1 Objectif

Description de la PROC LOGISTIC de SAS. Comportement sur les grandes bases.

Un étudiant est venu me voir une fois pour me demander si je comptais décrire l'utilisation de la « proc logistic » de SAS durant mon cours de régression logistique (Master SISE)[1]. Je lui ai dit qu'on utilisait suffisamment d'outils comme ça (R, SPAD, SIPINA, TANAGRA et le tableur EXCEL), je ne voyais pas trop l'intérêt de voir un logiciel supplémentaire. D'autant plus que le plus important finalement est de bien maîtriser la chaîne de traitements, de comprendre la finalité et les implications de chacune des étapes, et de savoir lire les résultats. Qu'importe l'outil, la démarche reste toujours la même. Et puis, tout à fait prosaïquement, les heures ne sont malheureusement pas extensibles à l'infini dans nos Universités. Je lui dis alors que SAS étant disponible dans nos salles informatiques, il ne tenait qu'à lui de s'exercer en récupérant les très nombreux tutoriels accessibles sur internet.

Après coup, je suis allé vérifier moi-même sur le web. Et je me suis rendu compte qu'ils ne sont pas si nombreux que ça finalement les tutoriels en français, avec des copies d'écran explicites, montrant de manière simple et didactique la chaîne complète de traitements allant de l'importation de données jusqu'à la récupération des résultats. Je me suis dit qu'il y avait là des choses à faire.

Dans ce tutoriel, nous décrivons l'utilisation de la « proc logistic » de **SAS 9.3**, sans et avec la sélection de variables. Nous en profiterons pour étudier ses performances (essentiellement la rapidité de calcul) sur une base de grande taille. Nous comparerons les valeurs obtenues avec celles de **Tanagra 1.4.43**.

2 Données

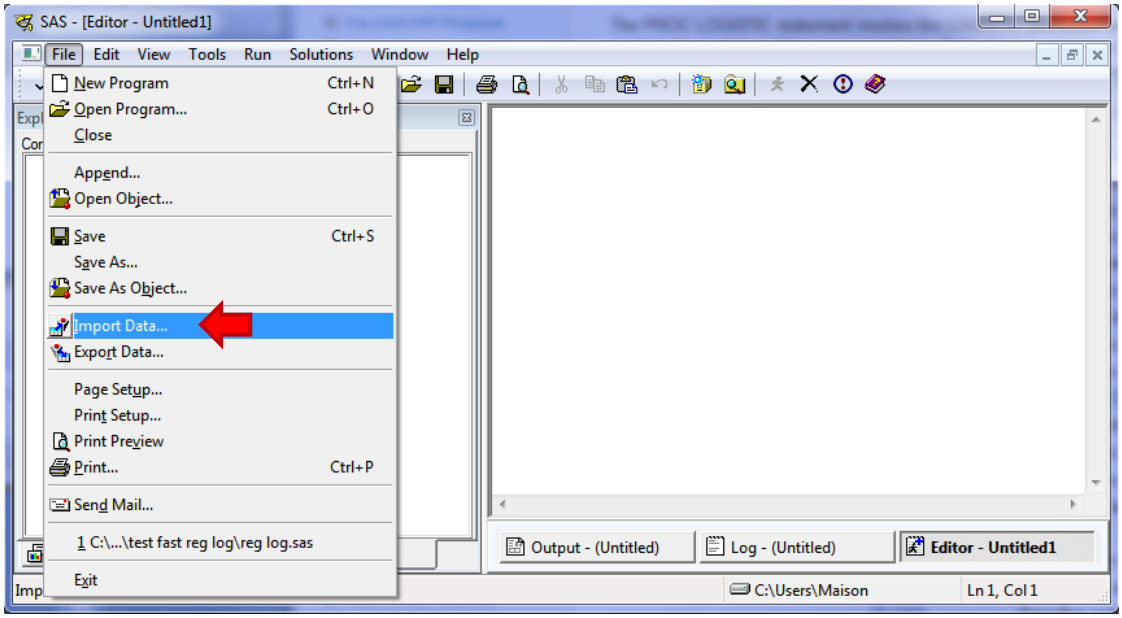
Nous utilisons une variante binaire de la base « wave » de Breiman et al. (1984). Nous l'avons mis à contribution dans un de nos anciens tutoriels où nous étudions le comportement de différents logiciels sur une (relativement) grande base de données (voir « Régression logistique sur les grandes bases » pour une description approfondie des données générées). Nous avons les mêmes caractéristiques (121 variables prédictives, 300.000 observations), mais pas exactement les mêmes valeurs puisque le générateur opère aléatoirement. Pour éviter cet aléa, nous l'initialisons arbitrairement dans ce tutoriel. Le code en langage R du programme principal utilisé est le suivant :

```
#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)
```

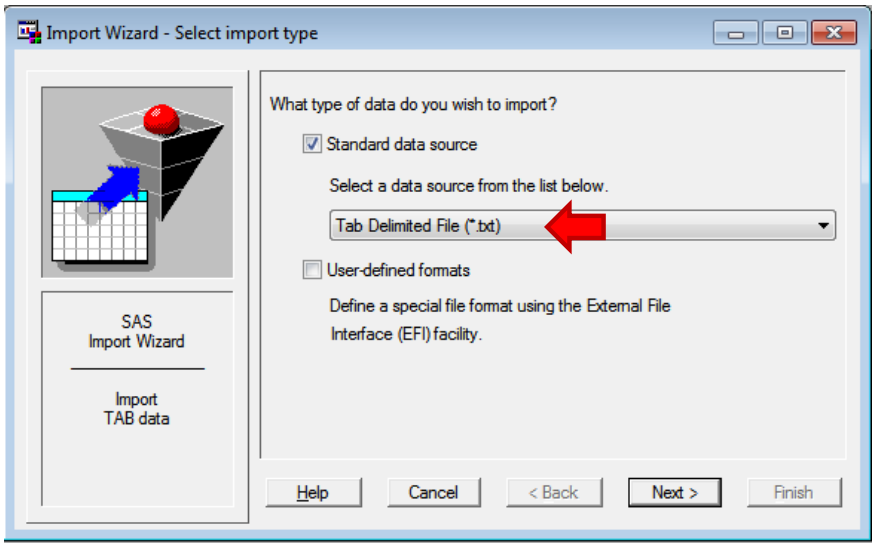
[1] http://dis.univ-lyon2.fr/?page_id=1288

3 Importation des données dans SAS

Première étape dans tout logiciel, il faut importer le fichier de données « **wavebin.txt** ». Il est au format texte avec séparateur tabulation en ce qui nous concerne[2]. Après avoir démarré SAS[3], nous actionnons le menu « **File / Import Data…** »[4].



Un assistant apparaît pour nous guider tout au long du processus. Dans un premier temps, nous devons spécifier le type du fichier « Tab Delimited File (*.txt) ».



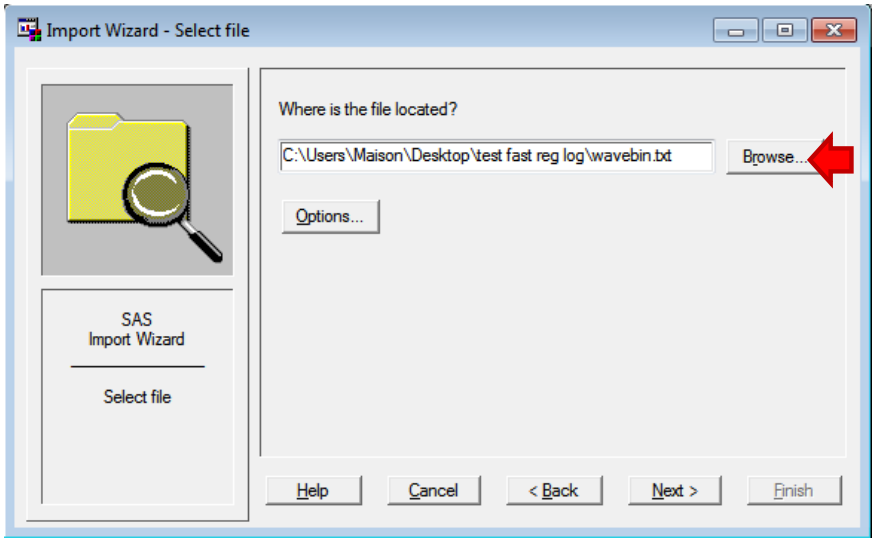
Nous cliquons sur NEXT. Nous devons maintenant spécifier le fichier.

---

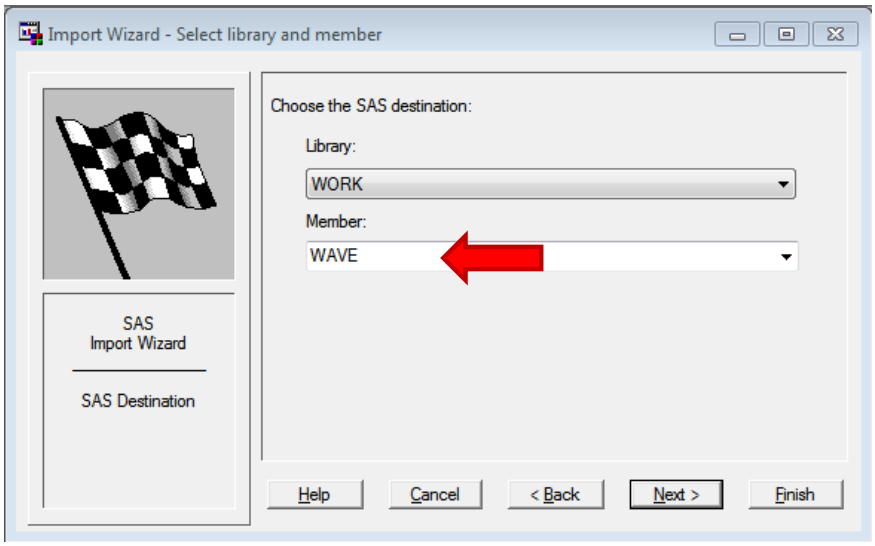
[2] Nous utilisons l'extension classique « .TXT ». Il semble qu'une nouvelle extension « .TSV » soit dédiée à ce type de fichier pour mieux le caractériser - http://www.cs.tut.fi/~jkorpela/TSV.html

[3] Je connais peu SAS. Il y a sûrement d'autres manières de faire plus performantes ou plus immédiates. J'en sais suffisamment en revanche pour arriver à mes fins. C'est le plus important après tout.

[4] J'utilise la version anglaise parce que ce tutoriel sera traduit en anglais par la suite. La transposition à SAS en français ne pose absolument aucun problème. Et les commandes SAS sont les mêmes.

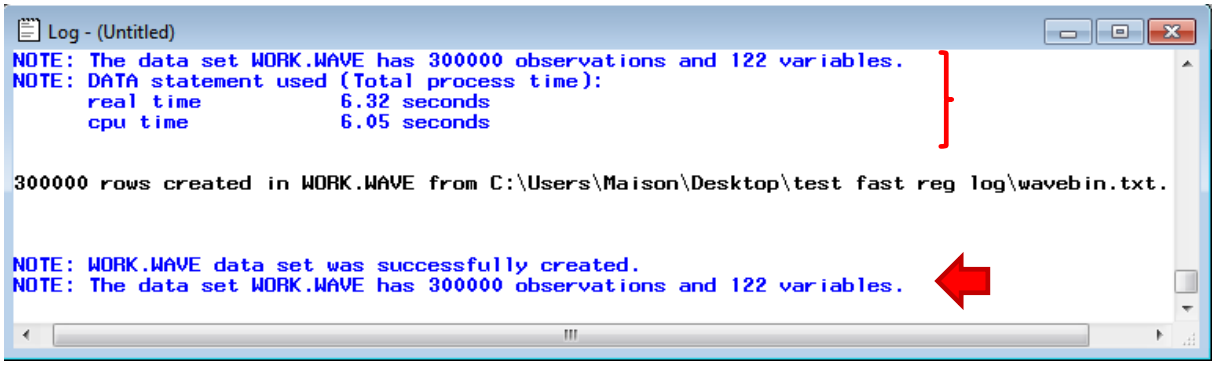


Enfin, SAS nous demande dans quelle banque nous souhaitons stocker le jeu de données. Pour simplifier, nous utilisons la banque WORK. Nous attribuons le nom WAVE.



Il ne reste plus qu'à cliquer sur le bouton FINISH.

Au bout de 6 secondes, les données sont importées. SAS détaille les opérations dans le fichier log (« Journal » en français).

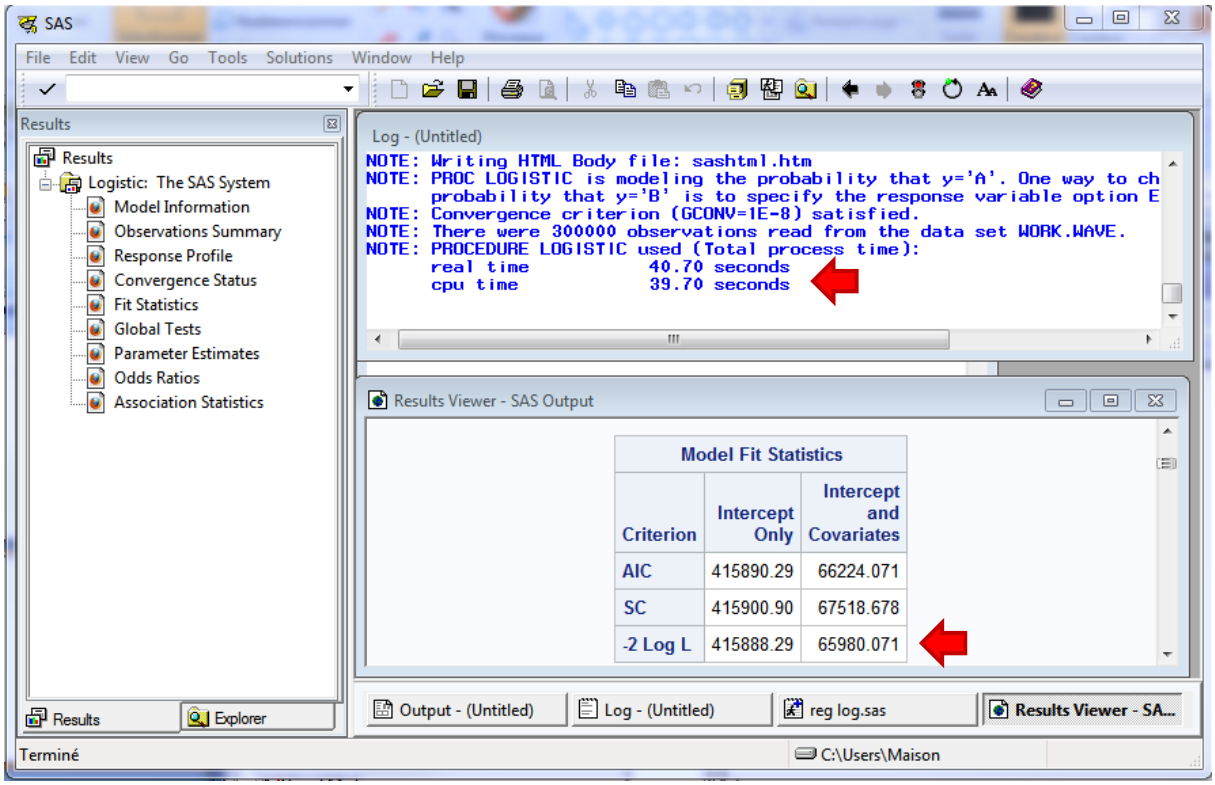


4 Régression logistique sur toutes les variables

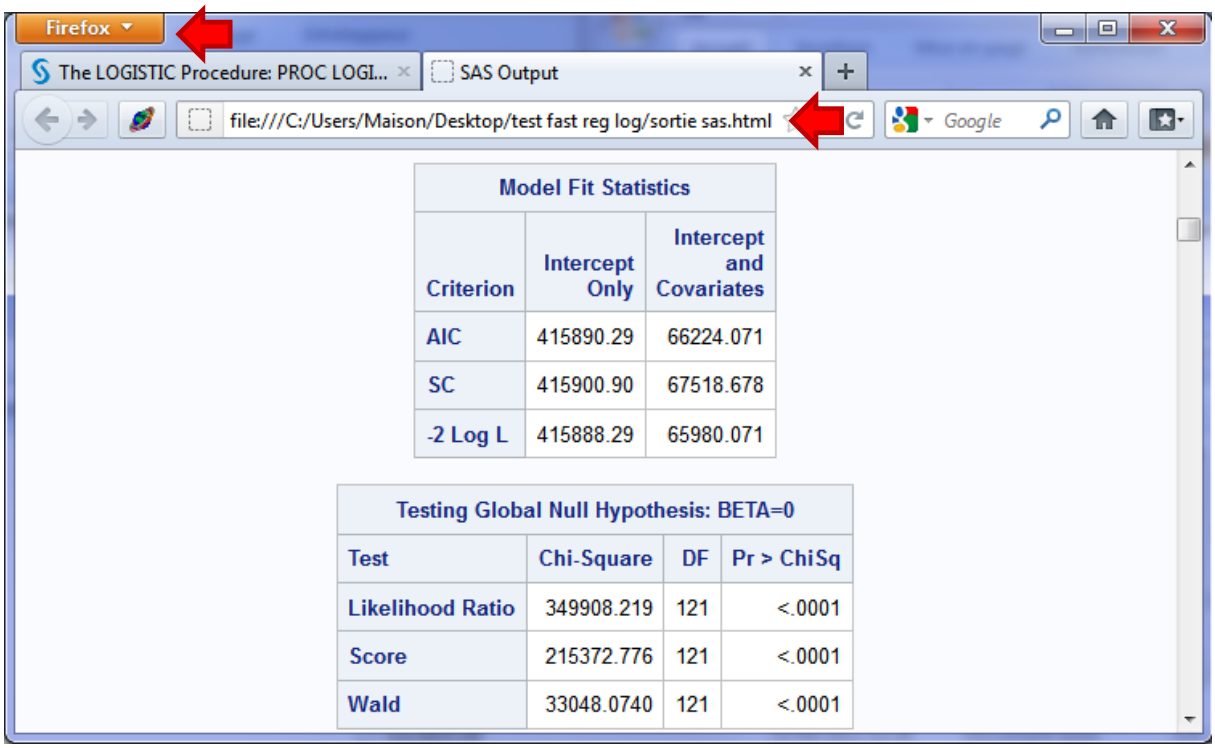
Nous souhaitons expliquer les valeurs de la cible « y » à partir de toutes les autres variables. Les instructions SAS sont les suivantes :

```
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;
```

Les résultats s'affichent au bout de 40 secondes.



Nous pouvons les sorties au format HTML en actionnant le menu « File / Save As » (en veillant à bien activer la fenêtre « Results Viewer »). Le document peut être consulté dans un navigateur web (exactement comme pour Tanagra !).



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

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	349908.219	121	<.0001
Score	215372.776	121	<.0001
Wald	33048.0740	121	<.0001

Voyons le détail des sorties de SAS, et comparons-les avec ceux de Tanagra.

4.1 Evaluation globale

Manifestement, le modèle est globalement très significatif.

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 fournit plusieurs tests : le test du rapport de vraisemblance, le test du score et le test de Wald. Le premier est le plus puissant. Les autres sont plus conservateurs (favorisent H0).

4.2 Coefficients

Nous n'affichons que les 5 premiers coefficients. Ici également, les résultats sont cohérents. Cependant, les conditions de convergence n'étant pas les mêmes, les estimations des écarts-type et, par conséquent, la statistique de Wald peuvent être différents à la 4ème décimale. Nous retrouverons le même phénomène quel que soit le logiciel utilisé pour la régression logistique.

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 et intervalles de confiance

Enfin, nous avons les odds-ratio et leur intervalle de confiance à 95%. Ici également, les résultats sont identiques.

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 Régression logistique avec sélection de variables

Nous souhaitons maintenant opérer une sélection forward à 1% de manière à ne conserver que les variables explicatives pertinentes. SAS utilise le test de score dans ce contexte, tout comme Tanagra. Nous nous attendons donc à obtenir exactement la même séquence de sélection et, à la sortie, le même modèle final.

La commande SAS devient la suivante :

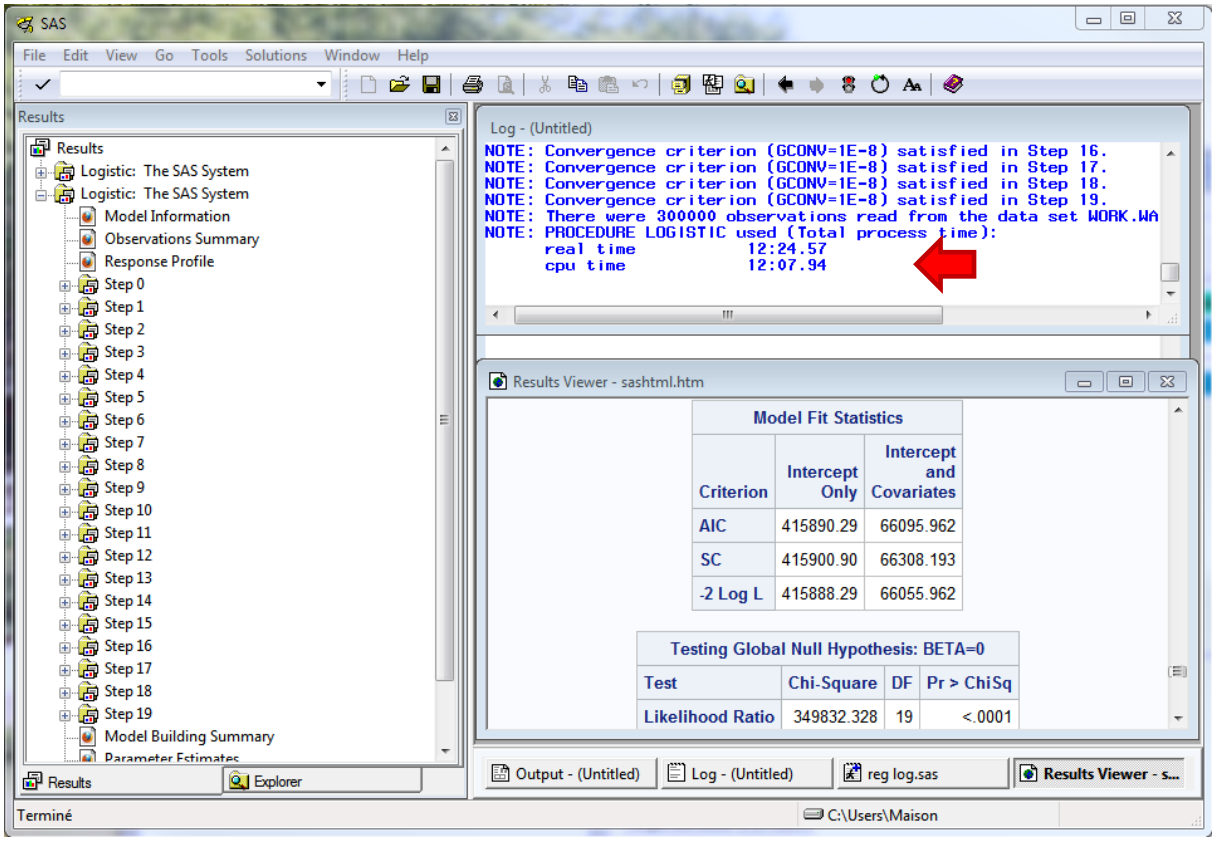
```
proc logistic data = wave;
model y = V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15 V16 V17 V18
V19 V20 V21 rnd_1 rnd_2 rnd_3 rnd_4 rnd_5 rnd_6 rnd_7 rnd_8 rnd_9 rnd_10
rnd_11 rnd_12 rnd_13 rnd_14 rnd_15 rnd_16 rnd_17 rnd_18 rnd_19 rnd_20
rnd_21 rnd_22 rnd_23 rnd_24 rnd_25 rnd_26 rnd_27 rnd_28 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;
```

Notons le rôle des options SELECTION et SLENTRY.

Au bout de 12 minutes et 7 secondes, un modèle comportant 19 variables est proposé.

Une variable explicative non pertinente **seulement** (COR_32) a été incorporée dans le modèle. C'est assez remarquable compte tenu de la taille élevée de l'échantillon qui tend à rendre positif tous les tests de significativité.



Nous constatons que SAS et TANAGRA reposent sur le même mécanisme de sélection.

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 Comparaison des performances

Les résultats en termes statistiques sont identiques, qu'en est-il au niveau des performances de calcul ? Nous avons comparé les temps de traitement de SAS et TANAGRA à chaque étape. Nous

obtenons les valeurs suivantes (« n » est la taille de l'échantillon, « p » est le nombre de variables explicatives candidates).

Wave (n = 300.000, p = 121)	**SAS 9.3**	**TANAGRA 1.4.43**
Importation des données	**6 sec.**	9 sec.
Modèle complet	**40 sec.**	74 sec.
Sélection de variables	12 mn et 7 sec.	**10 mn et 48 sec.**

SAS est plus rapide sauf, curieusement, lors de la sélection de variables. Je ne vois pas bien pourquoi en vérité. Est-ce dû une implémentation différente du test du score utilisé pour la sélection ? Il faudrait creuser un peu plus pour en savoir d'avantage. L'autre atout de SAS est qu'il ne semble pas charger la totalité des données en mémoire. C'est un argument fort lorsque la volumétrie devient vraiment importante.

Ceci étant, petite coquetterie de programmeur, je trouve que TANAGRA n'a absolument pas à rougir de ses performances.

7 Conclusion

Dans ce tutoriel, nous avons présenté de manière très succincte la PROC LOGISTIC de SAS dédiée à la régression logistique. L'outil propose de nombreuses options. Il faut étudier attentivement la documentation pour tenter d'en cerner les immenses possibilités. Néanmoins, pour un usage standard sur des bases de taille modérée, y compris pour les enseignements, des outils gratuits tels que Tanagra et R (ex. « Introduction à R – Régression Logistique ») peuvent largement faire l'affaire.