



La régression logistique généralisée avec la procédure LOGISTIC

The Power to Know.™

Sommaire

<i>I / Régression logistique généralisée</i> _____	3
a. Introduction _____	3
b. Présentation de l'exemple à étudier _____	3
<i>II / Modélisation avec la proc LOGISTIC</i> _____	4
a. Syntaxe de la proc LOGISTIC dans le cas d'une régression généralisée _____	4
b. Analyse de la sortie _____	4
<i>III / Ecriture des équations logistiques généralisées</i> _____	8
a. Enoncé des formules: _____	8
b. Ecriture des équations logistiques de notre exemple: _____	8
<i>IV / Calcul des probabilités et prédiction:</i> _____	9
a. Obtention des probabilités prédites: _____	9
b. Retrouver les probabilités prédites, par le calcul: _____	9
c. Comment faire des prévisions à partir de nouvelles données ? _____	11
1. Utilisation de l'option PREDPROBS= : _____	11
2. Utilisation des options OUTEST= et INEST=: _____	12
3. Le scoring avec SAS® System 9: _____	13

La régression logistique généralisée avec la procédure LOGISTIC

I / Régression logistique généralisée

a. 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 8.2, sera détaillé de l'écriture du modèle jusqu'à l'obtention de prédictions. Un bref aperçu de la proc LOGISTIC dans **SAS® System 9** sera également donné.

b. 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.

II / Modélisation avec la proc LOGISTIC

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

La syntaxe que 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, c'est la modalité située en dernière position dans l'ordre alphabétique qui 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 age.

b. 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', dans le cas présent.

①	The LOGISTIC Procedure	
	Model Information	
	Data Set	WORK.CONFISERIE
	Response Variable	Type
	Number of Response Levels	3
	Number of Observations	12
	Frequency Variable	count
	Sum of Frequencies	130
	Model	generalized logit
	Optimization Technique	Fisher's scoring

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

②

Response Profile		
Ordered Value	Type	Total Frequency
1	bonbon	40
2	chocolat	41
3	caramel	49

Logits modeled use Type='chocolat' as the reference category.

Juste après, le tableau ‘Class Level Information’ 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. (Vous trouverez plus de renseignements dans l’article Allo support N°10, intitulé ‘La gestion des variables catégorielles dans la proc LOGISTIC’).

③

Class Level Information		
Class	Value	Design Variables
		1
Sexe	filles	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** indique que les deux effets sexe et âge entrent en considération dans le modèle (p-value > 0.05).

④

Model Convergence Status		
Convergence criterion (GCONV=1E-8) satisfied.		
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates

AIC	288.537	275.480	
SC	294.272	292.685	
-2 Log L	284.537	263.480	
Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.0577	4	0.0003
Score	19.8919	4	0.0005
Wald	17.5469	4	0.0015
Type III Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
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.

⑤

Analysis of Maximum Likelihood Estimates							
Parameter	Type	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	
Intercept	bonbon	1	-0.3606	0.3461	1.0856	0.2975	
Intercept	caramel	1	-1.2521	0.4216	8.8208	0.0030	
Sexe	garçon	1	0.5017	0.4736	1.1220	0.2895	
Sexe	garçon	1	1.5977	0.4742	11.3522	0.0008	
Age	ado	1	0.3765	0.4534	0.6895	0.4063	
Age	ado	1	1.2724	0.4685	7.3778	0.0066	

Les estimations des **Odds ratio** 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.

⑥

Odds Ratio Estimates					
Effect	Type	Point Estimate	95% Wald Confidence Limits		
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 : **$\exp(2*\text{estimate})$** .
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: **$\exp(\text{estimate})$** .

III / Ecriture des équations logistiques généralisées

a. Enoncé des formules:

La formule générale des équations généralisées est donnée ci-dessous :

$$\log \frac{p_i}{p_{k+1}} = \beta_i' X, \quad i \in \{1, \dots, k+1\},$$

où $k+1$ est le nombre de modalités de la variable Réponse.

b. Ecriture des équations logistiques de notre exemple:

Ce sont les paramètres estimés, obtenus dans la sortie ⑤ précédente, qui vont permettre d'écrire les équations logistiques du modèle obtenu. Ces mêmes paramètres peuvent être récupérés dans une table SAS, grâce à l'option OUTEST=.

La variable réponse 'type' ayant trois modalités, avec 'chocolat' comme modalité de référence, nous obtenons les 2 équations suivantes:

$$\text{Log}([\text{Pr}(\text{Type}=\text{bonbon}) / \text{Pr}(\text{Type}=\text{chocolat})]) = -0.3606 + 0.5017 * (\text{Sexe}=\text{garçon}) + 0.3765 * (\text{Age}=\text{ado})$$

$$\text{Log}([\text{Pr}(\text{Type}=\text{caramel})/\text{Pr}(\text{Type}=\text{chocolat})]) = -1.2521 + 1.5977 * (\text{Sexe}=\text{garçon}) + 1.2724 * (\text{Age}=\text{ado})$$

Ces équations logistiques s'adaptent, pour chaque combinaison des variables explicatives:

Pour Sexe=garçon et Age=ado:

$$\text{Log}([\text{Pr}(\text{Type}=\text{bonbon}) / \text{Pr}(\text{Type}=\text{chocolat})]) = -0.3606 + 0.5017 + 0.3765 = 0.5176$$

$$\text{Log}([\text{Pr}(\text{Type}=\text{caramel})/\text{Pr}(\text{Type}=\text{chocolat})]) = -1.2521 + 1.5977 + 1.2724 = 1.618$$

Pour Sexe=garçon et Age=enfant :

$$\text{Log}([\text{Pr}(\text{Type}=\text{bonbon}) / \text{Pr}(\text{Type}=\text{chocolat})]) = -0.3606 + 0.5017 = 0.1411$$

$$\text{Log}([\text{Pr}(\text{Type}=\text{caramel})/\text{Pr}(\text{Type}=\text{chocolat})]) = -1.2521 + 1.5977 = 0.3456$$

Pour Sexe=filles et Age=ado:

$$\text{Log}([\text{Pr}(\text{Type}=\text{bonbon}) / \text{Pr}(\text{Type}=\text{chocolat})]) = -0.3606 + 0.3765 = 0.0159$$

$$\text{Log}([\text{Pr}(\text{Type}=\text{caramel})/\text{Pr}(\text{Type}=\text{chocolat})]) = -1.2521 + 1.2724 = 0.0203$$

Pour Sexe=filles et Age=enfant:

$$\text{Log}([\text{Pr}(\text{Type}=\text{bonbon}) / \text{Pr}(\text{Type}=\text{chocolat})]) = -0.3606$$

$$\text{Log}([\text{Pr}(\text{Type}=\text{caramel})/\text{Pr}(\text{Type}=\text{chocolat})]) = -1.2521$$

IV / Calcul des probabilités et prédiction:

a. 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_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

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.

Une question qui revient souvent est de savoir comment retrouver, par le calcul, les probabilités prédites.

b. Retrouver les probabilités prédites, par le calcul:

La résolution des équations logistiques généralisées, nous amène à la formule des probabilités prédites suivante:

$$p_i = \frac{\exp(\beta_i' X)}{\sum_i \exp(\beta_i' X)}, \quad i \in \{1, \dots, k+1\},$$

où $k+1$ est le nombre de modalités de la variable Réponse

β_{k+1} étant considéré à zéro (vecteur zéro), on en déduit:

$$p_{k+1} = \frac{1}{\sum_i \exp(\beta_i' X)}$$

Pour obtenir les probabilités prédites, il suffit donc d'appliquer ces formules, pour chaque combinaison de variables explicatives.

Ainsi,

pour Sexe=garçon et Age=ado, on obtient:

$$\Pr(\text{Type}=\text{bonbon}) = \frac{\exp(0.5176)}{1 + \exp(0.5176) + \exp(1.618)} = 0.21733$$

$$\Pr(\text{Type}=\text{caramel}) = \frac{\exp(1.618)}{1 + \exp(1.618) + \exp(0.5176)} = 0.65315$$

$$\begin{aligned} \Pr(\text{Type}=\text{chocolat}) &= 1 - \Pr(\text{Type}=\text{caramel}) - \Pr(\text{Type}=\text{bonbon}) \\ &= \frac{1}{1 + \exp(0.5176) + \exp(1.618)} = 0.12952 \end{aligned}$$

Plus facilement, cela peut se coder de la façon suivante :

```
data prob(drop=logbsurc logssurc count);
  set out;
  /* Première équation logistique: */
  logbsurc = -0.3606 + 0.5017 * (Sexe='garçon') + 0.3765 * (Age='ado');

  /* Deuxième équation logistique: */
  logssurc = -1.2521 + 1.5977 * (Sexe='garçon') + 1.2724 * (Age='ado');

  /* Calcul des probabilités prédites: */
  PCalc_bonbon = exp(logbsurc)/(1+exp(logbsurc)+exp(logssurc));
  PCalc_caramel = exp(logssurc)/(1+exp(logbsurc)+exp(logssurc));
  PCalc_chocolat=1-PCalc_bonbon-PCalc_caramel;
run;

title 'Comparaison entre les probabilités obtenues par l'option PREDPROBS=
et celles calculées manuellement';
proc print data=prob;
run;
```

On vérifie ainsi aisément que les probabilités calculées à partir du modèle obtenu sont identiques (aux différences d'arrondis près) à celles calculées par l'option PREDPROBS=I :

Comparaison entre les probabilités obtenues par l'option PREDPROBS=et celles calculées manuellement											
Obs	Type	Sexe	Age	_FROM_	_INTO_	IP_bonbon	IP_chocolat	IP_caramel	PCalc_ bonbon	PCalc_ caramel	PCalc_ chocolat
1	chocolat	garçon	enfant	chocolat	caramel	0.32306	0.28056	0.39638	0.32307	0.39638	0.28055
2	chocolat	garçon	ado	chocolat	caramel	0.21732	0.12951	0.65317	0.21733	0.65315	0.12952
3	caramel	garçon	enfant	caramel	caramel	0.32306	0.28056	0.39638	0.32307	0.39638	0.28055
4	caramel	garçon	ado	caramel	caramel	0.21732	0.12951	0.65317	0.21733	0.65315	0.12952
5	bonbon	garçon	enfant	bonbon	caramel	0.32306	0.28056	0.39638	0.32307	0.39638	0.28055
6	bonbon	garçon	ado	bonbon	caramel	0.21732	0.12951	0.65317	0.21733	0.65315	0.12952
7	chocolat	filles	enfant	chocolat	chocolat	0.35159	0.50425	0.14416	0.35159	0.14417	0.50425
8	chocolat	filles	ado	chocolat	caramel	0.33460	0.32932	0.33607	0.33460	0.33608	0.32932
9	caramel	filles	enfant	caramel	chocolat	0.35159	0.50425	0.14416	0.35159	0.14417	0.50425
10	caramel	filles	ado	caramel	caramel	0.33460	0.32932	0.33607	0.33460	0.33608	0.32932
11	bonbon	filles	enfant	bonbon	chocolat	0.35159	0.50425	0.14416	0.35159	0.14417	0.50425
12	bonbon	filles	ado	bonbon	caramel	0.33460	0.32932	0.33607	0.33460	0.33608	0.32932

c. 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 une 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	filles	enfant	23	chocolat	chocolat	0.35159	0.50425	0.14416
8	chocolat	filles	ado	6	chocolat	caramel	0.33460	0.32932	0.33607
9	caramel	filles	enfant	3	caramel	chocolat	0.35159	0.50425	0.1441610
10	caramel	filles	ado	14	caramel	caramel	0.33460	0.32932	0.33607
11	bonbon	filles	enfant	8	bonbon	chocolat	0.35159	0.50425	0.14416
12	bonbon	filles	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 lancée 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='filles') 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='filles') 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 (ABSFCNV=, 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.

3. Le scoring avec SAS® System 9:

A partir de SAS® System 9, quelques modifications de syntaxe apparaissent dans la proc LOGISTIC.

Les prédictions peuvent maintenant être obtenues grâce à l'instruction **SCORE** et les paramètres du modèle sont désormais stockés, grâce à l'option **OUTMODEL=**.

```
proc logistic data=Confiserie outmodel=sasuser.ConfModel;  
  freq count;  
  class sexe(ref='fille') Age(ref='enfant') / param=ref;  
  model type(ref='chocolat') = sexe Age / link=glogit;  
  score out=Score1;  
run;
```

Le modèle est réutilisable grâce à l'option **INMODEL=** pour calculer les prédictions sur de nouvelles données:

```
proc logistic inmodel=sasuser.ConfModel;  
  score data=Confiserie out=Score2;  
run;
```

Pour plus de renseignements, un exemple est présenté dans la documentation en ligne de SAS® System 9 (Exemple 40.13: Scoring Data Sets with the SCORE Statement).

Depuis la version 8, l'utilisation de la procédure LOGISTIC tend à se simplifier: les variables de classe peuvent maintenant être traitées à l'intérieur de la procédure, grâce à l'instruction CLASS, et des régressions logistiques généralisées peuvent être menées sur des variables réponses nominales. L'aperçu de la proc LOGISTIC sous SAS® System 9 évolue dans ce sens, en proposant une syntaxe de plus en plus intuitive et simplifiée.

Blandine Colas
Ingénieur Consultant



SAS France
Domaine de Grégy - BP 5
77166 Grégy-sur-Yerres
Tél. : 01 60 62 11 11
Fax : 01 60 62 11 99

SAS Europe, Middle East & Africa
P.O. Box 10 53 40
Neuenheimer Landstr. 28-30
D-69043 Heidelberg, Germany
Tel: +49 6221 4160, Fax: +49 6221 474850

SAS, le Système SAS® sont les marques déposées de SAS Institute Inc., Cary NC, USA.
Les autres noms de produits ou concepts sont des marques déposées des sociétés respectives.

www.sas.com/france