LA PROCÉDURE LOGISTIC : PROCÉDURE DE BASE EN PERPÉTUELLE ÉVOLUTION

La procédure LOGISTIC est apparue dès les premières versions de SAS. Depuis, des évolutions arrivent dans chaque version. Nous allons vous présenter le rôle et les différentes régressions et modèles de cette procédure mais aussi les nouveautés des dernières maintenances.

Caractéristiques :

Catégories : SAS/STAT
OS : Windows, Unix
Version : SAS® 9.4
Vérifié en mars 2015

Sommaire

1. Introduction théorique 2
2. Régression logistique sous SAS 2
   2.1. Instruction CLASS : intégration des variables catégorielles 2
      2.1.1. Présentation des variables indicatrices 2
      2.1.2. Comment changer la modalité de référence d'une variable catégorielle ? 3
      2.1.3. Ecriture de l'équation logistique 4
   2.2. Instruction MODEL 5
   2.3. Instruction SCORE 5
      2.3.1. La première méthode – calcul du modèle et des prédictions dans une seule procédure LOGISTIC 6
      2.3.2. Deuxième méthode – calcul du modèle et des prédictions dans des procédures LOGISTIC séparées 6
      2.3.3. Remarque 6
   2.4. Instruction STRATA 7
3. Régression logistique généralisée, autre possibilité offerte par cette procédure 8
   3.1. Introduction 8
   3.2. Présentation de l'exemple à étudier 8
   3.3. Modélisation avec la proc LOGISTIC 8
      3.3.1. Syntaxe de la proc LOGISTIC dans le cas d'une régression généralisée 8
      3.3.2. Analyse de la sortie 9
      3.3.3. Calcul des probabilités et prédiction 11
         3.3.3.1. Obtention des probabilités prédites 11
         3.3.3.2. Comment faire des prévisions à partir de nouvelles données ? 12
      1. Utilisation de l'option PREDPROBS= 12
      2. Utilisation des options OUTEST= et INEST= 12
4. Les dernières nouveautés 13
   4.1. Derniers ajouts à la PROC LOGISTIC 13
   4.2. La procédure SURVEYSELECT : une ramification nouvelle 14
5. En cas de problème 14
   5.1. Eléments à transmettre au Support Clients 14
6. Liens utiles 15
   6.1. PROCEDURE LOGISTIC 15
   6.2. PROCEDURE SURVEYSELECT 15
7. Conclusion 15

1. INTRODUCTION THÉORIQUE

La procédure permet de réaliser des régressions logistiques aussi appelées modèle logit. Ces dernières sont des régressions dont la variable à expliquer est qualitative (binaire, ordinale ou nominale). Il s'agit d'un cas particulier d'un modèle linéaire généralisé. La relation entre la variable à expliquer et les variables explicatives n'est pas linéaire mais dépend d'une fonction, la fonction de lien. De ce fait, la probabilité, pour un modèle binaire avec une loi logistique de lien, de prendre la valeur 1 s'écrit :

$$p(1|X) = \frac{e^{b_0+b_1x_1+...+b_Jx_J}}{1+e^{b_0+b_1x_1+...+b_Jx_J}}$$

On remarque bien ici la non-linéarité du modèle.

2. RÉGRESSION LOGISTIQUE SOUS SAS

SAS offre plusieurs possibilités dans les dernières versions pour estimer ce genre de modèle mais la plus complète et répandue reste la procédure LOGISTIC.

Nous allons dans cette partie détailler les principales instructions de la procédure et leur rôle respectif, mais aussi continuer certaines explications théoriques en lien avec ces instructions.

2.1. Instruction CLASS : intégration des variables catégorielles

Considérons les données suivantes, correspondant à des malades soumis à un traitement A, B ou P (Placebo) pendant une certaine durée :

```
Data Maladie;
 input Trait $ Age Duree Gueri $ @@;
  datalines;
B 74 16 Non P 66 26 Oui A 71 17 Oui
A 62 42 Non P 74  4 Non P 70  1 Oui
B 66 19 Non B 59 29 Non A 70 28 Non
A 69  1 Non P 83  1 Oui B 75 30 Oui
P 77 29 Oui A 70 12 Non B 70  1 Non
B 67 23 Non A 76 25 Oui P 78 12 Oui
;
run ;
```

On souhaite expliquer la guérison du malade en fonction de son âge, du type et de la durée du traitement subi.

L'intégration, au niveau de la procédure LOGISTIC, de variables catégorielles, telles que la variable 'Trait', passe par l'utilisation de variables indicatrices.

2.1.1. Présentation des variables indicatrices

Rappelons qu'il est possible de remplacer une variable catégorielle à k modalités, par (k-1) variables indicatrices, codées chacune sur un chiffre.

Concrètement, si l'on considère la variable qualitative 'Trait' à trois modalités, A, B et P, celle-ci peut être remplacée par deux variables indicatrices au choix.

Si on garde 'P' comme valeur de référence, on peut choisir, par exemple, les variables indicatrices TraitA et TraitB, de la manière suivante :

TraitA= 1, si Trait='A'
0, sinon
et
TraitB= 1, si Trait='B'
0, sinon

Ces deux variables suffisent à exprimer les trois modalités 'A', 'B' et 'P', puisqu'on déduit la modalité 'P' si TraitA=0 et TraitB=0.

Les variables indicatrices prennent, ici, les valeurs classiques 0 et 1, mais d'autres valeurs sont tout à fait envisageables.

Depuis la version 8 (avant cela vous deviez créer vous-même manuellement ces variables indicatrices), la proc LOGISTIC s'est enrichie de l'instruction CLASS, qui permet de s'affranchir de cette étape. Les variables catégorielles sont directement prises en compte, sans manipulation préalable.

L'option PARAM, disponible dans cette procédure, propose plusieurs méthodes de création de variables indicatrices, à savoir :
EFFECT, GLM, ORTHPOLY, POLY et REF.
L'aide en ligne détaille chacune de ces méthodes ici.

Par défaut, la méthode est 'EFFECT', ce qui signifie que :
- la valeur de référence est la dernière modalité par ordre alphabétique.
- Les modalités sont codées avec des variables indicatrices prenant les valeurs 0 et 1 (et -1, pour la valeur de référence).

Exemple:

```
proc logistic data=Maladie;
 class trait;
 model gueri= trait age duree;
run;
```

On retrouve la façon dont les variables indicatrices sont créées dans le tableau '**Class Level Information**' de la sortie :

Class Level Information (Méthode EFFECT)

		Design Variables	
Class	Value	1	2
Trait	A	1	0
	B	0	1
	P	-1	-1

Si on examine ce tableau, deux variables indicatrices (design variables) ont été automatiquement créées, et P est la modalité de référence.

Les différentes méthodes proposées par l'option PARAM sont détaillées dans la documentation en ligne, Nous détaillerons seulement ici un cas, correspondant à une question fréquemment posée :

2.1.2. Comment changer la modalité de référence d'une variable catégorielle ?

Il est possible de choisir la modalité que l'on souhaite pour référence, grâce aux options PARAM= et REF= :

Exemple:

```
proc logistic data=maladie;
 class trait(param=ref ref= 'A');
 model gueri= trait age duree;
run;
```

Le tableau 'Class Level Information' alors obtenu est le suivant :

Class Level Information (Méthode REF)

		Design Variables	
Class	Value	1	2
Trait	A	0	0
	B	1	0
	P	0	1

La valeur de référence ('A') est, ici, codée en (0, 0) au lieu de (-1, -1) avec la méthode EFFECT.

2.1.3. Ecriture de l'équation logistique

La différence avec l'écriture d'une équation logistique sans variables catégorielles réside dans le fait que ce sont les variables indicatrices qui interviennent dans le modèle et pas directement la variable de classe.

De fait, ce sont les estimations des paramètres des variables indicatrices que l'on retrouve au niveau du tableau des paramètres estimés de la sortie :

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	28.9357	16.4530	3.0930	0.0786
Trait	A	1	0.5334	1.0439	0.2610	0.6094
Trait	B	1	1.8255	1.3165	1.9227	0.1656
Age		1	-0.3747	0.2202	2.8967	0.0888
Duree		1	-0.1145	0.0816	1.9711	0.1603

La lecture de ce tableau, permet l'écriture de l'équation logistique suivante :

Logit(p(Gueri)) = intercept + α_{TraitA}*TraitA + α_{TraitB}*TraitB + α_{Age}*Age + α_{Duree}*Duree,

où les α sont les 'Estimate' de la sortie.

Comme on le voit ci-dessus, l'écriture de l'équation dépend des variables indicatrices choisies : suivant ce choix, on obtient des équations différentes.
Cependant, il est important de noter que ces équations représentent un modèle identique, et que les prédictions qui en découlent sont également identiques quelle que soit l'écriture adoptée.

L'avantage d'utiliser l'instruction CLASS au niveau de la proc LOGISTIC est immédiat, puisque, comme nous l'avons vu, le code utilisé est synthétique et simplifié.
Au-delà de cela, le fait que la gestion des variables catégorielles se fasse au sein même de la procédure permet, maintenant, de réaliser des prédictions très facilement, par la simple utilisation des options OUTEST= et INEST=.

On estime une première table et on sort les estimateurs dans une table formatée que l'on nommera « A » via l'option OUTEST=. Ensuite toutes les tables similaires en termes de structure (par exemple, base sur les données de N, les tables des années N+1, N+2, au même format, sont estimables) peuvent être estimées avec les mêmes estimateurs en spécifiant INEST=A.

2.2. Instruction MODEL

Cette instruction reprend les normes des autres procédures statistiques de SAS. De ce fait, nous allons nous concentrer sur les points particuliers. Pour les points qui vous sembleraient non adressés, nous vous invitons à consulter la documentation en ligne sur ce sujet ici.

L'option EVENT= permet de choisir la modalité de la variable binaire à expliquer qui sera modélisée lors de la régression.

L'option ORDER= permet de choisir l'ordre de tri pour les niveaux de la variable réponse, selon l'ordre d'apparition dans la table d'entrée (DATA) par exemple. Il existe quatre choix.

La principale différence avec la procédure REG est l'option LINK=. Cette dernière permet de déterminer la fonction de lien entre les probabilités de la variable réponse et les prédicteurs linéaires. Par défaut, LINK=LOGIT.
Cette option permet aussi de choisir le type de régression logistique que l'on veut faire, il faut donc la changer seulement si on sait ce que l'on veut et pourquoi un changement est nécessaire.
Par ailleurs, pour les fonctions LOGIT et PROBIT, appartenant à la même famille de régression, même si les coefficients estimés sont différents, les probabilités associées sont identiques (la fonction de lien étant elle aussi différente).

Avec SAS®9, la procédure LOGISTIC a continué à s'enrichir en mettant à disposition deux nouvelles instructions SCORE et STRATA.

2.3. Instruction SCORE

Comme son nom l'indique, l'instruction SCORE permet le calcul des prédictions (scoring) d'après un modèle donné.

Il existe 2 méthodes pour utiliser l'instruction SCORE. Pour les illustrer, le jeu de données suivant sera utilisé dans un contexte de régression logistique généralisée :

```
data matable;
   input hopital traitement $ gravite $ @@;
   cards;
1 a none     1 a slight     1 a moderate
1 b none     1 b slight     1 b moderate
1 c none     1 c slight     1 c moderate
1 d none     1 d slight     1 d moderate
2 a none     2 a slight     2 a moderate
2 b none     2 b slight     2 b moderate
2 c none     2 c slight     2 c moderate
2 d none     2 d slight     2 d moderate
3 a none     3 a slight     3 a moderate
3 b none     3 b slight     3 b moderate
3 c none     3 c slight     3 c moderate
3 d none     3 d slight     3 d moderate
;
```

2.3.1. La première méthode – calcul du modèle et des prédictions dans une seule procédure LOGISTIC

La table utilisée pour élaborer le modèle est spécifiée dans l'option DATA= de la procédure LOGISTIC, alors que la nouvelle table « a » pour laquelle on souhaite obtenir les prédictions est, elle, spécifiée dans l'option DATA= de l'instruction SCORE.

```
proc logistic data=matable;
   class traitement hopital;
   model gravite(order=data) = traitement hopital / link=glogit;
   score data=a out=sortie;
   run;

proc print;
   run;
```

Avec la version 8, il était également possible de faire du scoring tout en élaborant le modèle, au sein d'une seule procédure LOGISTIC, mais il était pour cela nécessaire de concaténer la nouvelle table à la première.

2.3.2. Deuxième méthode – calcul du modèle et des prédictions dans des procédures LOGISTIC séparées

Dans un premier temps, le modèle est calculé et stocké dans la table définie par l'option OUTMODEL=

```
proc logistic data=matable outmodel=model;
   class traitement hopital;
   model gravite(order=data) = traitement hopital / link=glogit;
   run;
```

Le modèle stocké dans une table peut ensuite être appliqué à un nouveau jeu de données, grâce à l'option INMODEL= et à l'instruction SCORE :

```
proc logistic inmodel=model;
   score data=a out=out;
   run;

proc print;
   run;
```

2.3.3. Remarque

La table générée par l'option OUTMODEL, contenant les informations du modèle, est une table destinée uniquement au scoring au sein de la procédure LOGISTIC.

Pour utiliser les informations du modèle en dehors de la procédure LOGISTIC, les méthodes utilisées dans les versions antérieures sont toujours disponibles, notamment les options INEST et OUTEST.

2.4. Instruction STRATA

L'instruction STRATA permet de réaliser des régressions logistiques conditionnelles.

Les régressions logistiques conditionnelles sont beaucoup utilisées dans le cas d'études cas-témoins (case-control).
Au lieu de modéliser, par exemple, sur une population le fait d'être malade ou non par différents facteurs qualitatifs ou quantitatifs, on associe (apparie) à un cas malade un ou plusieurs cas témoins non-malades mais de profil similaire (même âge, même sexe,...). Cela donne lieu à une régression logistique par profil.
La procédure LOGISTIC permettait jusqu'à présent de traiter les cas d'appariement 1:1 (un cas pour un témoin). Elle est à présent, grâce à l'instruction STRATA, en mesure de traiter les cas de données appariées m:n (m cas pour n témoins).

Ainsi, en reprenant l'exemple évoqué plus haut, une régression logistique par profil âge pourra être réalisée en spécifiant la variable âge dans l'instruction STRATA :

```
data castemoin;
input age cas $ x1 x2;
datalines;
41 cas 3 5
41 temoin 4 9
41 temoin 1 4
42 temoin 3 9
42 cas 5 8
43 cas 7 9
43 temoin 2 3
43 temoin 4 7
43 temoin 2 4
…
;
run;

proc logistic data=castemoin;
class cas;
strata age;
model cas=x1 x2;
run;
```

L'instruction STRATA peut contenir autant de variables qu'il est nécessaire pour définir le profil désiré.

Un exemple complet de procédure LOGISTIC avec l'instruction STRATA est présenté dans la documentation en ligne ici.
Cet exemple est une étude de cas-témoins 1:1, mais il s'applique aussi à une étude cas-témoins m:n

A noter que la procédure PHREG permet également de réaliser les régressions logistiques conditionnelles m:n, mais la procédure LOGISTIC possède l'avantage de pouvoir utiliser les instructions CLASS et CONTRAST.

3. RÉGRESSION LOGISTIQUE GÉNÉRALISÉE, AUTRE POSSIBILITÉ OFFERTE PAR CETTE PROCÉDURE

3.1. Introduction

Depuis la version 8.2 de SAS, la procédure LOGISTIC permet, en plus des régressions logistiques binaires et ordinales, de réaliser des régressions logistiques généralisées (tout comme la procédure CATMOD).

La variable réponse est, dans ce cas, de type « nominale » et prend un nombre limité (>2) de valeurs. Dans ce document, un exemple, réalisé en version 9.4, sera détaillé de l'écriture du modèle jusqu'à l'obtention de prédictions.

3.2. Présentation de l'exemple à étudier

Ces données recensent les préférences que les enfants et les adolescents filles et garçons ont, en matière de sucreries.

```
data Confiserie;
      format Type $9.;
      input Sexe $ Age $ Type $ count @@;
      datalines;
   garçon  enfant chocolat  2   garçon  ado chocolat 10
   garçon  enfant caramel  13   garçon  ado caramel  19
   garçon  enfant bonbon    13  garçon  ado bonbon    3
   fille enfant chocolat    23  fille ado chocolat    6
   fille enfant caramel     3   fille ado caramel    14
   fille enfant bonbon      8   fille ado bonbon     16
   ;
run;
```

L'étude vise à exprimer le choix du type de sucrerie en fonction de l'âge et du sexe du sujet concerné. Ces données ne proviennent pas d'un questionnaire réel, mais ont été fabriquées pour les besoins de notre exemple.

3.3. Modélisation avec la proc LOGISTIC

3.3.1. Syntaxe de la proc LOGISTIC dans le cas d'une régression généralisée

La syntaxe qui est utilisée est la même que pour une régression binaire (variable réponse à deux modalités).
Par défaut, l'option LINK= de l'instruction MODEL est positionnée à LINK=LOGIT.
Pour accéder à la régression généralisée, il faut indiquer l'option LINK=GLOGIT.

```
   proc logistic data=Confiserie;
      freq count;
      class sexe(ref='fille') Age(ref='enfant') / param=ref;
      model type(ref='chocolat') = sexe Age / link=glogit;
   run;
```

D'après la table 'Confiserie', la variable réponse 'type' est nominale à trois modalités {bonbon, caramel, chocolat}.
Pour notre étude, c'est 'chocolat' qui a été choisi comme modalité de référence, grâce à l'option REF= de l'instruction MODEL. Si aucun choix n'avait été fait, la modalité située en dernière position dans l'ordre alphabétique aurait été la modalité de référence.
Les modalités de référence des variables explicatives catégorielles peuvent également être spécifiées par une option REF=, au niveau de l'instruction CLASS, comme c'est le cas ici, pour les variables sexe et âge.

3.3.2. Analyse de la sortie

Comme pour toutes les régressions, la sortie présente un bref tableau récapitulatif de l'étude menée, où apparaît notamment le type de la régression demandée : 'generalized logit' (logit généralisé en français), dans le cas présent.

Informations sur le modèle

Table	WORK.CONFISERIE
Variable de réponse	Type
Nombre de niveaux de réponse	3
Variable de fréquence	count
Modèle	logit généralisé
Technique d'optimisation	Newton-Raphson

Vient ensuite le tableau concernant le profil de la variable réponse 'Type' : on y retrouve ses trois modalités, et la modalité de référence :

Profil de réponse

Valeur ordonnée	Type	Fréquence totale
1	bonbon	40
2	caramel	49
3	chocolat	41

Les logits modélisés utilisent Type='chocolat' comme catégorie de référence

Juste après, le tableau 'Class Level Information' (Information sur les niveaux de classe) doit toujours être gardé en mémoire, puisqu'il indique les modalités de référence choisies pour chacune des variables de classe, et la façon dont ont été générées les variables indicatrices.

Informations sur les niveaux de classe

Classe	Valeur	Variables d'expérience
Sexe	**fille**	0
	garçon	1
Age	**ado**	1
	enfant	0

Les résultats asymptotiques qui suivent témoignent de la légitimité du modèle.

Le test global BETA=0 présente une p-value (Pr > ChiSq) inférieure à 0.05, ce qui signifie qu'au moins un des facteurs étudiés joue un rôle dans le choix du type de sucrerie.
La partie Analysis of Effects (Analyse des effets Type 3) indique que les deux effets sexe et âge entrent en considération dans le modèle (p-value > 0.05).

Etat de convergence du modèle

Critère de convergence (GCONV=1E-8) respecté.

Statistiques d'ajustement du modèle

Critère	Constante uniquement	Constante et Covariables
AIC	288.537	275.480
SC	294.272	292.685
-2 Log L	284.537	263.480

Test de l'hypothèse nulle globale : BETA=0

Test	Khi-2	DDL	Pr > Khi-2
Rapport de vrais	21.0577	4	0.0003
Score	19.8919	4	0.0005
Wald	17.5469	4	0.0015

Analyse des effets Type 3

Effet	DDL	Khi-2 de Wald	Pr > Khi-2
Sexe	2	12.2377	0.0022
Age	2	7.8266	0.0200

Les paramètres estimés apparaissent ensuite. Contrairement à la régression logistique binaire, on obtient plusieurs 'intercept' ainsi que plusieurs paramètres (un pour chaque modalité de la variable réponse, sauf pour la modalité de référence).
A côté de chaque variable de classe figure la modalité concernée.

Estimations par l'analyse du maximum de vraisemblance

Paramètre		Type	DDL	Estimation	Erreur type	Khi-2 de Wald	Pr > Khi-2
Intercept		**bonbon**	1	-0.3606	0.3461	1.0856	0.2975
Intercept		**caramel**	1	-1.2521	0.4216	8.8208	0.0030
Sexe	**garçon**	**bonbon**	1	0.5017	0.4736	1.1220	0.2895
Sexe	**garçon**	**caramel**	1	1.5977	0.4742	11.3522	0.0008
Age	**ado**	**bonbon**	1	0.3765	0.4534	0.6895	0.4063
Age	**ado**	**caramel**	1	1.2724	0.4685	7.3778	0.0066

Les estimations des Odds ratio (rapport de cotes) ci-dessous, nous permettent d'avancer qu'un garçon a environ 5 (4.942) fois plus de chances de choisir une confiserie de type 'caramel' plutôt que 'chocolat' par rapport à une fille.
De même, un adolescent a 3.5 (3.570) fois plus de chances de choisir une confiserie de type 'caramel' plutôt que 'chocolat', par rapport à un enfant.

Estimations des rapports de cotes

Effet	Type	Valeur estimée du point	95% Intervalle de confiance de Wald	
Sexe garçon vs fille	**bonbon**	1.651	0.653	4.178
Sexe garçon vs fille	**caramel**	4.942	1.951	12.518
Age ado vs enfant	**bonbon**	1.457	0.599	3.544
Age ado vs enfant	**caramel**	3.570	1.425	8.941

Les odds ratio se calculent habituellement par la formule suivante : exp(2*estimate).
Cependant, cette formule dépend de la paramétrisation choisie pour la variable de classe à expliquer. Ici, la méthode utilisée est 'REF' (reference cell coding), et la formule à appliquer est : exp(estimate).

3.3.3. Calcul des probabilités et prédiction

3.3.3.1. Obtention des probabilités prédites

Pour obtenir les probabilités de choisir chaque réponse, il suffit de préciser l'option PREDPROBS=I (Individual), au niveau de l'instruction OUTPUT, qui crée une table en sortie :

```
proc logistic data=Confiserie;
   freq count;
   class sexe(ref='fille') Age(ref='enfant') / param=ref;
   model type(ref='chocolat') = sexe Age / link=glogit;
   output out=out predprobs = I;
run;
proc print data=out;
run;
```

La table 'out' en sortie se présente comme suit :

Obs.	Type	Sexe	Age	count	_FROM_	_INTO_	IP_bonbon	IP_caramel	IP_chocolat
1	chocolat	garçon	enfant	2	chocolat	caramel	0.32306	0.39638	0.28056
2	chocolat	garçon	ado	10	chocolat	caramel	0.21732	0.65317	0.12951
3	caramel	garçon	enfant	13	caramel	caramel	0.32306	0.39638	0.28056
4	caramel	garçon	ado	19	caramel	caramel	0.21732	0.65317	0.12951
5	bonbon	garçon	enfant	13	bonbon	caramel	0.32306	0.39638	0.28056
6	bonbon	garçon	ado	3	bonbon	caramel	0.21732	0.65317	0.12951
7	chocolat	fille	enfant	23	chocolat	chocolat	0.35159	0.14416	0.50425
8	chocolat	fille	ado	6	chocolat	caramel	0.33460	0.33607	0.32932
9	caramel	fille	enfant	3	caramel	chocolat	0.35159	0.14416	0.50425
10	caramel	fille	ado	14	caramel	caramel	0.33460	0.33607	0.32932
11	bonbon	fille	enfant	8	bonbon	chocolat	0.35159	0.14416	0.50425
12	bonbon	fille	ado	16	bonbon	caramel	0.33460	0.33607	0.32932

Cette table reprend les données de départ, auxquelles viennent s'ajouter des variables automatiques :

- _FROM_: contient la valeur formatée de la réponse observée
- _INTO_ : contient la valeur formatée de la réponse prédite (correspondant à la plus forte probabilité)
- pour chaque modalité de la variable réponse, des variables _IP_, correspondant aux probabilités individuelles prédites.

3.3.3.2. Comment faire des prévisions à partir de nouvelles données ?

La procédure SCORE n'est pas utilisable dans le cas d'une régression multinomiale, mais il reste, malgré tout, deux solutions que nous allons développer :

1. Utilisation de l'option PREDPROBS=

Il est possible de faire de la prédiction en ajoutant simplement les nouvelles données, pour lesquelles la variable réponse est manquante, aux données servant à construire le modèle.
Les observations ainsi ajoutées ne seront pas utilisées pour la construction du modèle, comme en témoigne une note dans la sortie :

```
NOTE: 1 observation was deleted due to missing values for the response or explanatory variables.
```

Cependant, les prédictions seront calculées pour chacune d'entre elles. Ci-dessous, la prédiction calculée pour l'observation 13 est 'caramel', car la probabilité d'obtenir un caramel, pour un adolescent garçon (IP_caramel) est plus élevée que celle d'obtenir 'bonbon' ou 'chocolat'.

Prédiction, avec l'option PREDPROBS= :

Obs	Type	Sexe	Age	count	_FROM_	_INTO_	IP_bonbon	IP_chocolat	IP_caramel
1	chocolat	garçon	enfant	2	chocolat	caramel	0.32306	0.28056	0.39638
2	chocolat	garçon	ado	10	chocolat	caramel	0.21732	0.12951	0.65317
3	caramel	garçon	enfant	13	caramel	caramel	0.32306	0.28056	0.39638
4	caramel	garçon	ado	19	caramel	caramel	0.21732	0.12951	0.65317
5	bonbon	garçon	enfant	13	bonbon	caramel	0.32306	0.28056	0.39638
6	bonbon	garçon	ado	3	bonbon	caramel	0.21732	0.12951	0.65317
7	chocolat	fille	enfant	23	chocolat	chocolat	0.35159	0.50425	0.14416
8	chocolat	fille	ado	6	chocolat	caramel	0.33460	0.32932	0.33607
9	caramel	fille	enfant	3	caramel	chocolat	0.35159	0.50425	0.1441610
10	caramel	fille	ado	14	caramel	caramel	0.33460	0.32932	0.33607
11	bonbon	fille	enfant	8	bonbon	chocolat	0.35159	0.50425	0.14416
12	bonbon	fille	ado	16	bonbon	caramel	0.33460	0.32932	0.33607
13		garçon	ado	3		caramel	0.21732	0.12951	0.65317

2. Utilisation des options OUTEST= et INEST=

Pour cette deuxième solution, la prédiction se fait en deux temps :

- une première proc LOGISTIC est exécutée sur les données servant à fabriquer le modèle.
Les paramètres estimés du modèle sont stockés, sous forme de table, grâce à l'option OUTEST=

- une deuxième proc LOGISTIC réalise les prédictions sur de nouvelles données, en appliquant le modèle obtenu dans la première étape, grâce à l'option INEST=. L'option MAXITER est mise à zéro, afin que le modèle ne soit pas recalculé.

```
proc logistic data=Confiserie outest=outest;
   freq count;
   class sexe(ref='fille') Age(ref='enfant') / param=ref;
   model type(ref='chocolat') = sexe Age/ link=glogit;
   output out=out predprobs = I;
run;

proc logistic data=new inest=outest;
   class sexe(ref='fille') Age(ref='enfant') / param=ref;
   model type(ref='chocolat') = sexe Age / link=glogit maxiter=0;
   output out=newout predprobs = I;
run;
```

L'option MAXITER étant positionnée à 0, l'avertissement ci-dessous apparaît à la fois dans la log et dans la sortie :

```
Iteration limit reached without convergence.

WARNING: Convergence was not attained in 0 iterations. You may want to increase the maximum
number of iterations (MAXITER= option) or change the convergence criteria (ABSFCONV=, FCONV=,
GCONV=, XCONV= options) in the MODEL statement.

WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based
on the last maximum likelihood iteration. Validity of the model fit is questionable.
```

Cependant, le système poursuit le traitement, en se servant des paramètres estimés fournis (INEST=) pour calculer les prédictions.

Précautions d'emploi :

- Cette méthode doit être exclusivement utilisée pour calculer les probabilités prédites : toutes les statistiques basées sur la matrice de covariance, comme les intervalles de confiance des probabilités prédites, seront incorrectes, puisque la matrice de covariance du modèle d'origine n'est pas utilisée.

- En toute logique, les observations de la table, sur laquelle doivent se faire les prédictions (`new` dans notre exemple), comportent une variable réponse non renseignée.
Dans la pratique, il s'avère que, pour que la prédiction fonctionne, il est obligatoire que cette table contienne au moins quelques observations ayant une variable réponse renseignée.

4. LES DERNIÈRES NOUVEAUTÉS

4.1. Derniers ajouts à la PROC LOGISTIC

L'instruction EFFECT, dont la documentation en ligne est disponible ici, permet de construire des modèles plus riches que ce que l'instruction CLASS permet.
Tandis que l'instruction EFFECTPLOT produit un affichage graphique des modèles estimés.

La procédure est capable d'utiliser la méthode du maximum de vraisemblance pénalisée de Firth.

Bien entendu, d'autres améliorations sont aussi présentes dans d'autres instructions avec l'ajout de statistiques ou options supplémentaires. Nous vous invitons à consulter la page des nouveautés pour connaître l'ensemble de celles-ci pour le module ou domaine qui vous intéresse.

4.2. La procédure SURVEYSELECT : une ramification nouvelle

Les variables réponses catégorielles binaires, ordinales ou nominales sont très fréquentes dans les données d'enquête. L'analyse logistique est très souvent utilisée pour trouver les relations entre ces réponses discrètes et des variables explicatives.
La procédure LOGISTIC du module SAS/STAT permet de réaliser une analyse logistique sur des données provenant d'un échantillon aléatoire. Mais elle ne peut être utilisée dans le cas d'échantillons complexes collectés selon un plan d'enquête, avec éventuellement des stratifications, des classifications, ou/et des poids inégaux. En effet, la procédure LOGISTIC calcule les statistiques sous l'hypothèse que l'échantillon est issu d'une population infinie par simple tirage aléatoire, ce qui n'est pas toujours vérifié. Cela affecte alors les erreurs standards et du même coup tout ce qui est calculé à partir des fonctions de la matrice de covariance des paramètres.

Tout comme la procédure LOGISTIC, la procédure SURVEYLOGISTIC calcule les paramètres estimés par la méthode du maximum de vraisemblance, mais l'estimation de la variance se fait de façon différente. L'information concernant la structure de l'échantillon, les strates, les classes et les poids est incorporée (« Taylor expansion approximation »). Un ajustement est également fait pour réduire le biais en cas de petit échantillon, et le facteur de correction correspondant à une population finie est inclus si l'échantillon est sans remise.
La syntaxe de la procédure SURVEYLOGISTIC reprend en grande partie la syntaxe de la procédure LOGISTIC. Les instructions STRATA, CLUSTER et WEIGHT, et les options de la procédure RATE et TOTAL permettent de définir la structure de l'échantillon.

- L'instruction STRATA désigne la ou les variable(s) caractère(s) ou numérique(s) qui forme(nt) les strates des données. Si la variable est formatée, ce sont les valeurs formatées qui définissent les niveaux des strates. En ajoutant l'option LIST, un résumé sur la composition des strates sera donné.
- L'instruction CLUSTER nomme la ou les variable(s) qui forme(nt) les clusters de l'échantillon. S'il existe des strates, les clusters sont emboîtés dedans.
CLUSTER et STRATA peuvent apparaître plusieurs fois.
- L'instruction WEIGHT indique le poids affecté à chaque observation. C'est un nombre positif.
- L'option RATE est un nombre entre 0 et 1 (ou 1 et 100) indiquant le taux d'échantillonnage, ou une table de données avec une variable _RATE_ indiquant le taux d'échantillonnage de chaque strate.
- L'option TOTAL est le nombre total d'unités d'échantillonnage ou une table de données avec une variable _TOTAL_ indiquant le nombre total d'unités d'échantillonnage dans chaque strate.
Si l'une des deux options RATE ou TOTAL est utilisée, le facteur de correction pour une population finie (fcp) est appliqué.

Une illustration basée sur les données de l'exemple 98.1 de la documentation en ligne, présent ici, permet de bien comprendre le fonctionnement et le but de cette procédure

5. EN CAS DE PROBLÈME

5.1. Eléments à transmettre au Support Clients

Si vous rencontrez des problèmes lors de l'utilisation, vous pouvez nous écrire à support@sas.com, en attachant à votre message l'erreur reçue, mais aussi le programme et une table pour reproduire le plus fidèlement votre problématique.

6. LIENS UTILES

6.1. PROCEDURE LOGISTIC

La procédure LOGISTIC a été traitée dans de nombreuses ressources internes en français :
http://www.sas.com/offices/europe/france/services/support/faq/sasstat.html

6.2. PROCEDURE SURVEYSELECT

Les plans d'enquête supportés par la procédure SURVEYLOGISTIC sont indiqués dans la FAQ suivante :
What sample designs are supported by the survey procedures, and how do I code them?
http://support.sas.com/faq/039/FAQ03965.html

Pour obtenir les probabilités prédites, il faut utiliser la procédure LOGISTIC, comme l'indique la FAQ suivante :
How can I get predicted values from my model fit with PROC SURVEYLOGISTIC?
http://support.sas.com/faq/044/FAQ04490.html

7. CONCLUSION

La procédure LOGISTIC, ancêtre chez SAS, permet de réaliser toutes sortes de régressions logistiques. Cependant, malgré son statut d'ancienne procédure, celle-ci continue toujours à évoluer pour intégrer les nouveautés découvertes dans ce domaine.

Jérémy NOEL
Consultant Support Clients SAS France