1 Topic

Feature selection for categorical predictor variables in logistic regression.

The aim of the logistic regression is to build a model for predicting a binary target attribute from a set of explanatory variables (predictors, independent variables), which are numeric or categorical. They are treated as such when they are numeric. We must recode them when they are categorical. The dummy coding is undeniably the most popular approach in this context.

We take a simple example to clarify ideas. We have a variable X with 3 levels {A, B, C}. We create 2 dichotomous variables: X_A which takes the value 1 when X = A, 0 otherwise; X_B which takes the value 1 when X = B, 0 otherwise. We identify the case X = C when both X_A and X_B take the value 0. “C” is called the reference group. The choice of the reference group influences the interpretation of the regression coefficients. But it has no influence of the performance of the model i.e. regardless of the chosen reference group, the error rate of the model - for instance - is not modified.

The situation becomes more complicated when we perform a feature selection. The idea is to determine the predictors that contribute significantly to the explanation of the target attribute. There is no problem when we consider a numeric variable. It is either excluded or either kept in the model. But how to proceed when we handle a categorical explanatory variable? Should we treat the dichotomous variables associated to a categorical predictor as a whole that we must exclude or include into the model? Or should we treat the each dichotomous variable independently? How to interpret the coefficients of the selected dichotomous variables in this case?

In this tutorial, we study the approaches proposed by various tools: **R 3.1.2**, **SAS 9.3**, **Tanagra 1.4.50** and **SPAD 8.0**. We will see that feature selection algorithms rely on specific criteria according to the software. We will see also that they use different approaches when we are in the presence of the categorical predictor variables.

2 Dataset

We use the “heart-disease” dataset from the UCI Repository[1]. We transform the target attribute NUM into a binary variable with two levels {absence, presence}.

The predictors are numeric or categorical. Here are their descriptions:

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

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 - List of the predictor variables

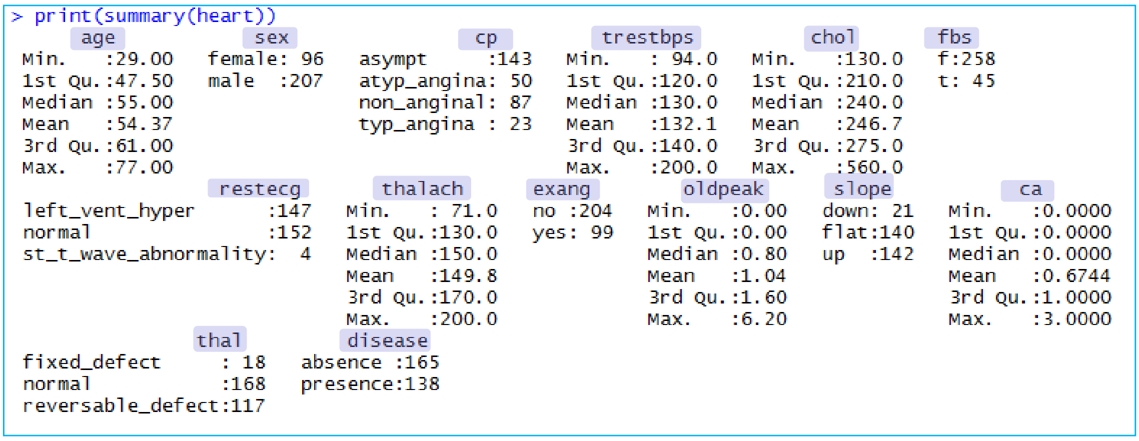
We can observe the levels of each categorical variable.

3 Logistic regression and variable selection in R

We import the « heart-c.xlsx » data file using the “xlsx” package.

```
#clear the memory
rm(list=ls())
#modify the default directory
setwd("…")
#load the data file
library(xlsx)
heart <- read.xlsx(file = "heart-c.xlsx",sheetIndex = 1,header = T)
print(summary(heart))
```

We obtain the following summary. “Disease” is the target attribute.



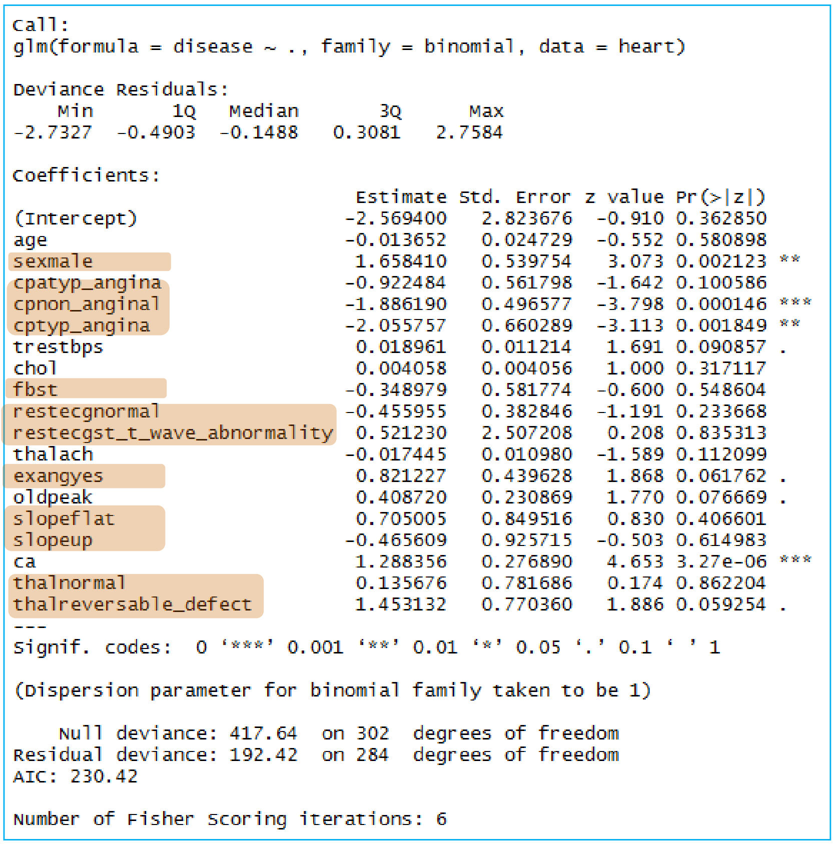
```
> print(summary(heart))
      age            sex                cp         trestbps          chol       fbs
 Min.   :29.00   female: 96   asympt      :143   Min.   : 94.0   Min.   :130.0   f:258
 1st Qu.:47.50   male  :207   atyp_angina: 50   1st Qu.:120.0   1st Qu.:210.0   t: 45
 Median :55.00                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 Regression with all the predictors

We perform the logistic regression with the glm() procedure (“stats” package). The tool can handle both numeric and categorical (R data type **factor**) predictors.

```
#regression disease vs. all the other variables of the data frame
lr.all <- glm(disease ~ ., data = heart, family = binomial)
#display the results
print(summary(lr.all))
```

We get the following classifier, the Akaike criterion equals 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 - Classifier with all the candidate variables - glm() under R

We observe that the categorical predictors are automatically coded. R chooses the first level (in the alphabetic order) as the reference group. Thus, for the categorical variables of our study, the reference groups are: SEX = FEMALE, CP = ASYMPT, FBS = F, RESTECG = LEFT_VENT_HYPER, EXANG = NO, SLOPE = DOWN, THAL = FIXED_DEFECT.

This choice is important for the interpretation of the results. For instance, we see that the coefficient for the dummy variable (SEX = MALE) is significant at the 1% level i.e. the coefficient is significantly different from 0. We observe also that its sign is positive. It means that, compared to the women (reference group), the men have more chance to be sick (disease = presence). To the contrary, if we have defined SEX = MALE as the reference group, we would get the same coefficient in absolute value, but with the opposite sign. It means that, compared to the men, the women have less chance to be sick.

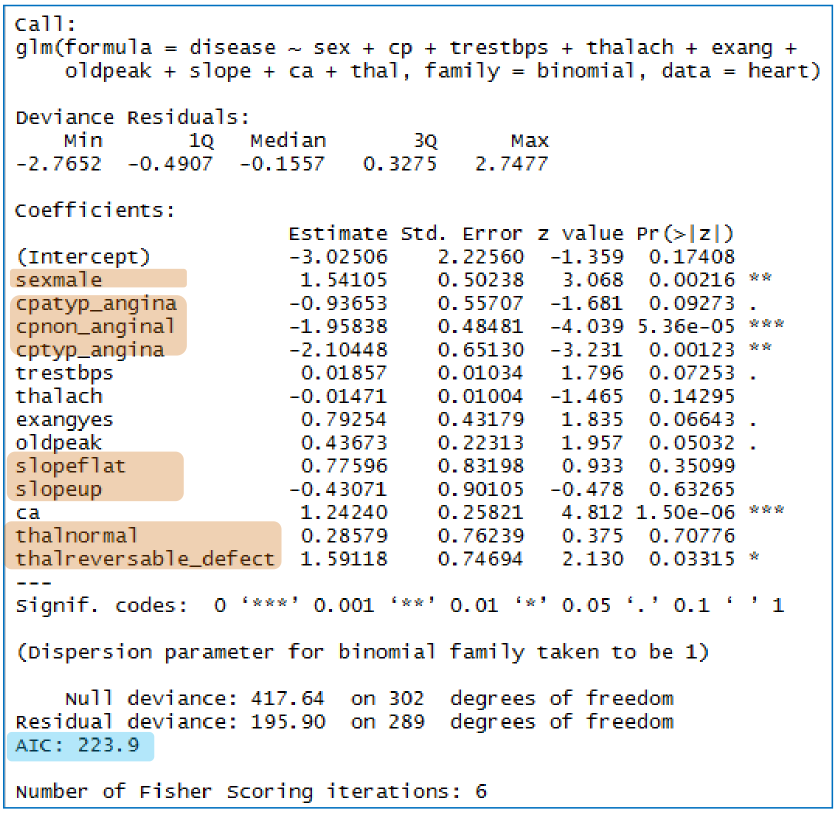
The choice of mode of reference has no influence on the prediction of the classifier. Whatever the reference chosen for each categorical predictor variable, we will get exactly the same prediction when we classify a new instance.

3.2 Backward selection - AIC criterion

In the backward elimination process, we start with all candidate variables. We test the deletion of each variable and measure the AIC criterion. We remove the variable which minimizes the AIC. We continue as long as the AIC decreases i.e. we stop when the deletion of the variable leads to an increase of the AIC criterion.

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

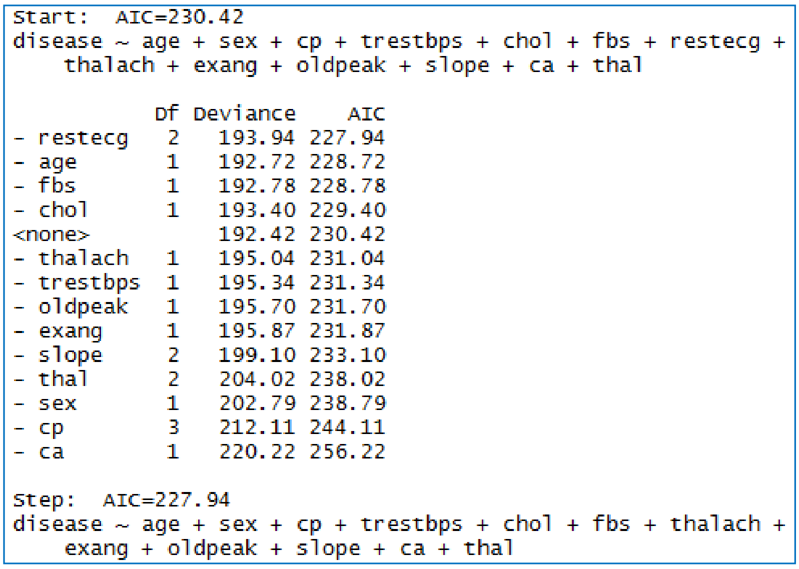
We obtain the following classifier:



```
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 - Classifier after the backward selection - stepAIC() under R

Compared with the regression incorporating all the variables (AIC = 230.42, Figure 2), the new classifier is better with a lowest AIC (AIC = 223.9, Figure 3). We note especially, and this is what interests us in this tutorial, that the dichotomous variables coming from the categorical predictors with more than 2 levels are treated as a whole: either they are all excluded (CP, RESTEG), or retained (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 - First step of the backward elimination process - StepAIC() under R

This behavior appears clearly when we inspect the details of the calculations. In the first step (Figure 4), stepAIC seeks to exclude the worst variable: RESTECG minimizes the AIC criterion (227.94). The two corresponding dichotomous variables are excluded at the same time as evidenced by the degree of freedom (DF - "degree of freedom" column, DF = 2 to RESTECG) in the table showing the intermediate results.

Conclusion: This behavior seems consistent with the initial specifications. We attempt to determine the best subset of variables for the regression. When a variable is expressed through a set of indicators, we cannot ungroup them.

We will see below that this reasoning can be called into question because, beyond the selection of variables, we are looking to build the most performing model. The possibility to deal individually with the indicators of categorical variables provides an additional freedom in searching for solutions, and leads to perhaps more efficient models. The challenge is to correctly interpret the classifier obtained.

4 Regression and variable selection in SAS

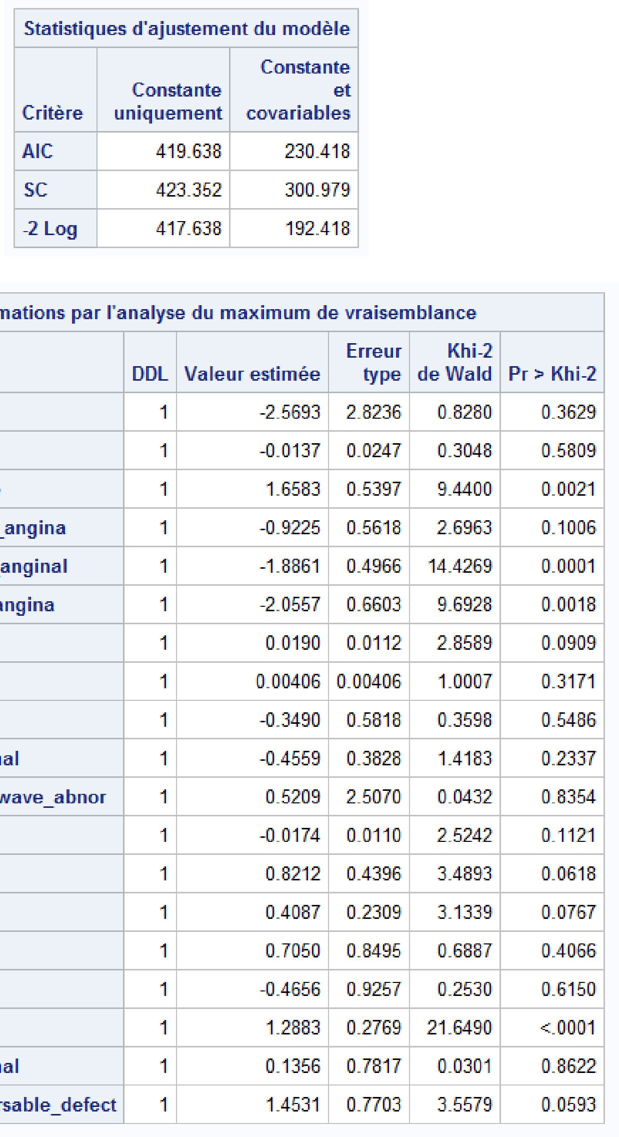
We analyze the variable selection in SAS. We go the heart of the matter. For a deeper description of the PROC LOGISTIC, please refer to a previous article ("Introduction to SAS proc logistic", July 2012).

4.1 Regression with all the variables

The HEART dataset is stored in the DATAREG bank that we created. We set the following commands to perform a logistic regression.

```
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;
```

The parameter CLASS indicates the categorical predictors. PARAM specifies the type of encoding to use and REF the reference group.



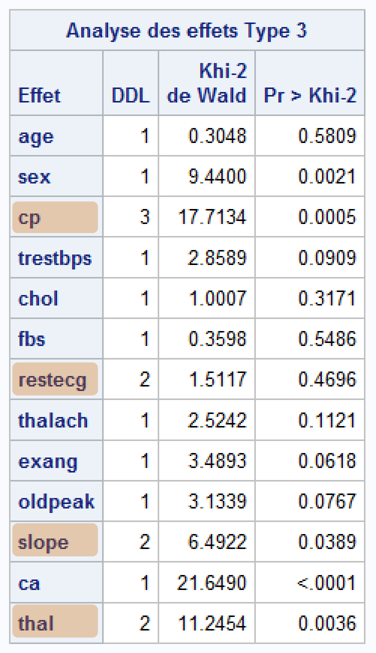
Statistiques d'ajustement du modèle		
Critère	Constante uniquement	Constante et covariables
AIC	419.638	230.418
SC	423.352	300.979
-2 Log	417.638	192.418

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 - Logistic regression in SAS

In our example, PARAM = REFERENCE recommends the dummy coding; with REF = FIRST, we specify that the first value (in alphabetical order) will correspond to the reference group.

Thus we have the default settings of R (section 3.1). The obtained classifiers are identical (Figure 5, AIC = 230.418). But SAS does not forget that some variables are indicators coming from a categorical predictor. It proposes a table where it checks the relevance of the variables (Figure 6).



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 - Type 3 - Analysis of effects - SAS

When the predictor variable is numeric or binary, the degrees of freedom, the test statistic and p-value are identical to those of the previous table (Figure 5). By contrast, when the variable is categorical with more than 2 levels, SAS performs a test for the simultaneous nullity of all the coefficients associated to its dichotomous variables.

Let us study the variable THAL à the 1% level. Apparently, in comparison with the reference group THAL = FIXED DEFECT, neither THAL = NORMAL, nor THAL = REVERSABLE DEFECT, individually seems not to lead a significant increase or decrease of the chance to be sick (Figure 5). But when we perform the global test for the variable i.e. we test if the coefficients are simultaneously zero, we reject the null hypothesis (Figure 6, p-value = 0.0036).

In fact, we observe that treating the indicators individually or as a whole related to categorical variable correspond to different behaviors.

4.2 Variable selection in SAS

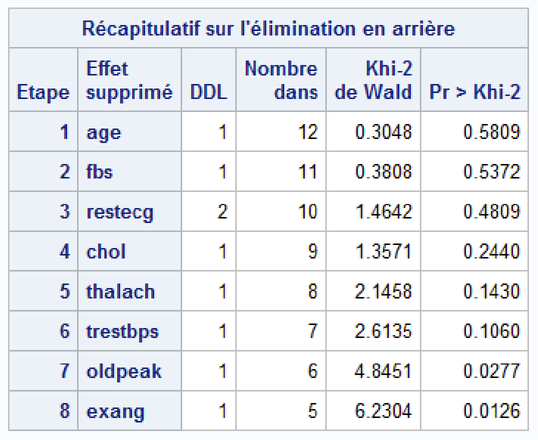
The backward selection is based on Wald test in SAS. Like R, the indicators coming from a categorical variable are processed as a group. The calculations are based on the type 3 analysis of effects (Figure 6) table. At each step SAS removes the less relevant variable.

The variable selection process works as follows: (1) SAS detects the least relevant variable, one that provides the highest p-value; (2) the variable is removed from the model if the p-value is greater than a threshold specified by the SLSTAY parameter. (3) SAS fits again the model on the dataset with the remaining variables, and then performs the steps (1) and (2) until all of the variables are relevant.

For our dataset, according to the table above (Figure 6), it would be AGE that would be eliminated at the first iteration with a p-value equal to 0.5809, superior to the parameter SLSTAY = 0.01. Here is the SAS code for the backward selection using the "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;
```

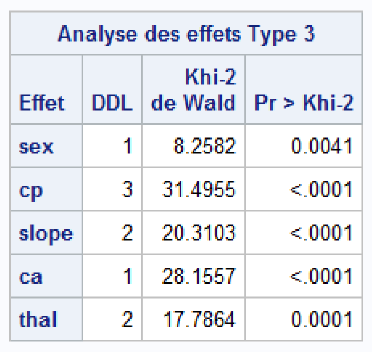
SAS describes in a table the variables removed at each step (Figure 7).



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

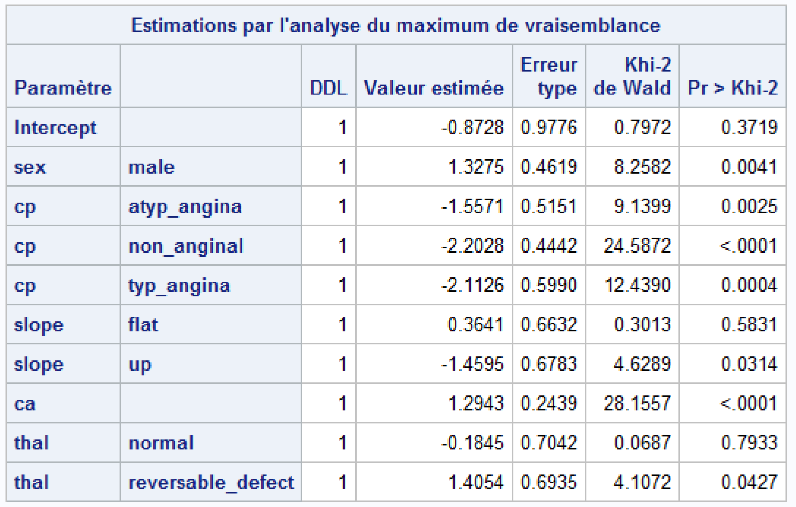
Figure 7 - Backward selection process in SAS

Finally, we obtain a classifier with 5 predictor variables: SEX, CP, SLOPE, CA and THAL. All the variables are statistically relevant at the 1% level (Figure 8). It cannot be otherwise. This is the consequence of this type of selection process.



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 - Type 3 - Analysis of effects at the end of the selection - SAS



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 - Coefficients of the model at the end of the selection - SAS

But some indicators may be non-significant (at the same significance level) into the table of coefficients (e.g. the indicators of SLOPE and THAL) (Figure 9). This is not inconsistent. SAS uses the Type 3 Analysis of effects to detect the variables to remove. We have seen above that an indicator may appear non-significant while the corresponding variable is significant (Figure 5 and Figure 6).

Note: About the AIC criterion, that of SAS is higher than that of R (323.167 vs. 223.9; the higher is the AIC, the worst is the model). This is not surprising. SAS does not use explicitly this criterion in the detection of the best model.

5 Regression in Tanagra

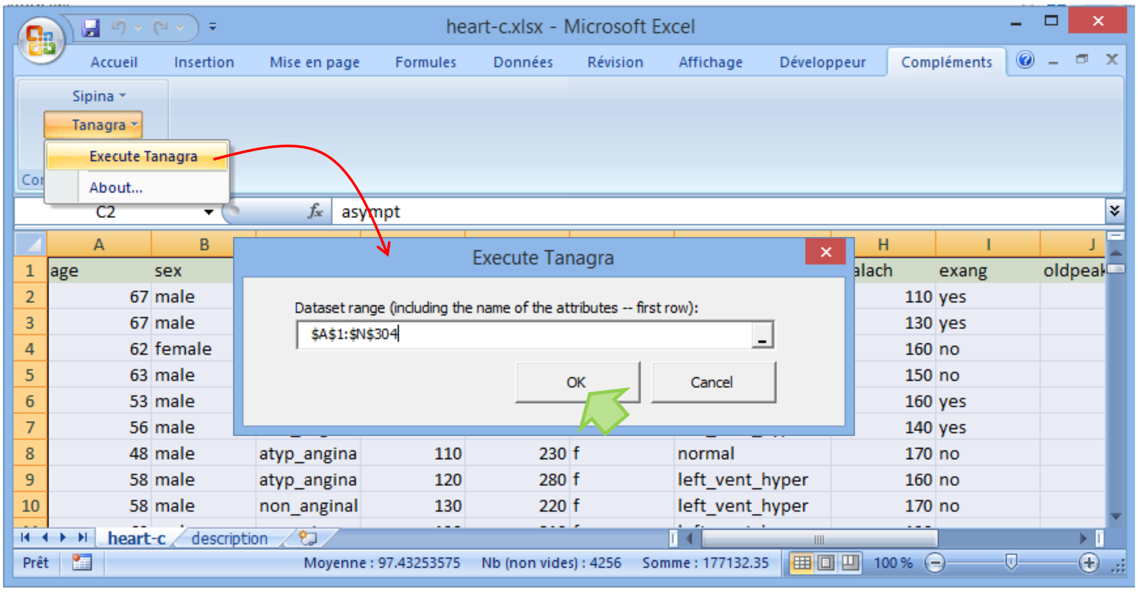
With Tanagra, we must recode explicitly the categorical predictors before performing a logistic regression[2]. There are several reasons for this: It seemed to me pedagogically desirable that students perform explicitly the coding for understanding the necessity and the influence of this step in regression; the user can introduce further analysis by using other types of coding (see the CLASS parameter into SAS); Tanagra relies on a specific strategy in variable selection process, it deals with the indicators as independent variables, the resulting classifier is of a different nature compared to previous approaches.

5.1 Regression with all the variables

5.1.1 Import the dataset

We load the file “heart-c.xlsx” into the Excel. We select the data range that we send to Tanagra using the **tanagra.xla** add-in.

[2] Tanagra Tutorial, « Dummy coding for categorical predictor variables », March 2016.



Execute Tanagra

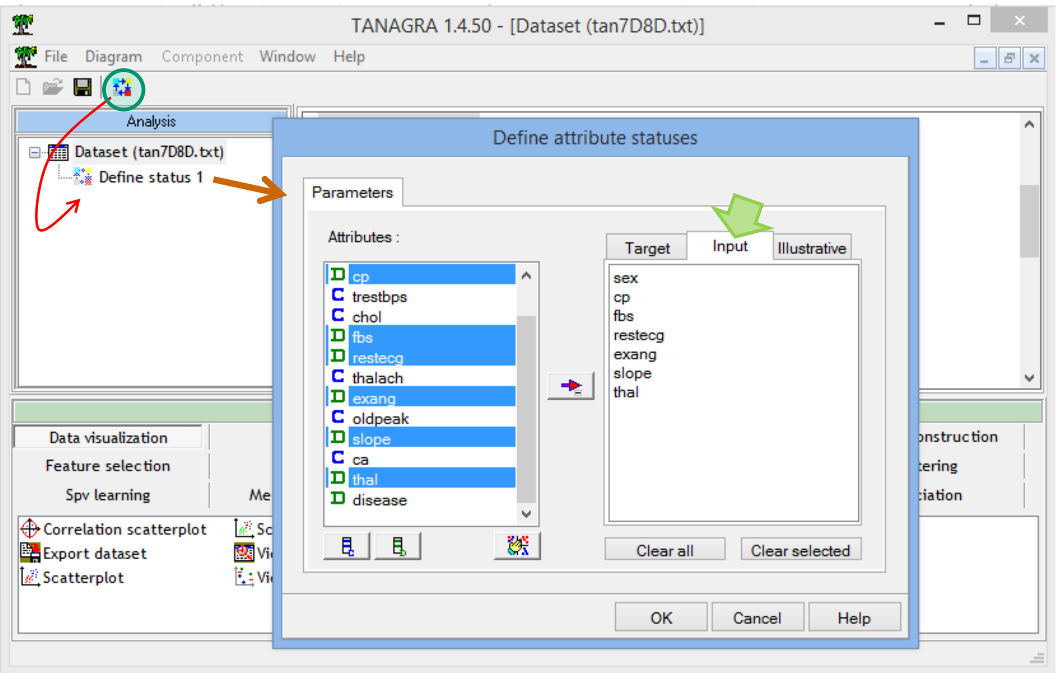
Dataset range (including the name of the attributes -- first row): $A$1:$N$304

	A	B						H	I	J
1	age	sex						alach	exang	oldpeak
2	67	male						110	yes	
3	67	male						130	yes	
4	62	female						160	no	
5	63	male						150	no	
6	53	male						160	yes	
7	56	male						140	yes	
8	48	male	atyp_angina	110	230	f	normal	170	no	
9	58	male	atyp_angina	120	280	f	left_vent_hyper	160	no	
10	58	male	non_anginal	130	220	f	left_vent_hyper	170	no	

Figure 10 - Send the dataset from Excel to Tanagra[3]

5.1.2 Dummy coding of categorical predictor variables

Tanagra is automatically launched. We have 14 variables and 303 instances. In order to code the categorical predictor variables into indicators, we must first to specify the variables to recode using the DEFINE STATUS component available into the toolbar.



Define attribute statuses

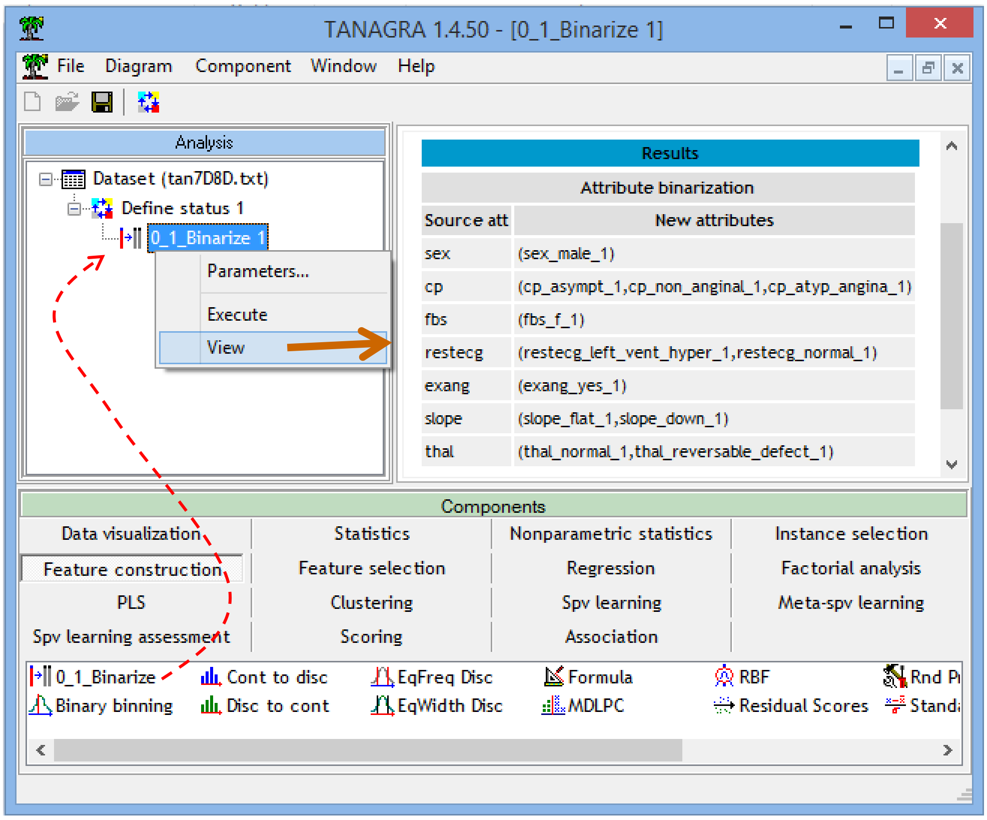
Attributes : cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, disease

Input: sex, cp, fbs, restecg, exang, slope, thal

We set as INPUT all the categorical variables except DISEASE which is the target attribute. Then we insert the 0_1_BINARIZE (tab FEATURE CONSTRUCTION) component into the diagram. It performs the coding process. We click on the VIEW pop-up menu to obtain the

[3] Tanagra Tutorial, “Tanagra add-in for Office 2007 and Office 2010”, August 2010.

results. We can visualize the set of indicators for each variable. The omitted value corresponds to the reference group.

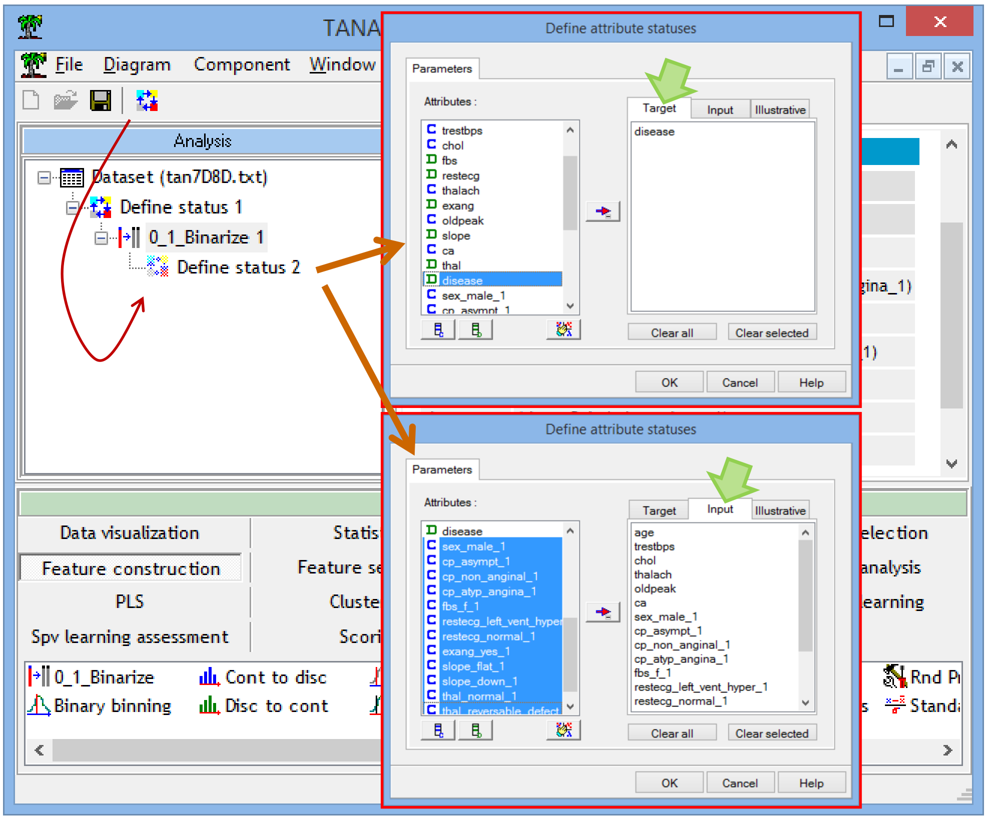


Results

Attribute binarization

Source att	New attributes
sex	(sex_male_1)
cp	(cp_asympt_1,cp_non_anginal_1,cp_atyp_angina_1)
fbs	(fbs_f_1)
restecg	(restecg_left_vent_hyper_1,restecg_normal_1)
exang	(exang_yes_1)
slope	(slope_flat_1,slope_down_1)
thal	(thal_normal_1,thal_reversable_defect_1)

5.1.3 Logistic regression



Define attribute statuses

Attributes : trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, disease, sex_male_1, cp_asympt_1

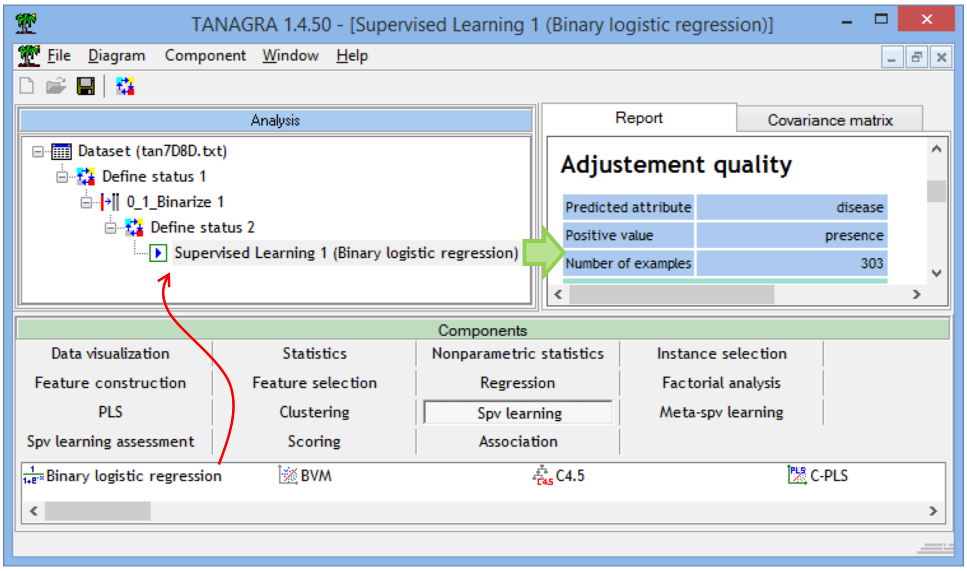
Target: disease

Define attribute statuses

Attributes : disease, sex_male_1, cp_asympt_1, cp_non_anginal_1, cp_atyp_angina_1, fbs_f_1, restecg_left_vent_hyper, restecg_normal_1, exang_yes_1, slope_flat_1, slope_down_1, thal_normal_1, thal_reversable_defect

Input: age, trestbps, chol, thalach, oldpeak, ca, sex_male_1, cp_asympt_1, cp_non_anginal_1, cp_atyp_angina_1, fbs_f_1, restecg_left_vent_hyper_1, restecg_normal_1

We insert again the DEFINE STATUS component to define the target attribute (TARGET = DISEASE) and the input ones (INPUT = all the numeric variables + the indicators from the categorical predictors). Then we add the BINARY LOGISTIC REGRESSION (SPV LEARNING tab) component. We click on the VIEW pop-up menu to get the results of the regression.



Adjustement quality

Predicted attribute	disease
Positive value	presence
Number of examples	303

We have exactly the same model that R and SAS as attested by the deviance (-2LL) i.e. applied on the same individuals, these classifiers will produce the same estimated class membership posterior probabilities and, a fortiori, the same predicted values. But, because the reference groups are not the same for the categorical predictors, the set of estimated coefficients are different.

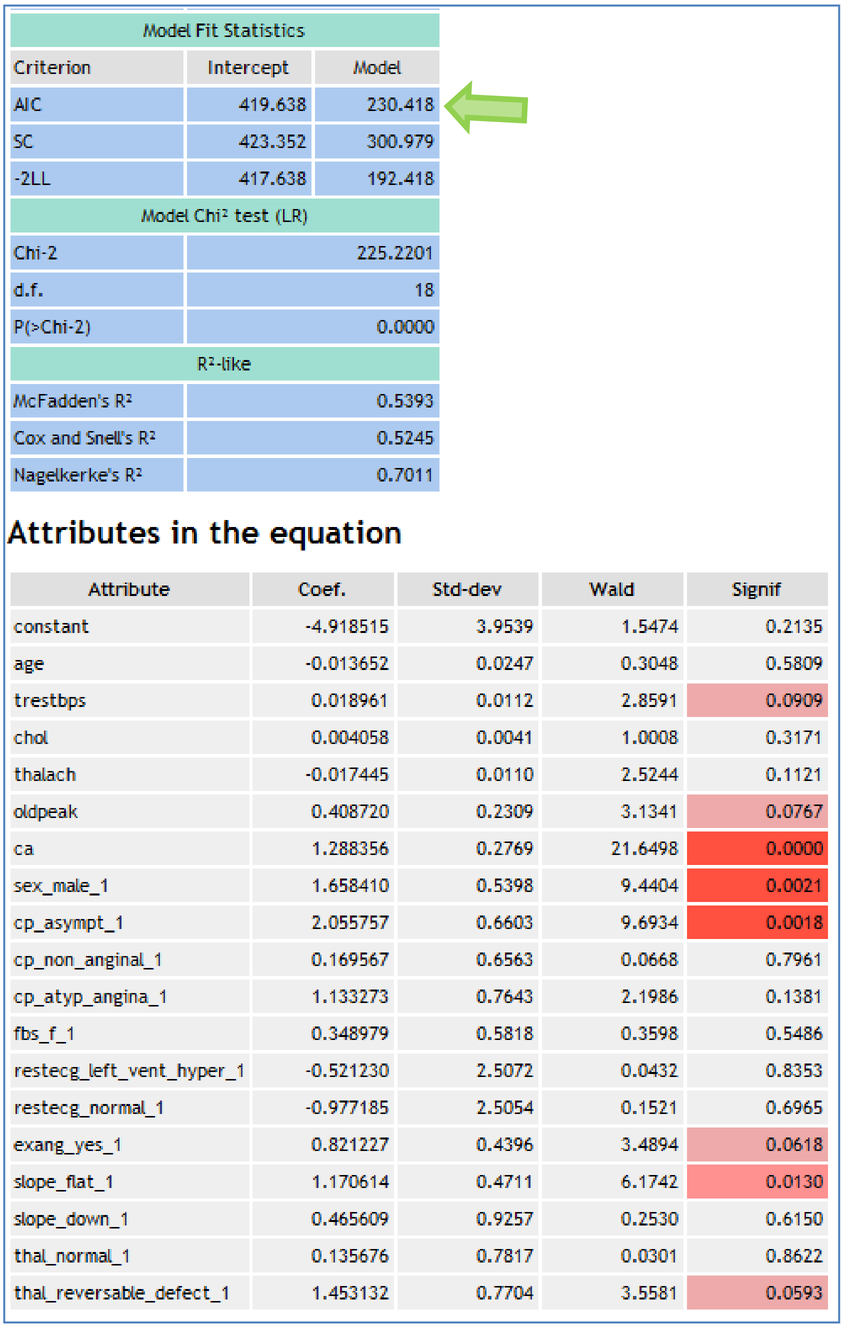
Let us compare the results of R (Figure 2) and Tanagra (Figure 11) for some categorical predictors:

- For the variable SEX, FEMALE is the reference group for the two tools R and Tanagra, the regression coefficient is the same 1.658410.
- For FS, R has selected the "F" modality as reference, Tanagra used "T", the coefficients are not the same, respectively - 0.348979 and 0.348979. Because FS is a binary variable, the coefficients have the same absolute value, but they are of opposite sign.
- For SLOPE which is not binary (SLOPE has 3 levels), R chooses "DOWN" while Tanagra used "UP" as reference group. The coefficients of R are $a_{FLAT/DOWN}$ = 0.705005 and $a_{UP/DOWN}$ = -0.465609; the coefficients for Tanagra are $a_{FLAT/UP}$ = 1.170614 and $a_{DOWN/UP}$ = 0.465609. We note that it is very easy to find the coefficients of R using those of Tanagra, indeed:

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

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

Consequently, regardless of the reference group used in the coding of the dummy variables, the behavior of the classifier and the interpretations of the coefficients are preserved.



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

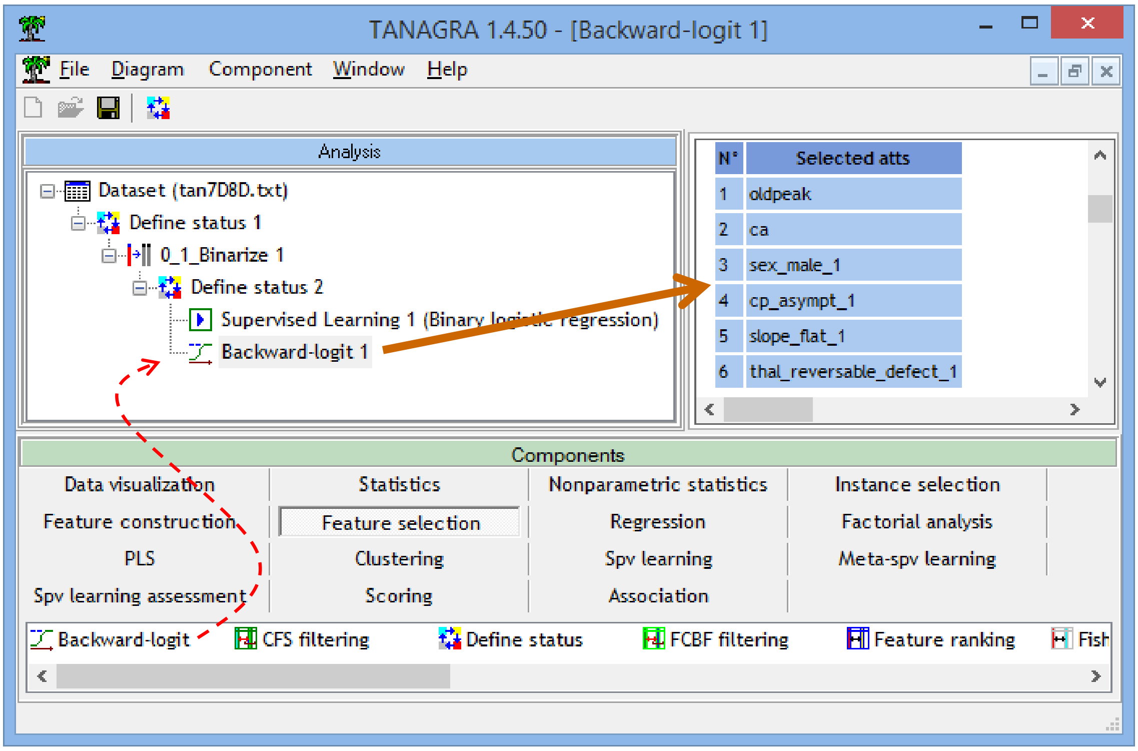
Attributes in the equation

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 - Regression with all the variables (numeric + indicators) - Tanagra

5.2 Backward selection in Tanagra

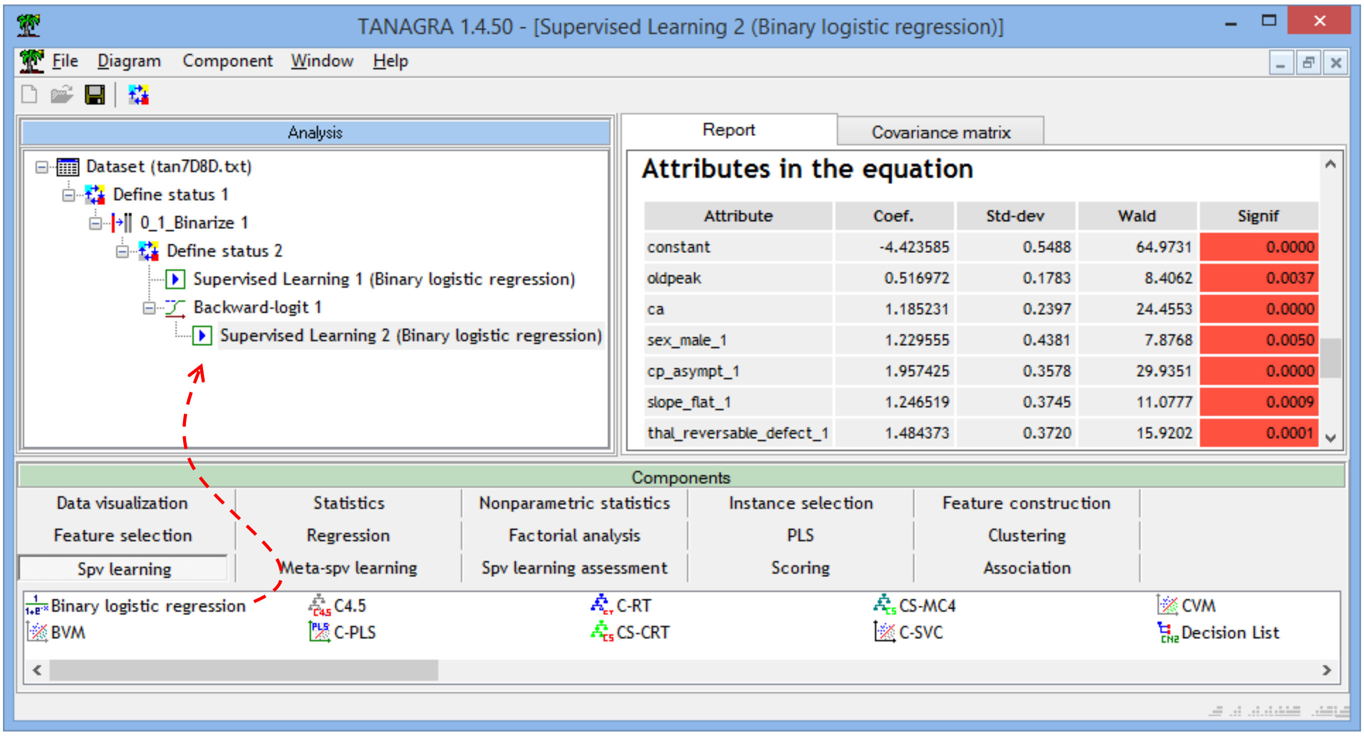
Tanagra has really a different behavior in the variable selection process. Indeed, the tool for the variable selection deals with the indicators without taking account of their relationship with the original variables. We add the BACKWARD-LOGIT component (FEATURE SELECTION) into the diagram. We insert the tool after the DEFINE STATUS 2 where we specified the candidate explanatory variables for the regression. The default level of significance is 1%.



N°	Selected atts
1	oldpeak
2	ca
3	sex_male_1
4	cp_asympt_1
5	slope_flat_1
6	thal_reversable_defect_1

Figure 12 - Bacwkard-Logit tool for the variable selection - Tanagra

We observe that 6 numeric or indicator variables are selected. We observe also that some indicators coming from the same variable are treated separately (e.g. CP, SLOPE, THAL).



Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-4.423585	0.5488	64.9731	0.0000
oldpeak	0.516972	0.1783	8.4062	0.0037
ca	1.185231	0.2397	24.4553	0.0000
sex_male_1	1.229555	0.4381	7.8768	0.0050
cp_asympt_1	1.957425	0.3578	29.9351	0.0000
slope_flat_1	1.246519	0.3745	11.0777	0.0009
thal_reversable_defect_1	1.484373	0.3720	15.9202	0.0001

Figure 13 - Regression on the selected variables - Tanagra

We perform the regression on the selected variables using the BINARY LOGISTIC REGRESSION component.

The AIC of the classifier equals 22.978 (not visible into the screenshot). Unlike SAS (Figure 9), all numeric and indicator variables are significant at the 1% level (Figure 13). It does not mean that this solution is better. This is simply the consequence of the selection strategy used by the tool that processes the indicator variables individually.

Note: Does this model make sense? The answer is yes. The missing levels become the reference group. We consider the variable SLOPE to support our argument. We have seen in the regression involving all indicators that coefficient $a_{DOWN/UP}$ = 0.465609 was not significant with a p-value equals 0.6150 (Figure 11). By removing the DOWN indicator variable, the system tells us that it makes sense to merge the terms DOWN and UP, as decision trees algorithms that perform merges of leaves during splitting process (e.g. CART or CHAID). Thus, the coefficient $a_{FLAT/\{UP,DOWN\}}$ = 1.246519 indicates the increase in chance to be sick associated to FLAT compared to the union of {UP, DOWN} levels.

6 Regression in SPAD - Comparison of the approaches

SPAD is a software package which provides tools for the whole data mining process[4]. We have study it in some circumstances previously (e.g. interactive decision tree learning). It works just like the other popular data mining / data science tools such as IBM SPSS Modeler or SAS Enterprise Miner.

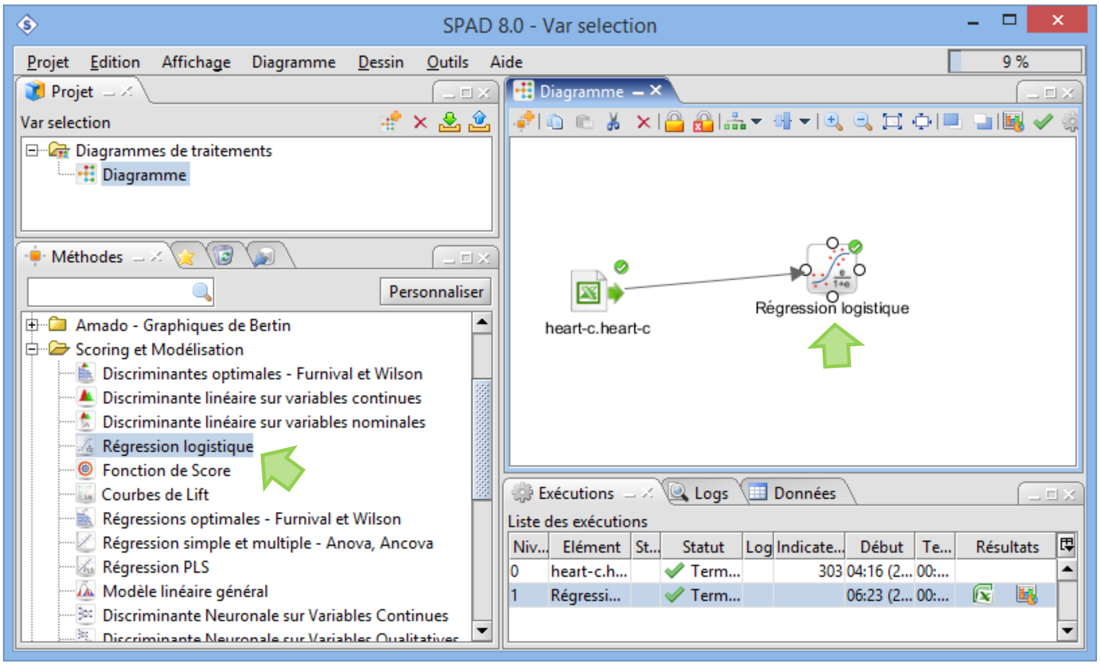
SPAD incorporates the two approaches highlighted above. It can achieve the backward selection based on the Wald test by processing as a group the indicators associated to categorical predictors, like SAS; it can also treat them individually, like Tanagra. In this section, we will see the settings to be specified to get the behavior that we want.

6.1 Logistic regression in SPAD

We must create a new project to start an analysis process. We use the FEUILLE EXCEL to import the data file. The operation is simple. We must nevertheless ensure that the CA column is encoded as real type. We set this in the METADONNEES tab of the dialog settings of the tool.

Then, we insert the REGRESSION LOGISTIQUE component (branch SCORING and MODELISATION), to which we link the Excel source. We click on the PARAMETRES menu to access the settings dialog box.

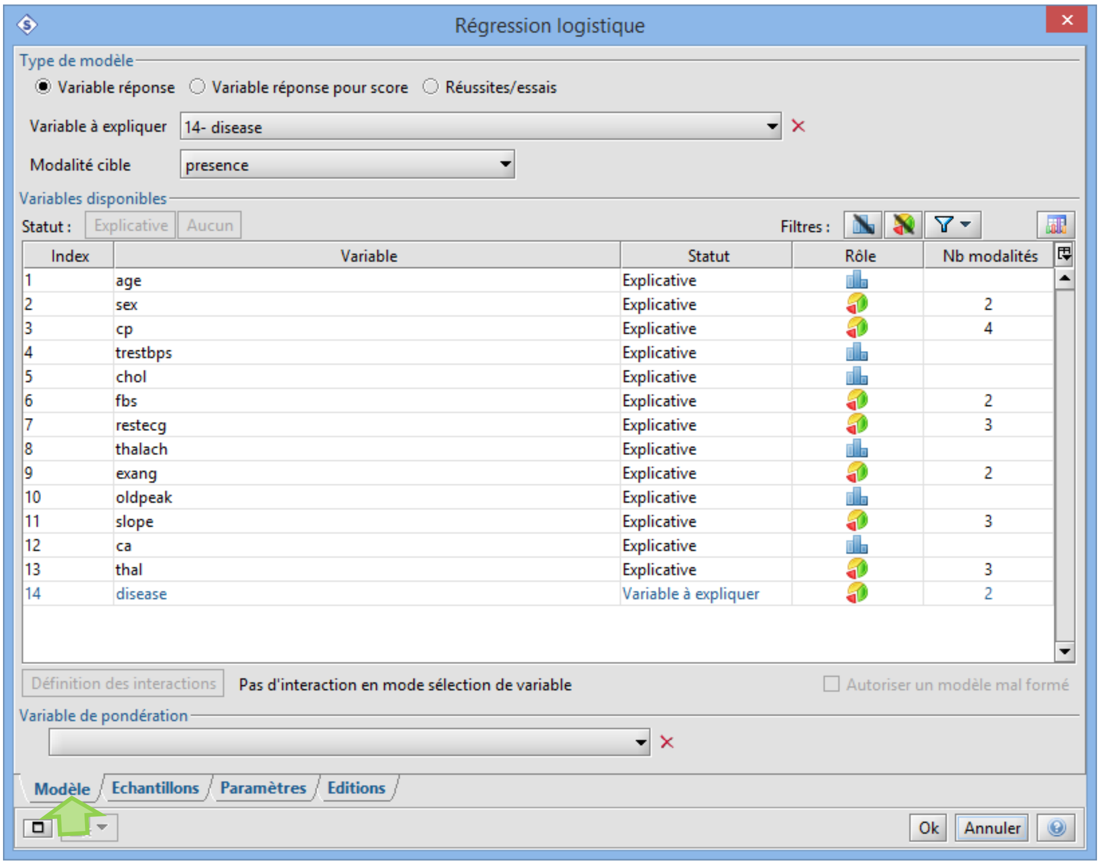
[4] COHERIS SPAD - http://www.coheris.com/produits/analytics/logiciel-data-mining/


SPAD 8.0 - Var selection
Projet
Edition
Affichage
Diagramme
Dessin
Outils
Aide
9 %
Projet
Var selection
Diagrammes de traitements
Diagramme
Méthodes
Personnaliser
Amado - Graphiques de Bertin
Scoring et Modélisation
Discriminantes optimales - Furnival et Wilson
Discriminante linéaire sur variables continues
Discriminante linéaire sur variables nominales
Régression logistique
Fonction de Score
Courbes de Lift
Régressions optimales - Furnival et Wilson
Régression simple et multiple - Anova, Ancova
Régression PLS
Modèle linéaire général
Discriminante Neuronale sur Variables Continues
Diagramme
heart-c.heart-c
Régression logistique
Exécutions
Logs
Données
Liste des exécutions

Niv...	Elément	St...	Statut	Log	Indicate...	Début	Te...	Résultats
0	heart-c.h...		✓ Term...		303	04:16 (2...	00:...	
1	Régressi...		✓ Term...			06:23 (2...	00:...	

Figure 14 - Diagram “Logistic Regression” in SPAD

Into the tab “MODELE”, we define the target attribute (DISEASE) and the predictors.



● Variable réponse ○ Variable réponse pour score ○ Réussites/essais

Variable à expliquer: 14- disease

Modalité cible: presence

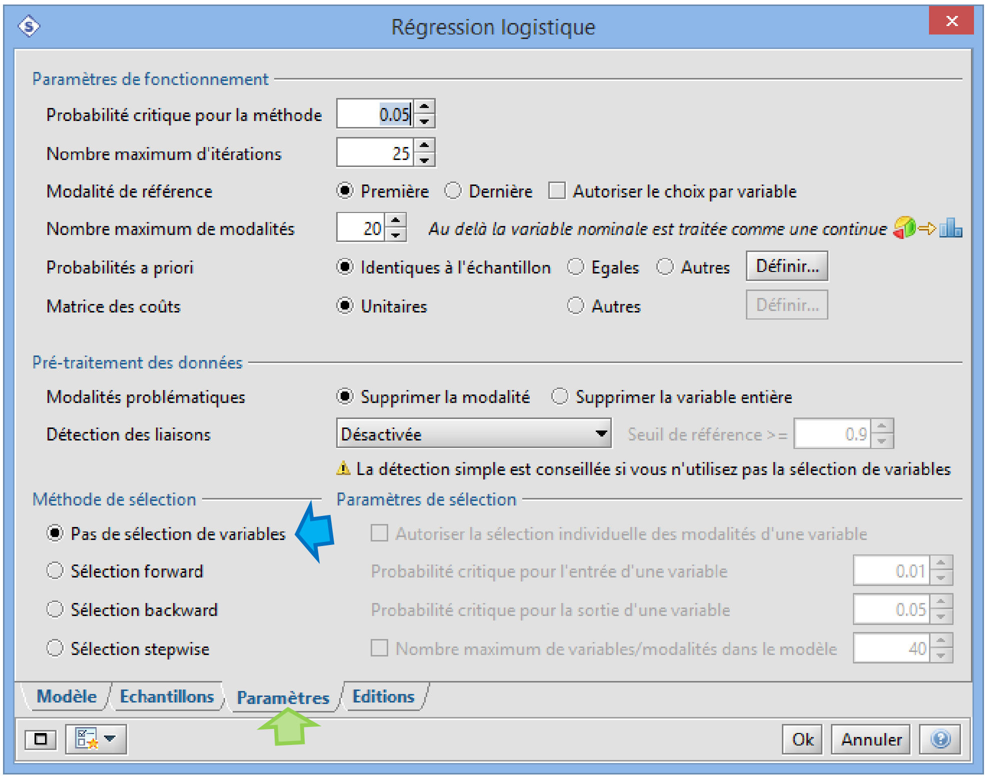
Index	Variable	Statut	Rôle	Nb modalités
1	age	Explicative		
2	sex	Explicative		2
3	cp	Explicative		4
4	trestbps	Explicative		
5	chol	Explicative		
6	fbs	Explicative		2
7	restecg	Explicative		3
8	thalach	Explicative		
9	exang	Explicative		2
10	oldpeak	Explicative		
11	slope	Explicative		3
12	ca	Explicative		
13	thal	Explicative		3
14	disease	Variable à expliquer		2

Pas d'interaction en mode sélection de variable

□ Autoriser un modèle mal formé

Figure 15 - Role of the variables into the analysis - SPAD

Into the tab “PARAMETRES”, we specify that we include all the candidate predictors into the model (*METHODE DE SELECTION = PAS DE SELECTION DE VARIABLES*) (Figure 16).



Régression logistique

Paramètres de fonctionnement

Probabilité critique pour la méthode: 0.05

Nombre maximum d'itérations: 25

Modalité de référence: ● Première ○ Dernière □ Autoriser le choix par variable

Nombre maximum de modalités: 20 *Au delà la variable nominale est traitée comme une continue*

Probabilités a priori: ● Identiques à l'échantillon ○ Egales ○ Autres Définir...

Matrice des coûts: ● Unitaires ○ Autres Définir...

Pré-traitement des données

Modalités problématiques: ● Supprimer la modalité ○ Supprimer la variable entière

Détection des liaisons: Désactivée Seuil de référence >= 0.9

⚠ La détection simple est conseillée si vous n'utilisez pas la sélection de variables

Méthode de sélection

● Pas de sélection de variables

○ Sélection forward

○ Sélection backward

○ Sélection stepwise

Paramètres de sélection

□ Autoriser la sélection individuelle des modalités d'une variable

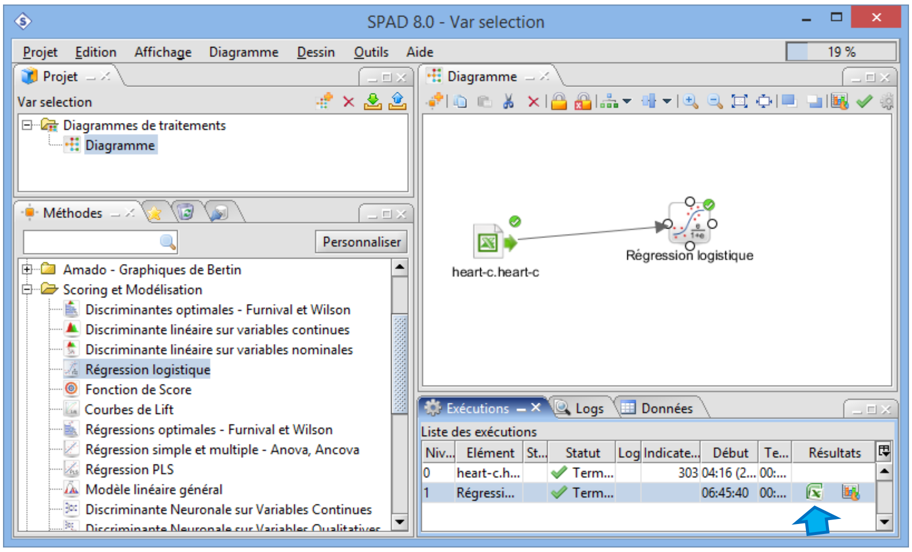
Probabilité critique pour l'entrée d'une variable: 0.01

Probabilité critique pour la sortie d'une variable: 0.05

□ Nombre maximum de variables/modalités dans le modèle: 40

Figure 16 - Settings - No variable selection - SPAD

We click on the OK button. The calculations are automatically launched. We can visualize the results by clicking on the Excel icon in the "Resultats" section (Figure 17).



SPAD 8.0 - Var selection

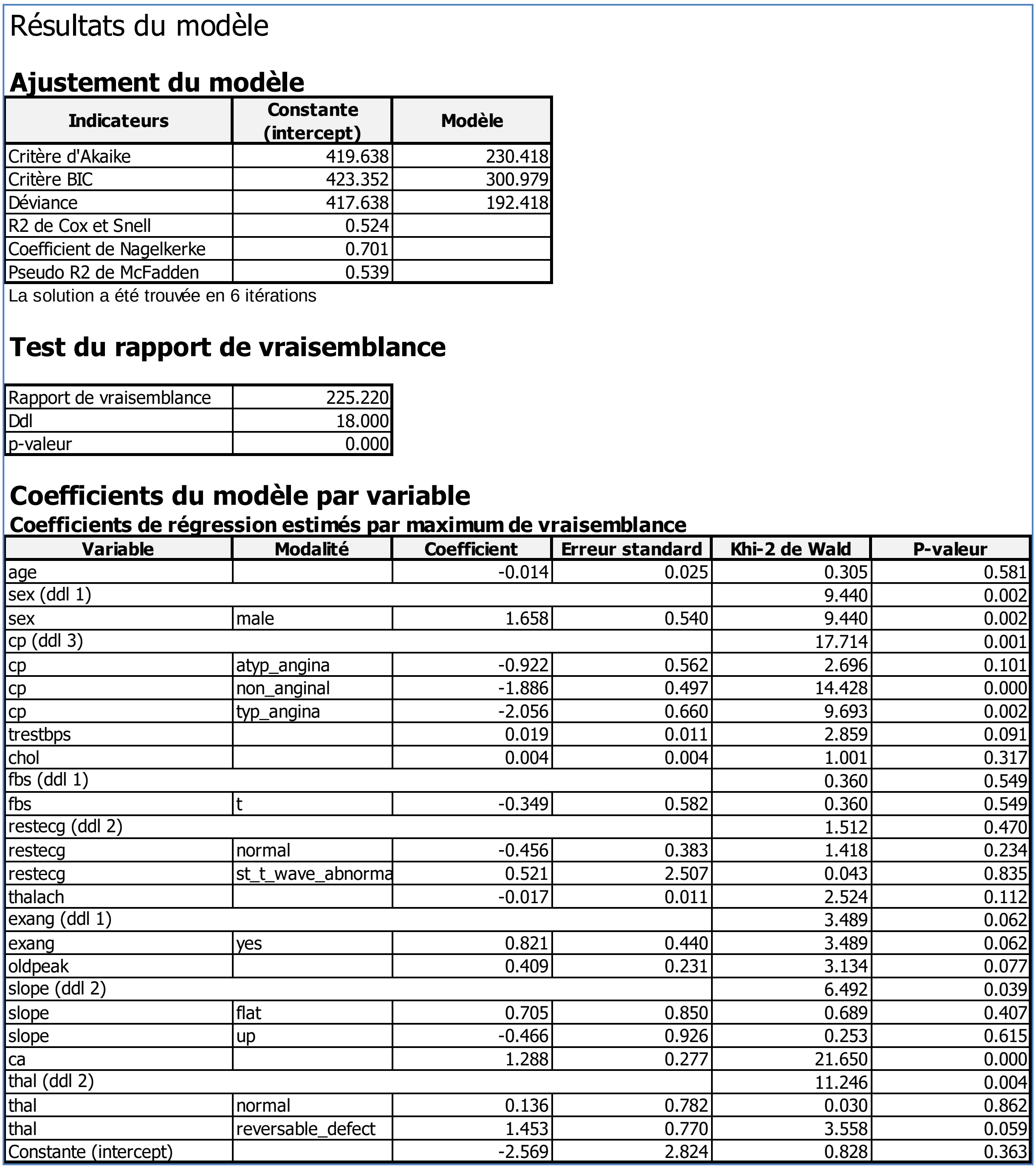
Projet
Edition
Affichage
Diagramme
Dessin
Outils
Aide
19 %
Projet
Var selection
Diagrammes de traitements
Diagramme
Méthodes
Personnaliser
Amado - Graphiques de Bertin
Scoring et Modélisation
Discriminantes optimales - Furnival et Wilson
Discriminante linéaire sur variables continues
Discriminante linéaire sur variables nominales
Régression logistique
Fonction de Score
Courbes de Lift
Régressions optimales - Furnival et Wilson
Régression simple et multiple - Anova, Ancova
Régression PLS
Modèle linéaire général
Discriminante Neuronale sur Variables Continues
Discriminante Neuronale sur Variables Qualitatives
Diagramme
heart-c.heart-c
Régression logistique
Exécutions
Logs
Données

Liste des exécutions

Niv...	Elément	St...	Statut	Log	Indicate...	Début	Te...	Résultats
0	heart-c.h...		✓ Term...		303	04:16 (2...	00:...	
1	Régressi...		✓ Term...			06:45:40	00:...	

Figure 17- Accessing to the results - SPAD

The outputs are splitted in several sheets. Into "REG_MODEL", we can visualize the overall evaluation of the model, and the table with the estimated coefficients of the classifier (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

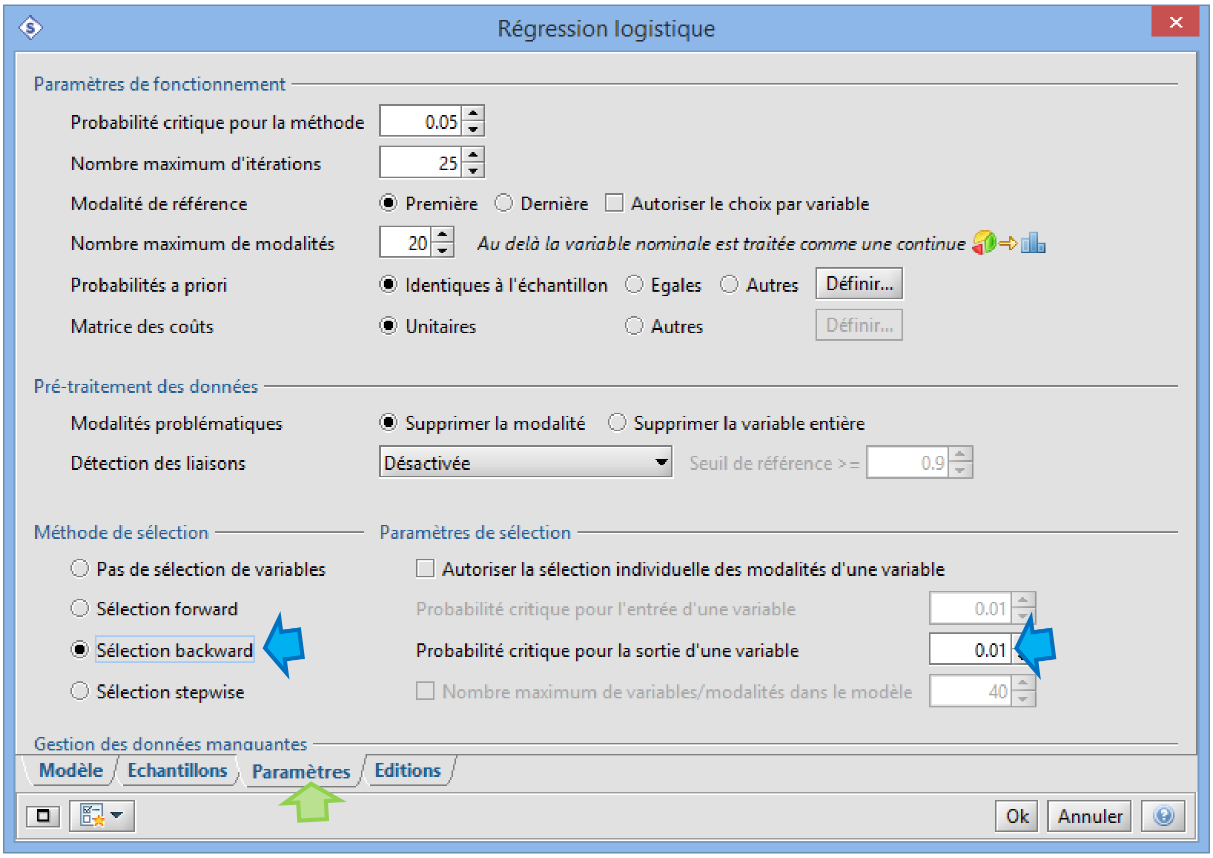
Figure 18 - Description of the classifier - SPAD

SPAD mixes the table of coefficients and the type 3 analysis of effects. Thus, we can identify at a glance the relevant indicators and the relevant variables.

For instance, at the 1% level, the two indicators from THAL are not relevant (THAL = NORMAL, p-value = 0.862; THAL = REVERSABLE_DEFECT, p-value = 0.059). But we cannot remove them from the model because we cannot consider that they are simultaneously equal to zero (THAL, p-value = 0.004). It may seem confusing as I said above. But we must not forget that these two indicators are not independent.

6.2 Backward selection I

First, we perform the backward elimination (at 1% level) where we treat as a whole the indicators coming from the same categorical predictors.



Régression logistique

Paramètres de fonctionnement

Probabilité critique pour la méthode: 0.05

Nombre maximum d'itérations: 25

Modalité de référence: ● Première ○ Dernière □ Autoriser le choix par variable

Nombre maximum de modalités: 20 *Au delà la variable nominale est traitée comme une continue*

Probabilités a priori: ● Identiques à l'échantillon ○ Egales ○ Autres Définir...

Matrice des coûts: ● Unitaires ○ Autres Définir...

Pré-traitement des données

Modalités problématiques: ● Supprimer la modalité ○ Supprimer la variable entière

Détection des liaisons: Désactivée Seuil de référence >= 0.9

Méthode de sélection

○ Pas de sélection de variables

○ Sélection forward

● Sélection backward

○ Sélection stepwise

Paramètres de sélection

□ Autoriser la sélection individuelle des modalités d'une variable

Probabilité critique pour l'entrée d'une variable: 0.01

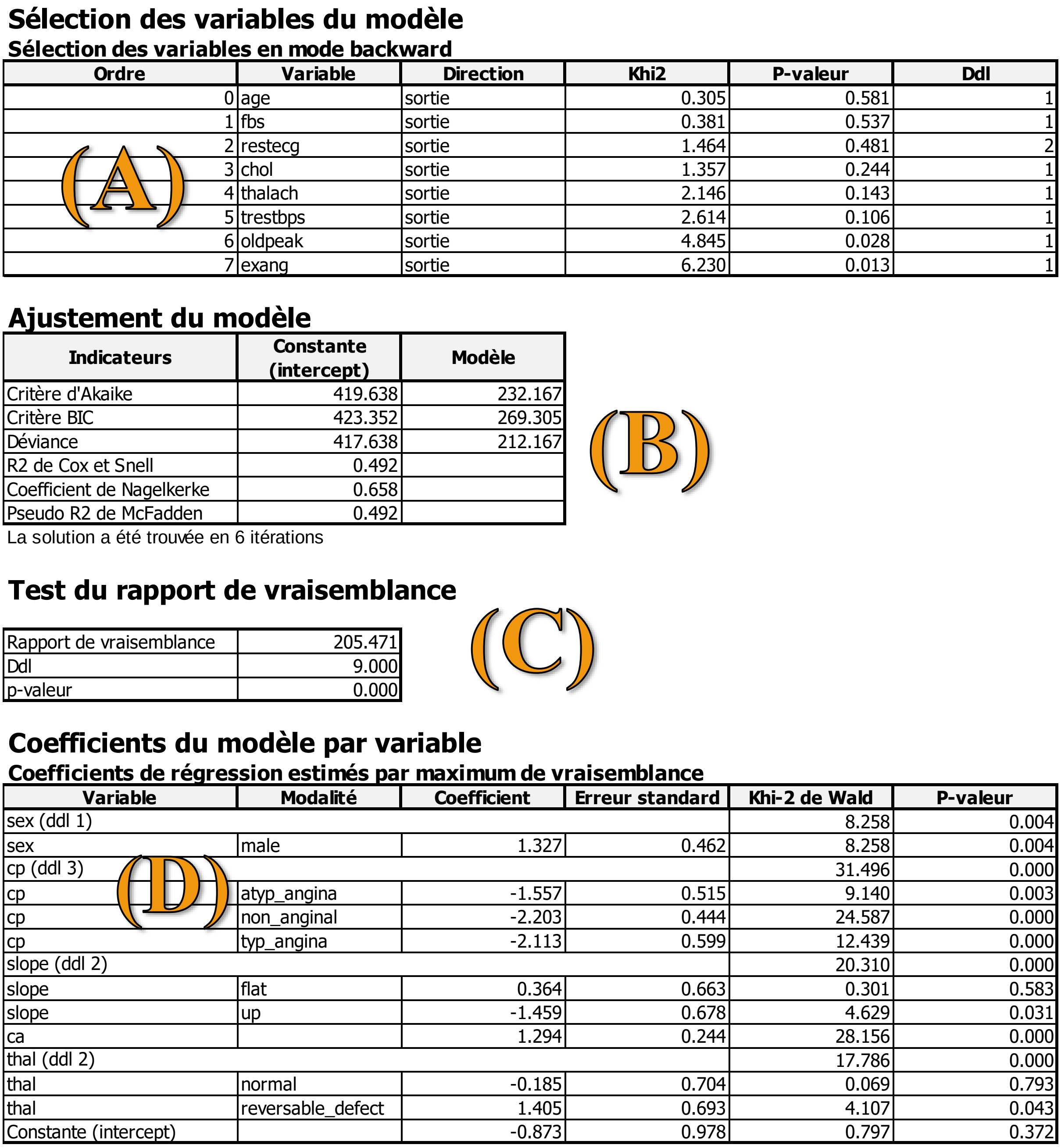
Probabilité critique pour la sortie d'une variable: 0.01

□ Nombre maximum de variables/modalités dans le modèle: 40

Gestion des données manquantes

We validate the settings by clicking on the button OK. Here are the details of the results (Figure 19):

- (A) The list of the variables removed at each step. This table is the same as the summary table provided by SAS (Figure 7).
- (B) The overall evaluation of the model. We observe among others that the AIC = 232.167.
- (C) The likelihood ratio test which assesses the global significance of the model.
- (D) The coefficients of the model, including the Type 3 Analysis of Effects.



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

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

Test du rapport de vraisemblance

Rapport de vraisemblance	205.471
Ddl	9.000
p-valeur	0.000

Coefficients du modèle par variable

Coefficients de régression estimés par maximum de vraisemblance

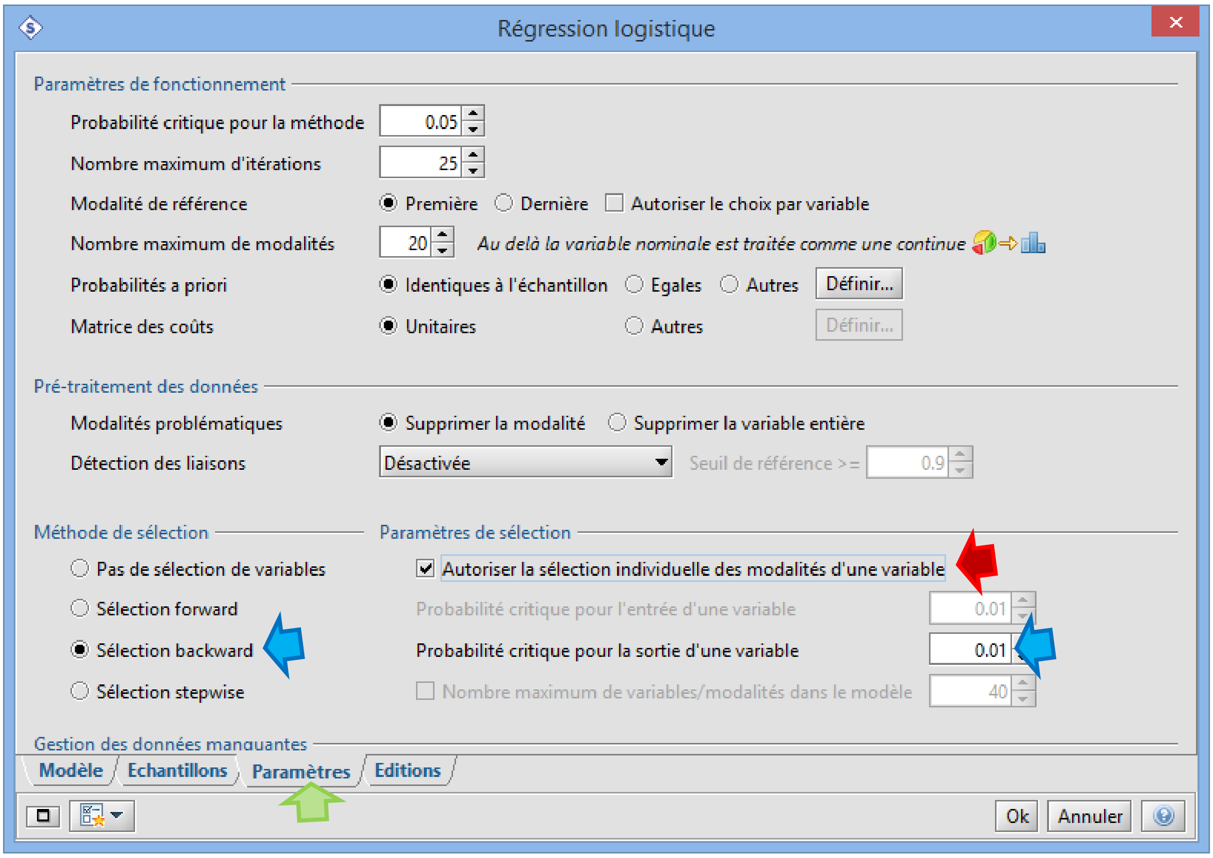
Variable	Modalité	Coefficient	Erreur standard	Khi-2 de Wald	P-valeur
sex (ddl 1)				8.258	0.004
sex	male	1.327	0.462	8.258	0.004
cp (ddl 3)				31.496	0.000
cp	atyp_angina	-1.557	0.515	9.140	0.003
cp	non_anginal	-2.203	0.444	24.587	0.000
cp	typ_angina	-2.113	0.599	12.439	0.000
slope (ddl 2)				20.310	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

Figure 19 - Backward elimination I - SPAD

The results are identical in all respects to those of SAS.

6.3 Backward selection II

In this section, we want to get the same behavior as Tanagra. We open again the settings dialog box. We specify that we treat individually the indicator variables i.e. in French: "*Autoriser la sélection individuelle des modalités d'une variable*" (red arrow).



Régression logistique

Paramètres de fonctionnement

Probabilité critique pour la méthode: 0.05

Nombre maximum d'itérations: 25

Modalité de référence: ● Première ○ Dernière □ Autoriser le choix par variable

Nombre maximum de modalités: 20 *Au delà la variable nominale est traitée comme une continue*

Probabilités a priori: ● Identiques à l'échantillon ○ Egales ○ Autres [Définir...]

Matrice des coûts: ● Unitaires ○ Autres [Définir...]

Pré-traitement des données

Modalités problématiques: ● Supprimer la modalité ○ Supprimer la variable entière

Détection des liaisons: Désactivée — Seuil de référence >= 0.9

Méthode de sélection

○ Pas de sélection de variables

○ Sélection forward

● Sélection backward

○ Sélection stepwise

Paramètres de sélection

☑ Autoriser la sélection individuelle des modalités d'une variable

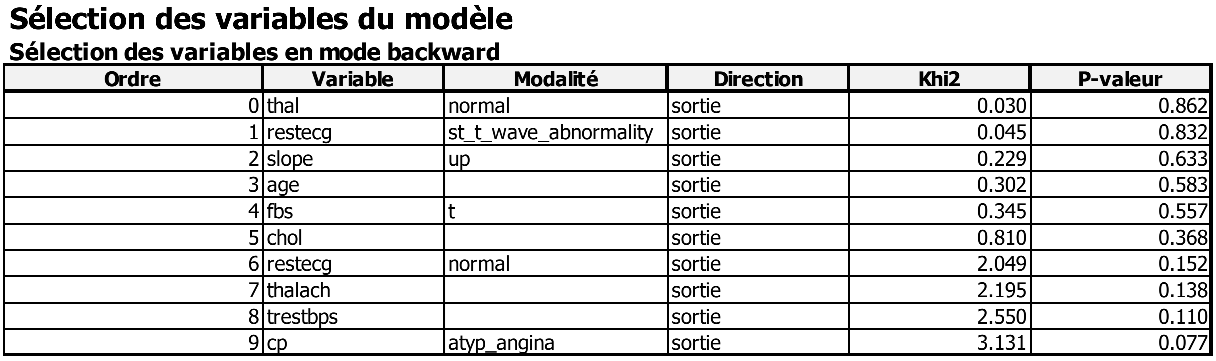
Probabilité critique pour l'entrée d'une variable: 0.01

Probabilité critique pour la sortie d'une variable: 0.01

□ Nombre maximum de variables/modalités dans le modèle: 40

Gestion des données manquantes

The selection table traces the elimination of the variables (numeric variables or indicator variables) (Figure 20).



Sélection des variables du modèle

Sélection des variables en mode backward

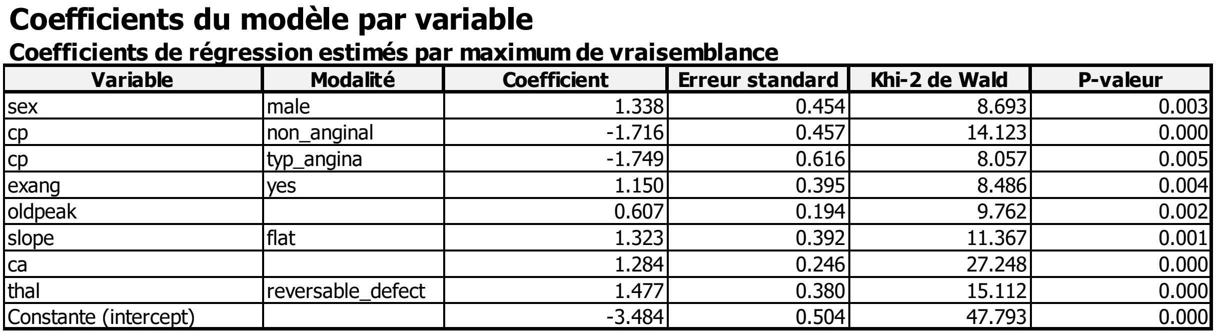
Ordre	Variable	Modalité	Direction	Khi2	P-valeur
0	thal	normal	sortie	0.030	0.862
1	restecg	st_t_wave_abnormality	sortie	0.045	0.832
2	slope	up	sortie	0.229	0.633
3	age		sortie	0.302	0.583
4	fbs	t	sortie	0.345	0.557
5	chol		sortie	0.810	0.368
6	restecg	normal	sortie	2.049	0.152
7	thalach		sortie	2.195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

Figure 20 - Backward elimination II - SPAD

Compared with Tanagra, as the references groups are not the same for categorical variables, the removed indicators may be different. Thus, at the end, we can get - and that is true for our dataset - different models despite identical settings (significance level for the Wald test).

Conclusion: In the case of variable selection context, the choice of the reference group has an influence on the characteristics of the final model when we process individually the indicator variables.

In SPAD, we get the following final classifier (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 - Classifier after the backward elimination II - SPAD

6.4 What is the best approach?

"What is the best approach?" is an inevitable question in the present study. Actually, there is not really a consensus about the best method in variable selection. The various approaches provide scenarios for solutions. All depends on what we do with the classifiers thereafter. The only objective evaluation would be to measure the performance of the resulting model (after the feature selection process) on a test set. But we must have a test set with sufficient number of instances to get a reliable evaluation of the error rate.

Drawing a parallel with decision trees, we could say that: treating the indicators from a categorical variable as a whole is similar to generate branches for each possible value of the variable during the splitting process of a node; treating individually the indicators enables to merge certain values as with CHAID induction tree algorithm (or others).

7 Conclusion

In this tutorial, we have studied the approaches implemented in various data mining tools for dealing with categorical predictors in variable selection context. All are based on the dummy coding of the categorical variables, but the indicators coming from one variable may be treated as whole or individually in the selection process. The resulting classifier can be different following the approaches.