sont significatives à 1% (Figure 13). Ca ne veut pas dire que c'est mieux, il s'agit là tout simplement de la conséquence de la stratégie de sélection opérée par le logiciel qui traite les indicatrices individuellement.

Est-ce que ce modèle a un sens ? La réponse est OUI. Les valeurs absentes deviennent les modalités de référence. Considérons la variable SLOPE pour étayer notre propos. Nous avons vu dans la régression impliquant toutes les indicatrices que le coefficient $a_{DOWN/UP}$ = 0.465609 n'était pas significatif du tout avec une p-value = 0.6150 (Figure 11). En retirant l'indicatrice DOWN, le système nous indique qu'il est judicieux de fusionner les modalités DOWN et UP, un peu à la manière des arbres de décision qui effectuent des fusions de feuilles lors d'une segmentation (ex. méthode CHAID et apparentés[5]). Ainsi, le coefficient $a_{FLAT/UP,DOWN}$ = 1.246519 indique le surcroît de risque associé à FLAT par rapport aux situations {UP, DOWN} réunies.

6 Régression sous SPAD - Comparaison des approches

SPAD met tout le monde d'accord en intégrant les deux approches. Il peut réaliser la sélection backward basée sur le test de Wald en traitant en bloc les indicatrices des explicatives catégorielles, à la manière de SAS, ou en les traitant de manière individuelle, à la manière de Tanagra. Nous verrons dans cette section les options à spécifier pour obtenir le comportement souhaité dans le logiciel.

6.1 Régression logistique sous SPAD

Le plus simple est de consulter le tutoriel consacré aux arbres interactifs pour la création d'un projet et l'importation d'un fichier Excel[6]. Attention, CA est une variable quantitative qu'il faut stocker en « réel ». Il faut l'indiquer explicitement dans les métadonnées lors de l'importation. Nous insérons ensuite le composant REGRESSION LOGISTIQUE (branche SCORING ET MODELISATION dans les METHODES), auquel nous relions la source Excel. Nous cliquons sur le menu contextuel PARAMETRES pour accéder au paramétrage.

[5] Ricco Rakotomalala, « Arbres de décision », Revue Modulad, n°33, pp. 163-187, 2005.

[6] Tutoriel Tanagra, « Nouveaux arbres interactifs dans SPAD 8 », août 2014.

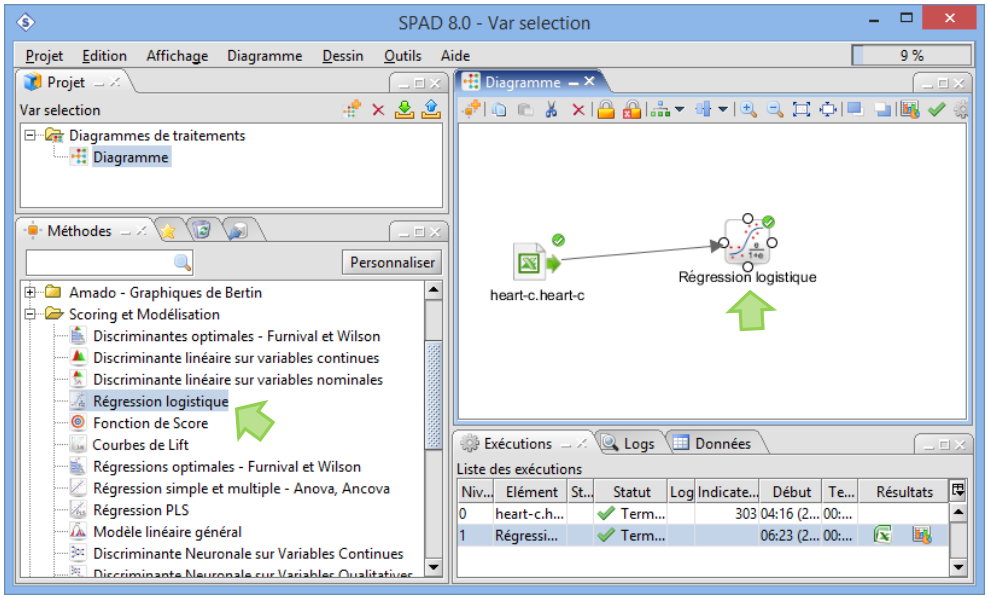


Figure 14 - Diagramme "Régression Logistique" sous SPAD

Dans l'onglet « Modèle », nous spécifions la cible (DISEASE) et les explicatives.

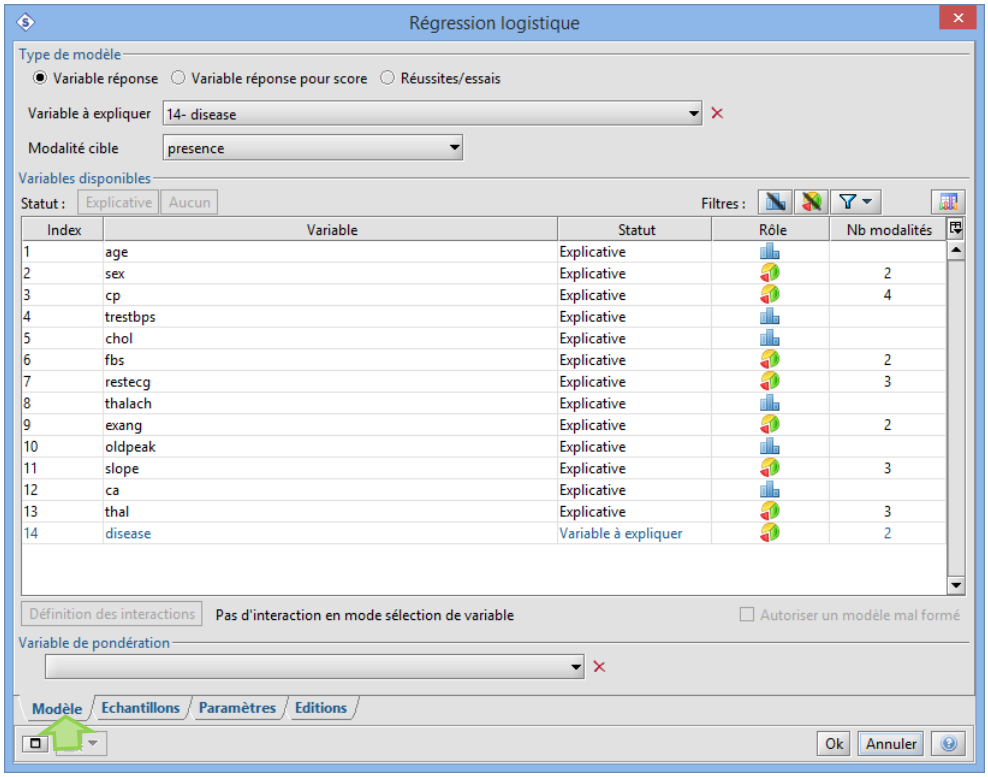


Figure 15 - Spécification des variables de l'analyse - SPAD

Nous souhaitons disposer d'une estimation crédible des performances en prédiction[7].

[7] Je décris dans le support suivant les schémas types d'échantillonnage permettant de disposer d'une estimation « crédible » des performances en prédiction des modèles en apprentissage supervisé : Tutoriel Tanagra, « Estimation de l'erreur de prédiction – Les techniques de ré-échantillonnage », février 2015.

Dans l'onglet « Paramètres », nous indiquons utiliser l'ensemble des variables explicatives disponibles (METHODE DE SELECTION = PAS DE SELECTION DE VARIABLES).

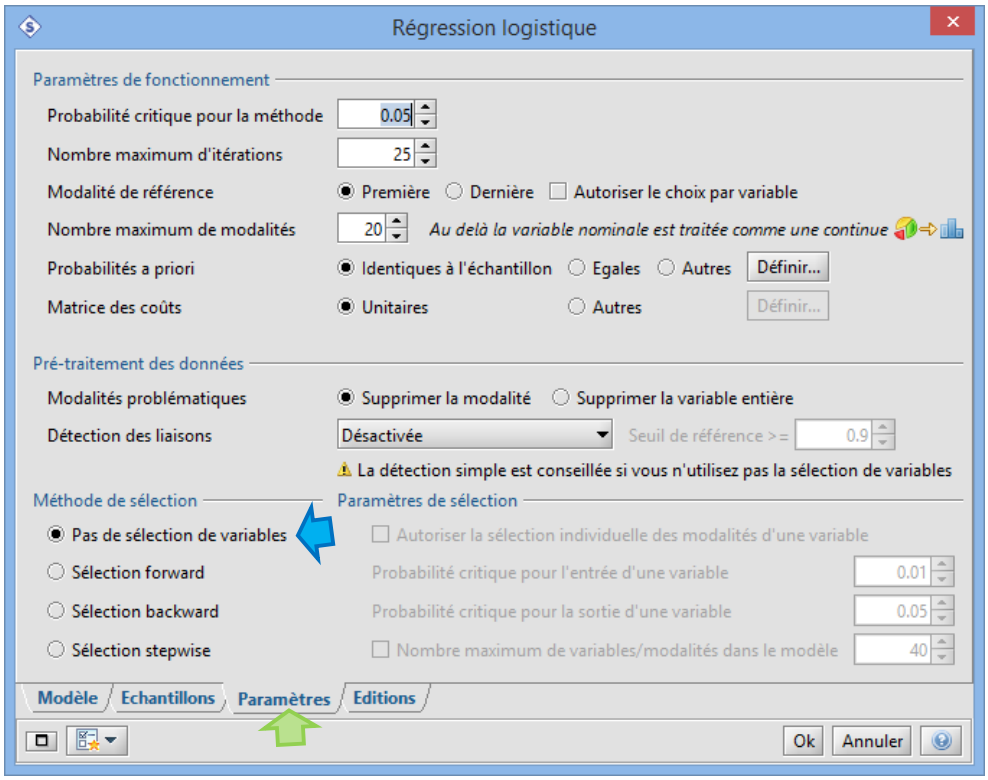


Figure 16 - Paramétrage - Pas de sélection de variables - SPAD

Nous validons. Les calculs sont automatiquement démarrés. Nous y avons accès en cliquant sur l'icône Excel dans la partie « Résultats ».

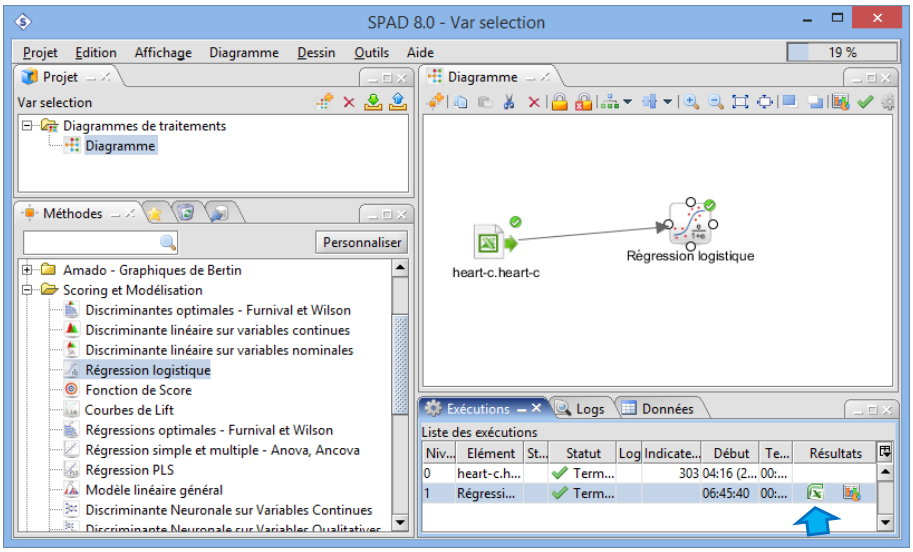


Figure 17- SPAD - Accès aux résultats de la régression logistique

Les sorties sont réparties sur plusieurs feuilles. Dans « **REG_MODEL** », nous disposons des indicateurs de qualité globale du modèle et du tableau des coefficients (Figure 18).

Résultats du modèle

Ajustement du modèle

Indicateurs	Constante (intercept)	Modèle
Critère d'Akaike	419.638	230.418
Critère BIC	423.352	300.979
Déviance	417.638	192.418
R2 de Cox et Snell	0.524	
Coefficient de Nagelkerke	0.701	
Pseudo R2 de McFadden	0.539	

La solution a été trouvée en 6 itérations

Test du rapport de vraisemblance

Rapport de vraisemblance	225.220
Ddl	18.000
p-valeur	0.000

Coefficients du modèle par variable

Coefficients de régression estimés par maximum de vraisemblance

Variable	Modalité	Coefficient	Erreur standard	Khi-2 de Wald	P-valeur
age		-0.014	0.025	0.305	0.581
sex (ddl 1)				9.440	0.002
sex	male	1.658	0.540	9.440	0.002
cp (ddl 3)				17.714	0.001
cp	atyp_angina	-0.922	0.562	2.696	0.101
cp	non_anginal	-1.886	0.497	14.428	0.000
cp	typ_angina	-2.056	0.660	9.693	0.002
trestbps		0.019	0.011	2.859	0.091
chol		0.004	0.004	1.001	0.317
fbs (ddl 1)				0.360	0.549
fbs	t	-0.349	0.582	0.360	0.549
restecg (ddl 2)				1.512	0.470
restecg	normal	-0.456	0.383	1.418	0.234
restecg	st_t_wave_abnorma	0.521	2.507	0.043	0.835
thalach		-0.017	0.011	2.524	0.112
exang (ddl 1)				3.489	0.062
exang	yes	0.821	0.440	3.489	0.062
oldpeak		0.409	0.231	3.134	0.077
slope (ddl 2)				6.492	0.039
slope	flat	0.705	0.850	0.689	0.407
slope	up	-0.466	0.926	0.253	0.615
ca		1.288	0.277	21.650	0.000
thal (ddl 2)				11.246	0.004
thal	normal	0.136	0.782	0.030	0.862
thal	reversable_defect	1.453	0.770	3.558	0.059
Constante (intercept)		-2.569	2.824	0.828	0.363

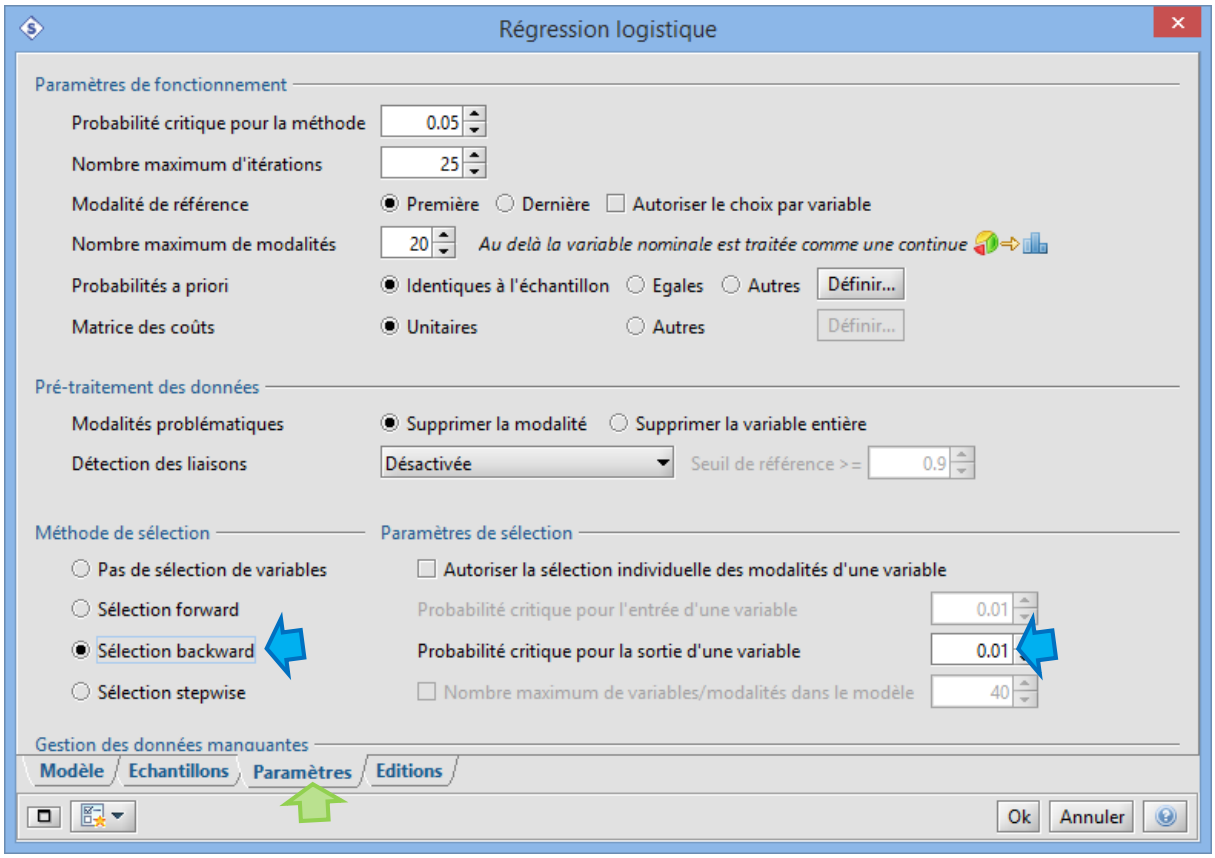
Figure 18 - Résultats du modèle - SPAD - Echantillon d'apprentissage (80%)

SPAD mixe le tableau des coefficients et l'analyse de type 3. Nous pouvons ainsi identifier en un coup d'œil les indicatrices et les variables pertinentes. Par exemple, au risque 1%, les deux indicatrices prises individuellement de THAL ne sont pas significatives (THAL = NORMAL, p-value = 0.862 ; THAL = REVERSABLE_DEFECT, p-value = 0.059), mais nous ne pouvons pas

les retirer car on ne peut pas considérer qu'elles sont nulles simultanément (THAL, p-value = 0.004). Ca peut paraître déroutant comme je l'ai dit plus haut. Il ne faut pas oublier simplement que ces deux indicatrices ne sont pas indépendantes.

6.2 Sélection backward - Traitement en blocs des indicatrices

Pour la régression backward à 1% avec un traitement en blocs des indicatrices des variables, nous retournons dans la boîte de paramétrage. Nous spécifions :



Nous validons, nous obtenons directement les résultats en cliquant sur l'icône Excel.

Les sorties intègrent la description du processus de sélection (Figure 19), nous disposons :

- (A) De la liste des variables éliminées à chaque étape, avec les statistiques de test, équivalent au récapitulatif de SAS (Figure 7).
- (B) Des indicateurs de qualité du modèle « Ajustement du modèle », nous observons – entres autres – un AIC = 232.167.
- (C) Du test de rapport de vraisemblance indiquant la significativité globale du modèle.
- (D) Du tableau des coefficients incluant les analyses de type 3.

Sélection des variables du modèle

Sélection des variables en mode backward

Ordre	Variable	Direction	Khi2	P-valeur	Ddl
0	age	sortie	0.305	0.581	1
1	fbs	sortie	0.381	0.537	1
2	restecg	sortie	1.464	0.481	2
3	chol	sortie	1.357	0.244	1
4	thalach	sortie	2.146	0.143	1
5	trestbps	sortie	2.614	0.106	1
6	oldpeak	sortie	4.845	0.028	1
7	exang	sortie	6.230	0.013	1

(A)

Ajustement du modèle

Indicateurs	Constante (intercept)	Modèle
Critère d'Akaike	419.638	232.167
Critère BIC	423.352	269.305
Déviance	417.638	212.167
R2 de Cox et Snell	0.492	
Coefficient de Nagelkerke	0.658	
Pseudo R2 de McFadden	0.492	

La solution a été trouvée en 6 itérations

(B)

Test du rapport de vraisemblance

Rapport de vraisemblance	205.471
Ddl	9.000
p-valeur	0.000

(C)

Coefficients du modèle par variable

Coefficients de régression estimés par maximum de vraisemblance

Variable	Modalité	Coefficient	Erreur standard	Khi-2 de Wald	P-valeur
sex (ddl 1)				8.258	0.004
sex	male	1.327	0.462	8.258	0.004
cp (ddl 3)				31.496	0.000
cp	atyp_angina	-1.557	0.515	9.140	0.003
cp	non_anginal	-2.203	0.444	24.587	0.000
cp	typ_angina	-2.113	0.599	12.439	0.000
slope (ddl 2)				20.310	0.000
slope	flat	0.364	0.663	0.301	0.583
slope	up	-1.459	0.678	4.629	0.031
ca		1.294	0.244	28.156	0.000
thal (ddl 2)				17.786	0.000
thal	normal	-0.185	0.704	0.069	0.793
thal	reversable_defect	1.405	0.693	4.107	0.043
Constante (intercept)		-0.873	0.978	0.797	0.372

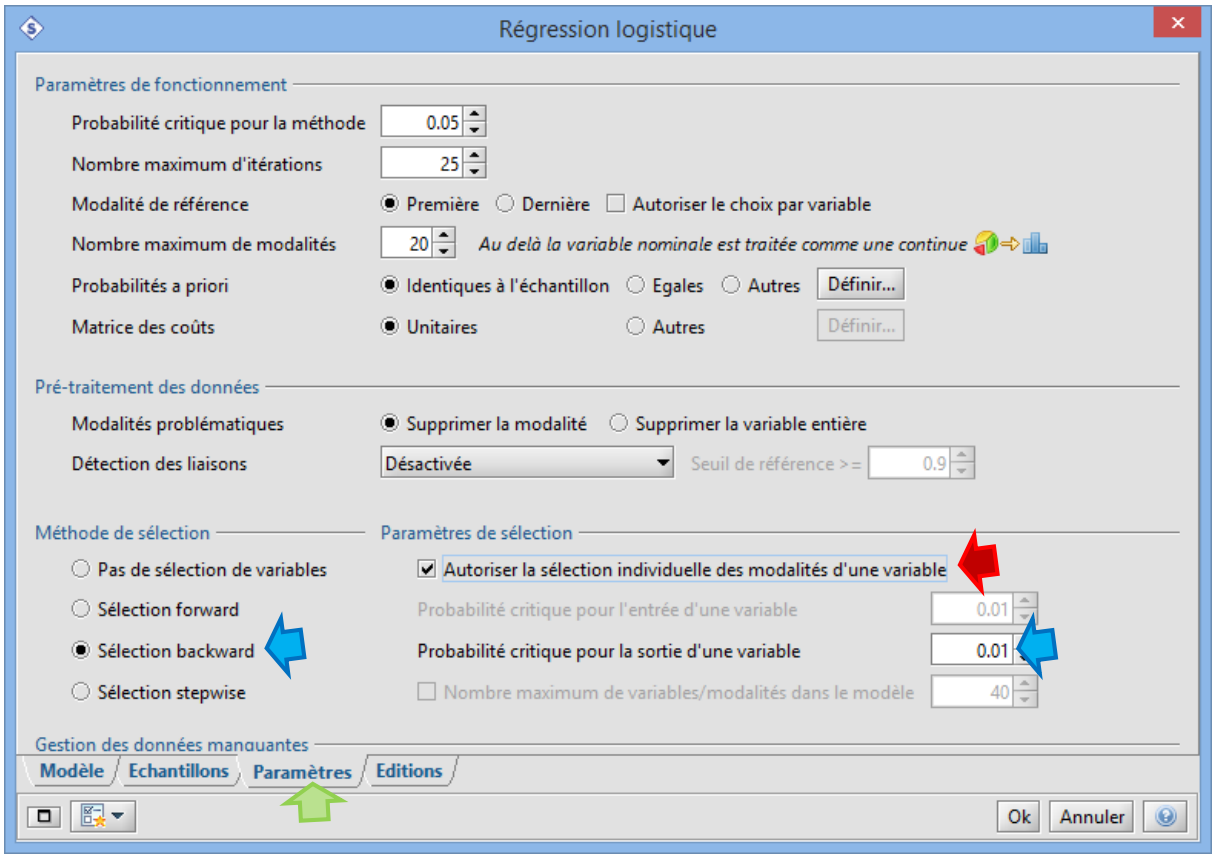
(D)

Figure 19 - Sélection backward - SPAD - Traitement en blocs des indicatrices

Les résultats rejoignent en tous points ceux de SAS.

6.3 Sélection backward - Traitement individuel des indicatrices

Pour traiter individuellement les indicatrices, nous revenons dans la boîte de paramétrage. En sus des indications précédent, nous précisons « Autoriser la sélection individuelle des modalités d'une variable ».



Le tableau de sélection retrace maintenant les sorties des variables quantitatives ou des indicatrices prises individuellement (Figure 20).

Sélection des variables du modèle

Sélection des variables en mode backward

Ordre	Variable	Modalité	Direction	Khi2	P-valeur
0	thal	normal	sortie	0.030	0.862
1	restecg	st_t_wave_abnormality	sortie	0.045	0.832
2	slope	up	sortie	0.229	0.633
3	age		sortie	0.302	0.583
4	fbs	t	sortie	0.345	0.557
5	chol		sortie	0.810	0.368
6	restecg	normal	sortie	2.049	0.152
7	thalach		sortie	2.195	0.138
8	trestbps		sortie	2.550	0.110
9	cp	atyp_angina	sortie	3.131	0.077

Figure 20 - Sélection backward - SPAD - Elimination des variables et indicatrices

Par rapport à Tanagra, comme les modalités de références ne sont pas les mêmes pour les variables catégorielles, les indicatrices éliminées peuvent être différentes. Ainsi, à la sortie, nous pouvons obtenir – et c'est le cas pour notre jeu de données – des modèles différents pour un paramétrage (niveau de signification) identique. Dans le cas de sélection de variables donc, avec le traitement individuel des indicatrices, le choix des modalités de référence pèse sur les caractéristiques du modèle final.

Sous SPAD, nous obtenons le modèle suivant (Figure 21) :

Coefficients du modèle par variable

Coefficients de régression estimés par maximum de vraisemblance

Variable	Modalité	Coefficient	Erreur standard	Khi-2 de Wald	P-valeur
sex	male	1.338	0.454	8.693	0.003
cp	non_anginal	-1.716	0.457	14.123	0.000
cp	typ_angina	-1.749	0.616	8.057	0.005
exang	yes	1.150	0.395	8.486	0.004
oldpeak		0.607	0.194	9.762	0.002
slope	flat	1.323	0.392	11.367	0.001
ca		1.284	0.246	27.248	0.000
thal	reversable_defect	1.477	0.380	15.112	0.000
Constante (intercept)		-3.484	0.504	47.793	0.000

Figure 21 - Modèle après sélection backward - SPAD

6.4 Traitement en bloc ou dissocié des indicatrices ?

Fuse alors du fond de la salle la question inévitable : « quelle est alors la bonne méthode, Monsieur ? ». Et je réponds toujours avec un petit air mystérieux « c'est un peu plus compliqué que ça… ». En réalité, la sélection de variables fournit des scénarios de solutions. Il nous appartient par la suite de les valider, soit par l'interprétation, soit en les appliquant sur un échantillon test pour estimer les performances en prédiction[8].

Pour faire un parallèle avec les arbres de décision, je dirais que la suppression de certaines indicatrices des variables catégorielles correspond au mécanisme de fusion des feuilles lors de la segmentation des sommets (méthode CHAID ou CART). Nous regroupons dans une même entité – groupe des « modalités de référence » ici en l'occurrence – les indicatrices qui présentent un comportement identique par rapport à la prédiction de la variable cible.

7 Conclusion

Dans ce tutoriel, nous avons étudié les stratégies de quelques logiciels lors de la sélection de variables en régression logistique, lorsque les explicatives comportent des variables catégorielles que nous exprimons à travers des indicatrices 0/1. Deux approches s'opposent : le traitement en bloc des indicatrices relatives à la même variable, et leur traitement individuel. Elles proposent des résultats différents et, surtout, pour la seconde méthode, la notion de modalité de référence est modifiée parce qu'elle peut désigner après sélection un ensemble de modalités. L'interprétation des coefficients du modèle évolue.

[8] Cela pourrait faire l'objet d'un autre tutoriel d'ailleurs….