1 Topic

Description of the SAS PROC LOGISTIC. Measuring its performance on large datasets.

In my courses at the University (http://dis.univ-lyon2.fr/), I use only free data mining tools (R, Tanagra, Sipina, Knime, Orange, etc.) and the spreadsheet applications (free or not). Sometimes, my students ask me if the commercial tools (e.g. SAS which is very popular in France) have different behavior, in terms of how to use, or for the reading of the results. I say them that some of these commercial tools are available on the computers of our department. They can learn how to use them by taking as a starting point the tutorials available on the Web.

But unfortunately, especially in the French language, they are not numerous about the logistic regression. We need a didactic document with clear screenshots which show how to: (1) import a data file into a SAS bank; (2) define an analysis with the appropriate settings; (3) read and understand the results.

In this tutorial, we describe the use of the SAS PROC LOGISTIC (**SAS 9.3**). We measure its quickness when we handle a moderate sized dataset. We compare the results with those of **Tanagra 1.4.43**.

2 Dataset

We use a binary version of the waveform database (Breiman and al., 1984). We have already used this data file in a previous tutorial[1]. We generate a dataset with the same characteristics (300,000 instances and 121 independent variables), but not the same values because the random number generator is initialized differently. We set the following source code under R.

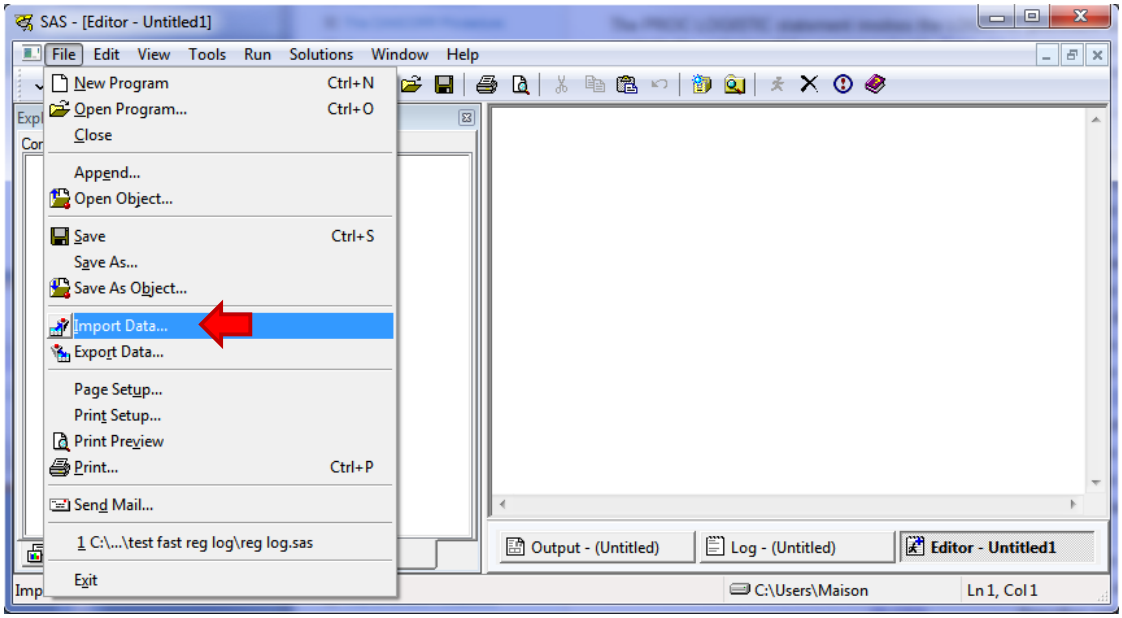
```
#generate and save a dataset
set.seed(1)
dataset.size <- 300000 #number of instances
nb.rnd <- 50 #number of random variables
nb.cor <- 50 #number of correlated variables
noise.level <- 1 #noise for correlated variables
data.wave <- generate.binary(dataset.size,nb.rnd,nb.cor,noise.level)
summary(data.wave)
#writting
write.table(data.wave,file="wavebin.txt",quote=F,sep="\t",row.names=F)
```

3 Importing the data file into SAS

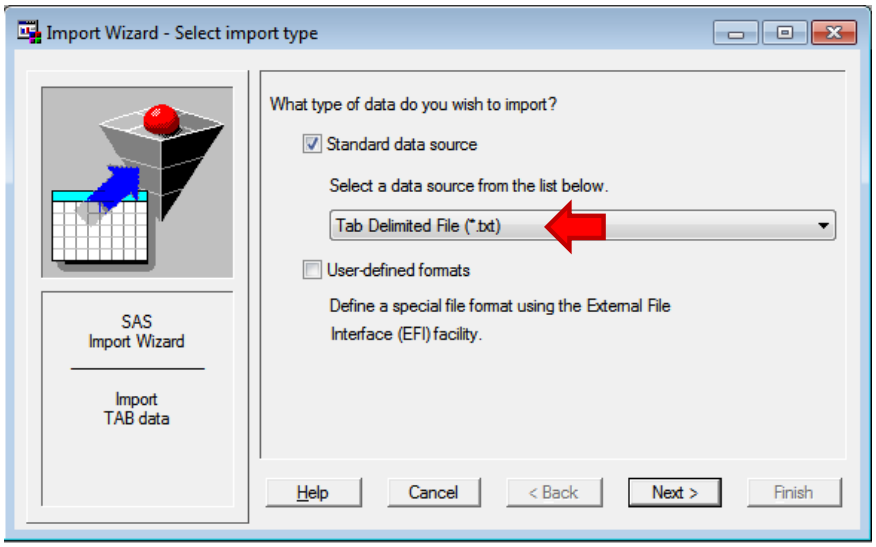
We have to import the data file "**wavebin.txt**" into a SAS data bank[2]. We launch SAS and we activate the FILE / IMPORT DATA menu.

[1] http://data-mining-tutorials.blogspot.fr/2012/02/logistic-regression-on-large-dataset.html

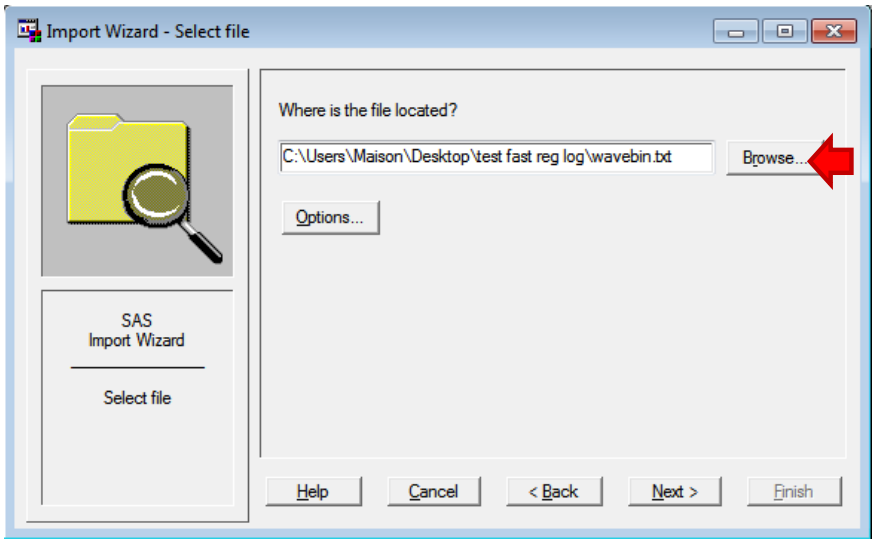
[2] Text file format with tab-separated - http://en.wikipedia.org/wiki/Tab-separated_values



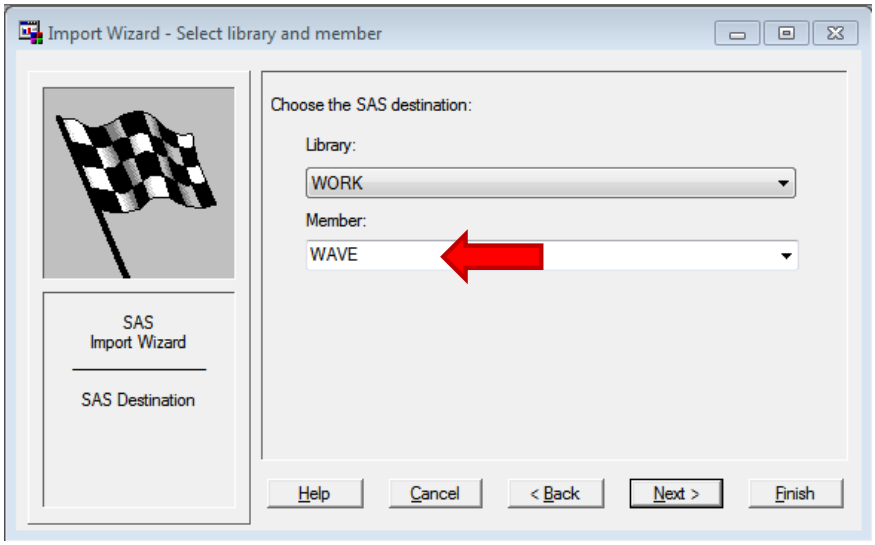
A wizard enables to set the kind of the data source [“Tab Delimited File (*.txt)”].



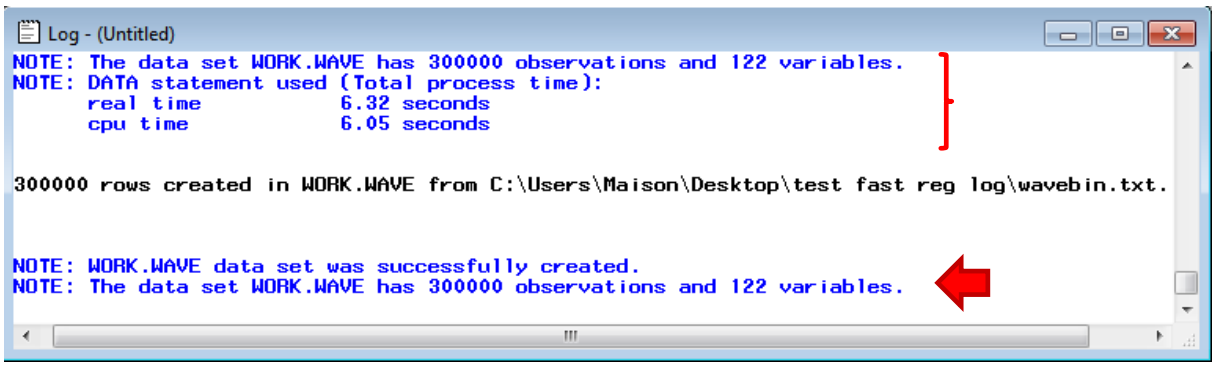
We click on the NEXT button. We set the file name.



Last, SAS asks us the name of the data bank where we want to insert the dataset. To simplify the management, we use the WORK data bank. We set WAVE as database name.



We validate by clicking on the FINISH button. The data file is imported in about 6 seconds.



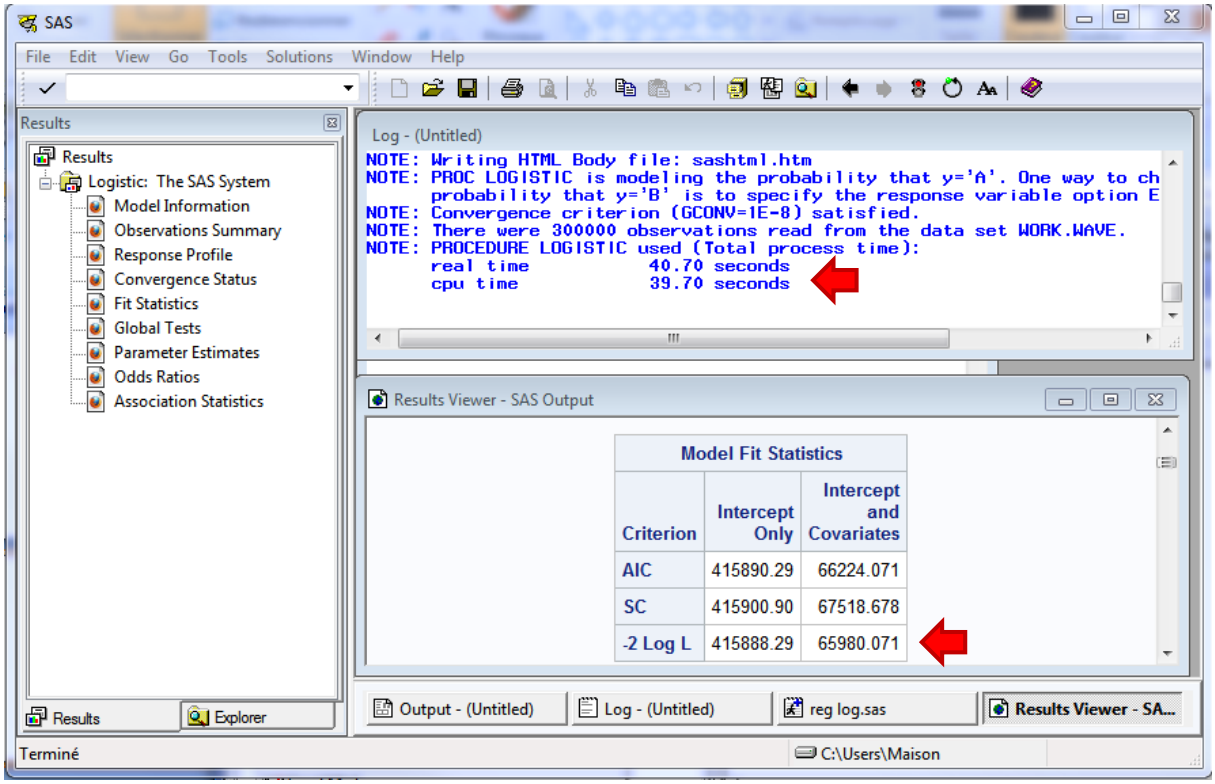
4 Logistic regression on all the candidate variables

We want to explain the dependent variable Y by all the other available variables using a logistic regression. We use the following command:

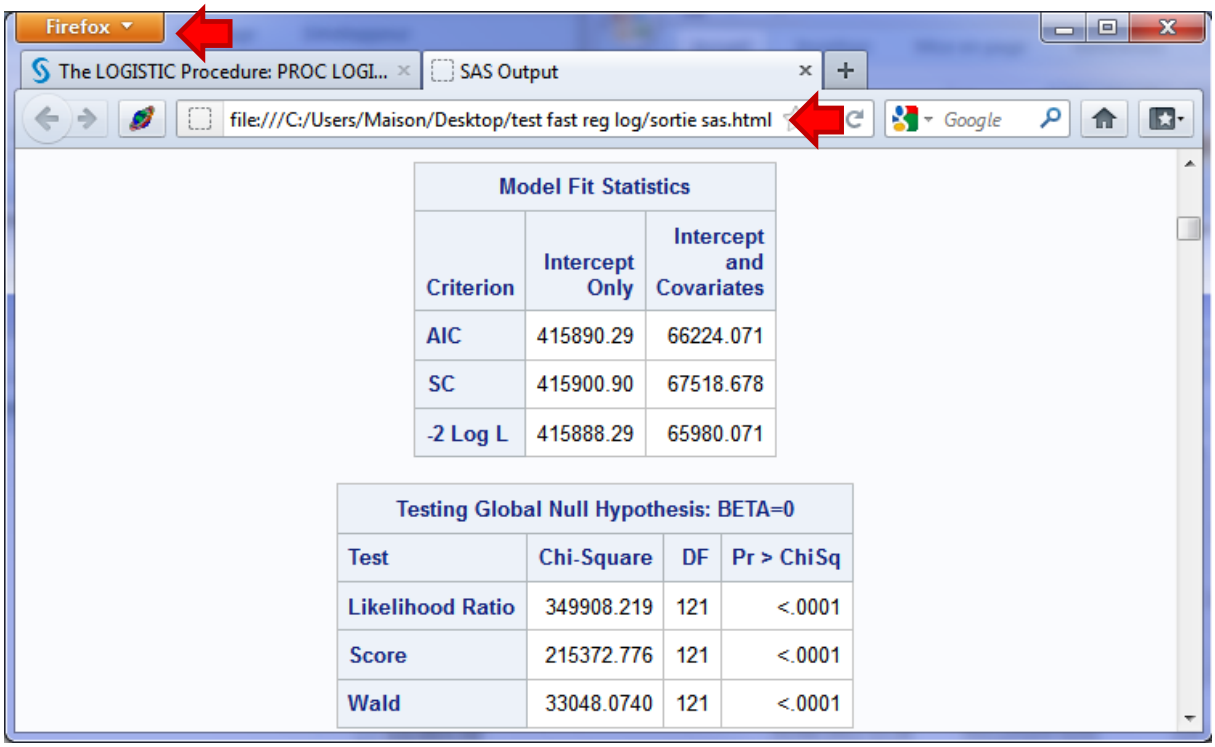
```
proc logistic data = wave;
model y = V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
V19 V20 V21 rnd_1 rnd_2 rnd_3 rnd_4 rnd_5 rnd_6 rnd_7 rnd_8 rnd_9 rnd_10
rnd_11 rnd_12 rnd_13 rnd_14 rnd_15 rnd_16 rnd_17 rnd_18 rnd_19 rnd_20
rnd_21 rnd_22 rnd_23 rnd_24 rnd_25 rnd_26 rnd_27 rnd_28 rnd_29 rnd_30
rnd_31 rnd_32 rnd_33 rnd_34 rnd_35 rnd_36 rnd_37 rnd_38 rnd_39 rnd_40
rnd_41 rnd_42 rnd_43 rnd_44 rnd_45 rnd_46 rnd_47 rnd_48 rnd_49 rnd_50 cor_1
cor_2 cor_3 cor_4 cor_5 cor_6 cor_7 cor_8 cor_9 cor_10 cor_11 cor_12 cor_13
cor_14 cor_15 cor_16 cor_17 cor_18 cor_19 cor_20 cor_21 cor_22 cor_23
cor_24 cor_25 cor_26 cor_27 cor_28 cor_29 cor_30 cor_31 cor_32 cor_33
cor_34 cor_35 cor_36 cor_37 cor_38 cor_39 cor_40 cor_41 cor_42 cor_43
cor_44 cor_45 cor_46 cor_47 cor_48 cor_49 cor_50;

run;
```

We obtain the results after about 40 seconds.



We can export the results into a HTML file format (the RESULTS VIEWER tab must be activated). The output can be visualized into a browser.



Let us describe the outputs of SAS. We compare them to those of Tanagra.

4.1 Global evaluation of the model

Clearly, the model is globally significant.

SAS

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	415890.29	**66224.071**
SC	415900.90	**67518.678**
-2 Log L	415888.29	**65980.071**

TANAGRA

Model Fit Statistics

Criterion	Intercept	Model
AIC	415890.29	**66224.071**
SC	415900.90	**67518.678**
-2LL	415888.29	**65980.071**

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	349908.219	121	<.0001
Score	215372.776	121	<.0001
Wald	33048.074	121	<.0001

Model Chi test (LR)	
Chi-2	349908.2193
d.f.	121
P(>Chi-2)	0

SAS provides more tests: likelihood ratio (LR), but also the score test, and the Wald test. The first one is the mode powerful (in the statistical sense).

4.2 Coefficients

We show only the first 5 coefficients here. The results are consistent. We note however that the estimation of the standard error, and consequently the Wald test, may be very slightly different. This is because the internal convergence conditions are not the same according the tools. We have the same phenomenon whatever the software used (free or commercial).

SAS

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.1435	0.1356	1.1208	0.2897
V1	1	-0.0213	0.0142	2.2625	0.1325
V2	1	-0.0905	0.0192	22.2763	<.0001
V3	1	-0.2344	0.0196	143.2749	<.0001
V4	1	-0.3267	0.0122	716.1681	<.0001
V5	1	-0.4271	0.0168	643.7802	<.0001

TANAGRA

Attributes in the equation

Attribute	Coef.	Std-dev	Wald	Signif
constant	-0.1435	0.1356	1.1208	0.2897
V1	-0.0213	0.0142	2.2625	0.1325
V2	-0.0905	0.0192	22.2763	0
V3	-0.2344	0.0196	143.2750	0
V4	-0.3267	0.0122	716.1688	0
V5	-0.4271	0.0168	643.7808	0

4.3 Odds-ratio and their confidence intervals

Last, we have the odds-ratio and their confidence intervals at the 95% confidence level.

SAS

Odds Ratio Estimates

Effect	Point Estim	95% Wald Confidence Limits	
V1	0.979	0.952	1.006
V2	0.914	0.880	0.948
V3	0.791	0.761	0.822
V4	0.721	0.704	0.739
V5	0.652	0.631	0.674

TANAGRA

Odds ratios and 95% confidence intervals

Attribute	Coef.	Low	High
V1	0.979	0.952	1.007
V2	0.914	0.880	0.949
V3	0.791	0.761	0.822
V4	0.721	0.704	0.739
V5	0.652	0.631	0.674

5 Logistic regression with variable selection

Now we want to perform a variable selection in order to obtain the relevant predictors only. We use the forward strategy based on the score test. The significance level used is 1%. The SAS command is the following:

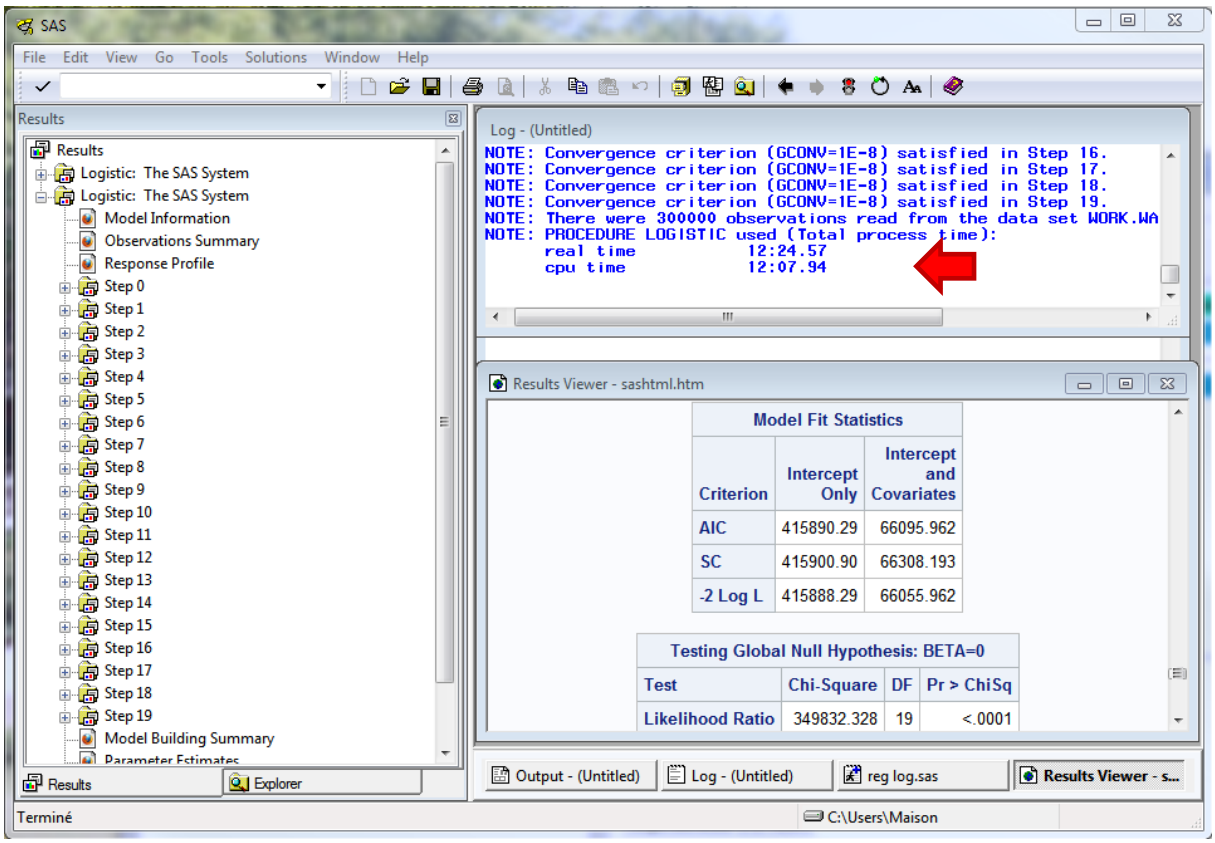
```
proc logistic data = wave;
model y = V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
V19 V20 V21 rnd_1 rnd_2 rnd_3 rnd_4 rnd_5 rnd_6 rnd_7 rnd_8 rnd_9 rnd_10
rnd_11 rnd_12 rnd_13 rnd_14 rnd_15 rnd_16 rnd_17 rnd_18 rnd_19 rnd_20
rnd_21 rnd_22 rnd_23 rnd_24 rnd_25 rnd_26 rnd_27 rnd_28 rnd_29 rnd_30
rnd_31 rnd_32 rnd_33 rnd_34 rnd_35 rnd_36 rnd_37 rnd_38 rnd_39 rnd_40
rnd_41 rnd_42 rnd_43 rnd_44 rnd_45 rnd_46 rnd_47 rnd_48 rnd_49 rnd_50 cor_1
cor_2 cor_3 cor_4 cor_5 cor_6 cor_7 cor_8 cor_9 cor_10 cor_11 cor_12 cor_13
cor_14 cor_15 cor_16 cor_17 cor_18 cor_19 cor_20 cor_21 cor_22 cor_23
cor_24 cor_25 cor_26 cor_27 cor_28 cor_29 cor_30 cor_31 cor_32 cor_33
cor_34 cor_35 cor_36 cor_37 cor_38 cor_39 cor_40 cor_41 cor_42 cor_43
cor_44 cor_45 cor_46 cor_47 cor_48 cor_49 cor_50 / selection = forward
slentry = 0.01;

run;
```

Let us note the options SELECTION and SLENTRY for the variable selection.

After 12 minutes and 7 seconds, a model with 9 variables is proposed.

Only one irrelevant attribute (COR_32) is included into the model. This is really noteworthy in view of the database size (number of instances), which tends to make significant all the variables, and the number of initial irrelevant attributes (RND and COR variables[3]).



We observe that both SAS and TANAGRA are based on the same variable selection mechanism (the score test for logistic regression).

[3] See http://data-mining-tutorials.blogspot.fr/2012/02/logistic-regression-on-large-dataset.html for the generation of the database.

SAS

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	V15	1	1	165232.509	<.0001
2	V7	1	2	46531.6437	<.0001
3	V14	1	3	16495.3124	<.0001
4	V8	1	4	11262.9599	<.0001
5	V6	1	5	7904.8714	<.0001
6	V16	1	6	6748.6872	<.0001
7	V17	1	7	3580.848	<.0001
8	V5	1	8	3365.3435	<.0001
9	V13	1	9	2604.9987	<.0001
10	V9	1	10	2548.4347	<.0001
11	V18	1	11	1258.052	<.0001
12	V4	1	12	1207.989	<.0001
13	V19	1	13	484.5763	<.0001
14	V3	1	14	468.3413	<.0001
15	V12	1	15	428.3135	<.0001
16	V10	1	16	465.539	<.0001
17	V2	1	17	115.1171	<.0001
18	V20	1	18	98.6673	<.0001
19	cor_32	1	19	7.4248	0.0064

TANAGRA

N°	AIC	Variable	CHI-SQUARE	p-value
1	AIC : 415890.29	V15	Chi-2 : 165232.509	p : 0.0000
2	AIC : 183858.53	V7	Chi-2 : 46535.214	p : 0.0000
3	AIC : 128113.30	V14	Chi-2 : 16495.702	p : 0.0000
4	AIC : 110328.05	V8	Chi-2 : 11262.962	p : 0.0000
5	AIC : 98389.26	V6	Chi-2 : 7904.881	p : 0.0000
6	AIC : 90081.47	V16	Chi-2 : 6748.742	p : 0.0000
7	AIC : 83018.83	V17	Chi-2 : 3580.968	p : 0.0000
8	AIC : 79339.54	V5	Chi-2 : 3365.587	p : 0.0000
9	AIC : 75894.39	V13	Chi-2 : 2604.999	p : 0.0000
10	AIC : 73239.77	V9	Chi-2 : 2548.435	p : 0.0000
11	AIC : 70642.63	V18	Chi-2 : 1258.052	p : 0.0000
12	AIC : 69374.46	V4	Chi-2 : 1207.989	p : 0.0000
13	AIC : 68157.23	V19	Chi-2 : 484.576	p : 0.0000
14	AIC : 67672.70	V3	Chi-2 : 468.342	p : 0.0000
15	AIC : 67204.61	V12	Chi-2 : 428.314	p : 0.0000
16	AIC : 66776.63	V10	Chi-2 : 465.539	p : 0.0000
17	AIC : 66311.30	V2	Chi-2 : 115.117	p : 0.0000
18	AIC : 66198.11	V20	Chi-2 : 98.667	p : 0.0000
19	AIC : 66101.39	cor_32	Chi-2 : 7.425	p : 0.0064

6 Comparison of the calculation times

The statistical results are the same. What about the computation time? We compare SAS and TANAGRA at each step of the process, we obtain the following results ("n" is the number of instances, "p" is the number of candidate predictors).

Wave (n = 300.000, p = 121)	**SAS 9.3**	**TANAGRA 1.4.43**
Data importation	**6 sec.**	9 sec.
Full model	**40 sec.**	74 sec.
Variable selection process	12 mn et 7 sec.	**10 mn et 48 sec.**

SAS is faster, this is not surprising. The ability of SAS to handle large dataset is well-known. But, surprisingly, it seems less quick for the variable selection process. I do not really understand why. Perhaps, the implementation of the score test requires more operations under SAS. Anyway, we note that Tanagra is suitably efficient on the moderate dataset such as we handle in this tutorial.

7 Conclusion

In this paper, we describe shortly the SAS PROC LOGISTIC. The tool incorporates many options. We must study carefully the documentation to understand all of them. For a standard usage (e.g. for the courses at the University), I think free tools such as Tanagra or R (or others[4]) are quite sufficient.

[4] http://data-mining-tutorials.blogspot.fr/2008/12/logistic-regression-software-comparison.html