Scénario : exempes de régression logistique

Résumé

Ce scénation présente trois exemples de régression logistique ou binomiale en introduisant les questions de choix de modèle traitées avec SAS ou R. L'objectif est donc de construire un modèle expliquant et / ou prévoyant prévision une variable dichotomique ou encore, de façon plus générale, qualitative ordinale. Différents modèles sont abordés selon le niveau d'interaction considéré entre les variables explicatives quantitatives ou qualitatives. Trois exemples sont traités pour illustrer les fonctionnalités des procédures de SAS et les fonctions R. Le premier exemple élémentaire analyse des données de cancer de la prostate, le deuxième une enquête sur les travail des femmes, le dernier les conséquences d'un accident de voiture selon trois facteurs.

1 Cancer de la prostate avec SAS

1.1 Données

Il y a quelques années, le traitement du cancer de la prostate dépendait de son extension ou non au niveau des ganglions du système lymphatique. Afin d'éviter une intervention chirurgicale (laparotomie) pour vérifier la contamination, des études ont tenté de la prévoir à partir de l'observation de variables explicatives. Dans ce but, 5 variables ont été observées sur 53 patients atteints d'un cancer de la prostate et sur lesquels une laparotomie a été réalisée afin de s'assurer de l'implication ou non du système lymphatique. Ces données sont extraites de Collet (1991).

Les variables considérées sont les suivantes :

âge du patient

acid niveau de "serum acid phosphatase",

radio résultat d'une analyse radiographique (0 : négatif, 1 : positif),

taille taille de la tumeur (0 : petite, 1 : grande),

gravite résultat de la biopsie (0 : moins sérieux, 1 : sérieux).

lymph La sixième variable indique l'implication (1) ou non (0) du système lymphatique.

L'objectif est donc prédictif (variable `lymph`).

Le fichier de données `prostate.dat` est accessible sur la page d'URL : `http://wikistat/data/`

```
/* Lire les données */
data sasuser.prost;
infile 'prostate.dat' dlm='09'x;
input age acid radio $ taille $ gravite $ lymph;
l_acid=log(acid);
run;
```

1.2 Traitements interactifs

Descriptif

Étudier (rapidement avec SAS/insight) les distributions des variables explicatives, vérifier que la distribution de la variable "acid" justifie une transformation par une fonction log.

```
Solutions > Analyse de données interactives >
          sasuser > prost
```

Modèle complet

Estimer un modèle binomial ou de régression logistique :

```
Analyse > fit
Sélect. lymph > Y
/* Attention Y doit être de type réel
   (interval donc 0,1) pour cette procédure */
Sélect. l_acid age radi o taille gravite
Expand  /*toutes les interactions d'ordre 2 dans X
Sélect. age*age l_acid*l_acid age*l_acid > Remove
Method > Binomial /* régression logistique */
```

```
/* Au lieu de Normal : modèle gaussien ou
   régression multilinéaire*/
Link function > Canonical
/* fonction logit dans le cas binomial */
Output > TypeIII(LR) Tests
OK > Apply
```

Le modèle est estimé et les tableaus de type III fournissent les statistiques des tests de Wald et du rapport de vraisemblance sur la significativité des paramètres du modèle.

Choix de modèle

À partir du modèle complet incluant les interactions d'ordre 2, mettre en œuvre une procédure de sélection par élimination en respectant les règles suivantes :

- ne supprimer un effet principal qu'à la condition qu'il n'intervienne plus dans des interactions,
- ne supprimer qu'un terme à la fois,
- utiliser conjointement les critères fournis par la décomposition (type III) du test de Wald et du test de rapport de vraisemblance pour choisir le facteur à éliminer.

Choisir, parmi les interactions ou effets principaux, celui pour lequel le test de Wald ($H_0 : b_j = 0$) (resp. le test du rapport de vraisemblance) est le moins significatif, c'est-à-dire avec la plus grande "prob value". Le retirer du modèle et recalculer l'estimation. Il suffit pour cela de sélectionner le nom de la variable ou de l'interaction dans le tableau (TYPE III) et d'exécuter la commande

```
Edit > delete
```

Choix

À l'issue de la sélection, deux modèles restent en compétition celui meilleur au sens du test de Wald et celui au sens du test du rapport de vraisemblance. Comment choisir parmi ces deux modèles ?

Remarque : actuellement, une simple échographie permet de délivrer un diagnostic avec beaucoup plus de fiabilité.

1.3 Les autres procédures

Retrouver automatiquement le même modèle à l'aide de la procédure `logistic` :

```
proc logistic data=sasuser.prost;
class radio taille gravite ;
model lymph = age l_acid radio taille gravite
   age*taille age*radio age*gravite
   l_acid*radio l_acid*taille l_acid*gravite
   radio*taille radio*gravite taille*gravite
   /selection=backward ;
run;
```

Comparer les estimations des paramètres obtenus par les deux modèles. Ceux-ci diffèrent. Pourquoi ?

Comparer avec les résultats de la procédure `genmod` :

```
proc genmod data=sasuser.prost;
class radio taille gravite ;
model lymph = age l_acid radio taille gravite
      age*taille age*radio age*gravite
      l_acid*radio l_acid*taille l_acid*gravite
      radio*taille radio*gravite taille*gravite
      / dist=bin type3;
run;
```

Commenter les effets des variables en notant les valeurs prises par les paramètres.

2 Panel d'enquête avec R

2.1 Les données

Les données (Jobson 1992) étudiées sont issues d'une enquête réalisée auprès de 200 femmes mariées du Michigan. Les variables considérées sont les suivantes : `THISYR`, la variable à expliquer, (Woui) si la femme travaille l'année en cours, (Wnon) sinon ; `CHILD1` code la présence (Boui) ou l'absence

(Bnon) d'un enfant de moins de 2 ans ; CHILD2 présence (Eoui) ou absence (Enon) d'un enfant entre 2 et 6 ans ; ASCEND l'ascendance noire (Anoi) ou blanche (Abla) ; les autres variables, âge (AGE), nombre d'années d'études (EDUC), revenu du mari (HUBINC) sont quantitatives.

Les données sont disponibles dans le fichier jobpanel.dat. Lire le fichier puis recoder les facteurs :

```
# Lecture des données:
type=c("character","character","numeric","numeric",
   "numeric","character","character","character")
panel=read.table("jobpanel.dat",colClasses=type,
   header=TRUE)
# Codage explicite des facteurs
panel[,"THISYR"]=factor(panel[,"THISYR"],
  levels=c("0","1"),labels=c("Wnon","Woui"))
panel[,"CHILD1"]=factor(panel[,"CHILD1"],
  levels=c("0","1"),labels=c("Bnon","Boui"))
panel[,"CHILD2"]=factor(panel[,"CHILD2"],
  levels=c("0","1"),labels=c("Enon","Eoui"))
panel[,"ASCEND"]=factor(panel[,"BLACK"],
  levels=c("0","1"),labels=c("Abla","Anoi"))
panel=panel[,-c(2,6)]
summary(panel)
```

2.2 Exploration

Uni-dimensionnelle

Vérifier les distributions des variables quantitatives, justifier la transformation.

```
hist(panel[,"HUBINC"])
panel[,"LHUBINC"]=log(1+panel[,"HUBINC"])
hist(panel[,"AGE"])
hist(panel[,"EDUC"])
```

Bi-dimensionnelle

Observer les liasons entre les variables.

```
# 2 quantitatives
plot(panel[,"AGE"],panel[,"LHUBINC"])
mosaicplot(THISYR~ASCEND,data=panel)
# 1 quali et 1 quantitative
boxplot(AGE~THISYR,data=panel)
boxplot(LHUBINC~THISYR,data=panel)
```

Multi-dimensionnelle

Transformer les variables quantitatives en qualitatives avant de calculer l'analyse multiple des correspondances à l'aide de la librairie FactoMineR.

```
# Transformations
panel[,"AGEq"]=cut(panel$AGE,breaks=
   quantile(panel[,"AGE"],probs = seq(0, 1, 1/2)),
   labels=c("A0","A1"),include.lowest = TRUE)
panel[,"HUBINCq"]=cut(panel$HUBINC,breaks=
   quantile(panel[,"HUBINC"],probs=seq(0,1,1/3)),
   labels=c("R0","R1","R2"),include.lowest=TRUE)
panel[,"EDUCq"]=cut(panel$EDUC,breaks=c(0,11,13,20),
   labels=c("Scol0","Scol1","Scol2"),
         include.lowest=TRUE)
# Analyse des correspondances des
# variables qualitatives
library(FactoMineR)
afcm=MCA(panel[,c(1,5:7,9:11)],graph=F)
# graphe avec individus
plot(afcm, choix="ind",habillage="quali")
#sans les individus
plot(afcm,habillage="quali",invisible="ind")
```

Tenter une interprétation des axes.

2.3 Modélisation par régression logistique

La prévision de la variable THISYR n'est pas l'objectif, celle-ci est par ailleurs fort mauvaise ; l'objectif est plutôt la simple explicaiton de cette variable par les autres afin de mettre en évidence l'influence des différents fac-

teurs. On se contentera donc d'estimer le modèle complet sans interaction en utilisant soit les variables quantitatives, soit celles quantitatives transformées en qualitatives par découpage en classes.

```
# avec les variables quantitatives
panel.glm=glm(THISYR~LHUBINC+AGE+EDUC+CHILD1+
   CHILD2+ASCEND,family=binomial,data=panel)
summary(panel.glm)
# avec tout qualitatif
panel.glm=glm(THISYR~HUBINCq+AGEq+EDUCq+CHILD1+
   CHILD2+ASCEND,family=binomial,data=panel)
summary(panel.glm)
```

Interpréter dans les deux cas l'influence des facteurs dans l'explication de la variable `THISYR`.

2.4 Modélisation par arbre binaire

Un arbre binaire de décision est estimé afin de compléter le tour d'horizon des outils disponibles pouvant aider à la modélisation ou l'explication de la variable `THISYR`.

```
# Estimation de l'arbre avec toutes les
# variables quantitatives
panel.tree=rpart(THISYR~LHUBINC+AGE+EDUC+
   CHILD1+CHILD2+ASCEND,data=panel,
   parms=list(split='information'),cp=0.0001)
# élagage par validation croisée
# recherche de la pénalisation optimale
library(e1071)
res=tune.rpart(THISYR~LHUBINC+AGE+EDUC+
   CHILD1+CHILD2+ASCEND,data=panel,
   cp=seq(0,0.05,length=30 ))
plot(res)
print(res)
# Elagage
panel.tree=rpart(THISYR~LHUBINC+AGE+EDUC+
   CHILD1+CHILD2+ASCEND,data=panel,
   parms=list(split='information'),cp=0.0137931)
plot(panel.tree)
text(panel.tree)

# Même chose avec les variables qualitatives
panel.tree=rpart(THISYR~HUBINCq+AGEq+EDUCq+
   CHILD1+CHILD2+ASCEND,data=panel,
   parms=list(split='information'),cp=0.0001)
# élagage par validation croisée
# recherche de la pénalisation optimale
library(e1071)
res=tune.rpart(THISYR~HUBINCq+AGEq+EDUCq+
   CHILD1+CHILD2+ASCEND,data=panel,
   cp=seq(0,0.05,length=30 ))
plot(res)
print(res)
# Elagage
panel.tree=rpart(THISYR~HUBINCq+AGEq+EDUCq+
   CHILD1+CHILD2+ASCEND,data=panel,
   parms=list(split='information'),cp=0.01724138)
plot(panel.tree)
text(panel.tree)
```

Comparer la construction de ces arbres. Comparer avec le choix des variables dans les nœuds avec les résultats de la régression logistique.

3 Ceinture de sécurité avec SAS

3.1 Les données

On s'intéresse aux résultats (Jobson, 1991) d'une étude préalable à la législation sur le port de la ceinture de sécurité dans la province d'Alberta à Edmonton au Canada. Un échantillon de 86 769 rapports d'accidents de voitures ont été compulsés afin d'extraire une table de contingence complète croisant :

1. Gravité des blessures : Gr0 : rien à Gr3 : fatales
2. Risque regroupe Gr3 à Gr1 d'un côté et Gr0 de l'autre.
3. Port de la ceinture : Coui/Cnon

4. Sexe du conducteur : Hom/Fem
5. Etat du conducteur : Ajeu /A_bu

```
data sasuser.ceinture;
infile 'ceinture.dat';
input grave $ ceinture $ sexe $ alcool $ effectif;
select (grave);
when('Gr1','Gr2','Gr3') risque='Rimp';
when('Gr0') risque='Rfai';
otherwise;
end;
run;
```

3.2 Modélisations

Plusieurs modélisations sont testées avec les procédures `genmod` et `logistic`.

Compte tenu de la nature de la variable à expliquer qui est qualitative ordinale, la première chose à faire est d'utiliser la procédure suivante qui estime, par défaut, une régression logistique ordinale.

```
proc logistic data=sasuser.ceinture ;
class sexe alcool ceinture grave;
model grave=sexe alcool ceinture ;
freq effectif;
run;
```

L'hypothèse d'homogénéité des rapports de cote est-elle acceptable ?

Par la suite, compte tenu du résultat de ce test et des effectifs très déséquilibrés des modalités de la variable `risque`, les données ont été simplifiées pour ne considérer que deux états de gravité : aucune blessure ou blessure plus ou moins grave à fatale. Les deux procédures sont exécutées ci-dessous. Vérifier que, si les paramètres estimés sont différents, les tests de significativité conduisent aux mêmes conclusions.

```
proc logistic data=sasuser.ceinture;
class sexe alcool ceinture;
model risque=sexe|alcool|ceinture@2 ;
freq effectif;
run;

proc genmod data=sasuser.ceinture;
class sexe alcool ceinture ;
model risque=sexe|alcool|ceinture@2 /type3
dist=bin;
freq effectif;
run;
```

Que pensez vous de la présence des interactions ?

Utiliser la procédure logistic pour réduire le modèle.

Le modèle ci-dessous excluant les interactions permet d'estimer les rapports de cote ou odds ratio et même de tracer la courbe ROC.

```
proc logistic data=sasuser.ceinture descending;
class sexe alcool ceinture;
model risque=sexe alcool ceinture/ outroc=roc1;
freq effectif;
run;
```

Interpréter l'influence ds facteurs sur la gravité de l'accident. Qu'est-ce qui est le plus dangereux ?

Que dire de la qualité d'ajustement du modèle ?