Analyse statistique de données qualitatives et quantitatives en sciences sociales : TP

RÉGRESSION LOGISTIQUE (MODÈLES CHAPITRE 1)

Modèles de régression logistique à réaliser

- Une explicative catégorielle
  - ➢Evaldemo2 en fonction de educ3
- Deux explicatives catégorielles
  - ➢Evaldemo2 en fonction de educ3 + hincfel2
  - ➢Evaldemo2 en fonction de hincfel2 + plinsoc3
- Une explicative continue
  - ➢Evaldemo2 en fonction de agea
- Explicatives mixtes
  - ➢Evaldemo2 en fonction de agea + polintr2

- $\log\left(\frac{\pi_k}{1-\pi_k}\right) = \theta_k = \mu + \alpha_k \;\; ; (\alpha_1 = 0)$
- $\log\left(\frac{\pi_{jk}}{1-\pi_{jk}}\right) = \theta_{jk} = \mu + \alpha_j + \beta_k + \gamma_{jk}$;
  $(\alpha_1 = \beta_1 = \gamma_{1k} = \gamma_{j1} = 0)$

- $\log\left(\frac{\pi_x}{1-\pi_x}\right) = \beta_0 + \beta_1 x$
- $\log\left(\frac{\pi_x}{1-\pi_x}\right) = (\beta_0 + \alpha_k) + (\beta_1 + \tau_k)x$;
  $(\alpha_1 = \tau_1 = 0)$

Régression logistique avec une explicative catégorielle

/* A partir de la base de données – observations en lignes et variables en colonnes */

proc logistic data = *SAS-dataset*;

class *explicative catégorielle* (ref = '*niveau de référence*') /param=ref;

model *réponse*(event= '*niveau de Succès*') = *explicative* /expb clodds = WALD;

run;

/* Alternative : à partir d'un tableau de contingence*/

proc logistic data = *SAS-dataset*; *Le SAS-dataset renvoie ici aux fréquences de la table de contingence;

class *explicative catégorielle* (ref = '*niveau de référence*') /param=ref;

weight count;

model *réponse*(event= '*niveau de Succès*') = *explicative* /expb clodds = WALD;

run;

L'instruction CONTRAST (si explicative à plus de 2 niveaux)

En pratique :

- Faire tourner une première fois la proc logistic sans insérer l'instruction « contrast » et examiner la sortie « Informations sur les niveaux de classe ».
- Nécessité de comprendre l'équation du modèle : $\log\left(\frac{\pi_k}{1-\pi_k}\right) = \theta_k = \mu + \alpha_k \;\; ; \; (\alpha_1 = 0)$

Class Level Information			
Class	Value	Design Variables	
educ3	High	1	0
	Low	0	1
	Middle	0	0

La sortie ci-contre doit être interprétée comme suit :

- Modèle pour « Middle » : $\log\left(\frac{\pi_1}{1-\pi_1}\right) = \theta_1 = \mu + \alpha_1 = \mu$
- Modèle pour « High » : $\log\left(\frac{\pi_2}{1-\pi_2}\right) = \theta_2 = \mu + \alpha_2$
- Modèle pour « Low » : $\log\left(\frac{\pi_3}{1-\pi_3}\right) = \theta_1 = \mu + \alpha_3$

- α_2 représente le contraste « High Vs Middle » et α_3 représente le contraste « Low Vs Middle »
- Si je souhaite à présent définir les contrastes suivants :
  - « Middle Vs High » : $-\alpha_2$
  - « Low Vs Middle » : α_3
  - « Low Vs High » : $\alpha_3 - \alpha_2$

```
contrast 'Middle Vs High' educ3 -1 0 /e estimate = both;
contrast 'Low Vs Middle'  educ3  0 1 /e estimate = both;
contrast 'Low Vs High'    educ3 -1 1/e estimate = both;
```

Régression logistique avec deux explicatives catégorielles

- Point de départ : on teste l'effet des deux explicatives et de leur interaction sur la réponse

proc logistic data = *SAS-dataset*;

class *expcat1* (ref = ' *niveau-référence* ') *expcat2* (ref = '*niveau-référence*')/param=ref;

model *réponse*(event= '*niveau de Succès*') = *exp1 exp2 exp1*exp2*/expb clodds = WALD;

run;

- Examiner la sortie « Type 3 Analysis of effects » pour procéder à la sélection du modèle.

Attention TYPE 3 ⇔ évalue le caractère significatif de l'effet d'un terme en contrôlant tous les autres simultanément (que ceux-ci soient inclus avant ou après dans l'équation) → le retrait d'un terme affecte les statistiques du Chi^2 et les p-valeurs calculées pour tous les termes restants !!!

Sélection du modèle

- Si le terme d'interaction est significatif, on conserve le modèle complet pour l'interprétation des paramètres; sinon, on supprime le terme d'interaction du modèle et on refait tourner le code :

proc logistic data = *SAS-dataset*;

class *expcat1* (ref = ' *niveau-référence* ') *expcat2* (ref = '*niveau-référence*')/param=ref;

model *réponse*(event= '*niveau de Succès*') = *exp1 exp2* /expb clodds = WALD;

run;

- Si les deux explicatives ont un effet significatif, on conserve ce modèle pour interpréter les paramètres; sinon, on reteste l'effet de chaque explicative séparément (cf. modèle 1).

Régression logistique avec une explicative continue

- 1ère étape : analyse exploratoire

Le modèle suivant n'a de sens que si l'hypothèse de linéarité est raisonnable

$$\log(\text{Cote}_x) = \text{logit}\pi_x = \log\left(\frac{\pi_x}{1-\pi_x}\right) = \beta_0 + \beta_1 x$$

- Objectif = visualiser au moyen d'un graphe si le logarithme de la cote évolue de manière linéaire avec la variable explicative continue (ici, l'âge). Si le graphe met clairement en évidence une association non linéaire, alors le modèle n'est pas adéquat (→ alternative : catégoriser la variable en classes sur base de l'observation du graphe).
- Pour faire le graphe X (expcont) Y (log(cote)):
  - Si effectif élevé → conserver le caractère continu de la variable (cf. code TP)
  - Si effectif insuffisant que pour pouvoir estimer les log(cote) pour chaque valeur (ou presque) de la variable explicative → catégorisation de l'explicative (cf. code cours théorique)

Régression logistique avec une explicative continue

- Si le graphe ne suggère pas une relation non linéaire, mise en œuvre du modèle :

proc logistic data = *SAS-dataset*;

model *réponse*(event= '*niveau de Succès*') = *explicative_continue* /expb clodds = WALD;

run;

/* L'instruction « class » disparaît car il n'y a plus d'explicative catégorielle dans le modèle*/

Régression logistique avec des explicatives mixtes

Procédure : idem supra

- Point de départ : on teste l'effet des deux explicatives et de leur interaction sur la réponse

proc logistic data = *SAS-dataset*;

class expcat (ref = ' *niveau-référence* ')/ param = ref;

model *réponse*(event= '*niveau de Succès*') = *expcont expcat expcont*expcat /*

expb clodds = WALD;

run;

- Examiner la sortie « Type 3 Analysis of effects » pour procéder à la sélection du modèle.
- Après avoir sélectionné le modèle → examen des estimations + IC

Alternative à la proc logistic : PROC GENMOD – pour tous les modèles vus supra

Advantage: can request likelihood ratio tests using options <type1> or <type3>.

proc genmod data = *SAS-dataset* <descending>;

class expcat (ref = ' *niveau-référence* ')/ param = ref;

model *réponse* = *exp1 exp2 exp1*exp2*/**dist = bin link = logit** lrci type1 type3;

run;

Régression logistique

/* <descending> force success=‘the reverse of the default' */