TP: Choix de modèle en régression logistique avec SAS

Résumé

Les questions de choix de modèle sont cette fois traitées dans le cadre de la régression logistique dans l'environnement de SAS. L'objectif est donc de construire un modèle de prévision pour 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. Deux exemples sont traités pour illustrer les fonctionnalités des procédures de SAS.

1 Cancer de la prostate

Tous les fichiers de données ou de lecture sont accessibles sur la page d'URL :
`http://www.math.univ-toulouse.fr/~besse/Wikistat/data/`

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`).

1.2 Traitements interactifs

```
/* 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. 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.
2. Estimer un modèle binomial ou régression logistique dans SAS/insight (analyse de données interactives et ouvrir `sasuser.prost`) :
   `Analyse>fit`
   Sélectionner en `Y` la variable à expliquer (`lymph`). Elle est nécessairement de type réel (interval donc 0,1) pour cette procédure.
   Sélectionner toutes les variables explicatives (`l_acid, age, radio, taille, gravite`) puis `expand` pour considérer toutes les interactions d'ordre 2 dans `X`; Attention, supprimer `age*age`, `l_acid*l_acid` et `age*l_acid` inutiles à ce niveau.
   Cliquer sur `Method` et choisir `Binomial` pour la régression logistique au lieu de `Normal` qui est le choix par défaut correspondant au modèle gaussien ou régression multilinéaire. Laisser la fonction *lien canonique* qui, dans le cas binomial, est justement la fonction logit. Ajouter dans les résultats le tableau contenant les statistiques de test du rapport de vraisemblance : `Output > TypeIII(LR) Tests`.
   `OK,Apply` Le modèle est estimé et les tableaux de type III fournissent les statistiques des tests de Wald et du rapport de vraisemblance sur la significativité des paramètres du modèle.

3. 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 `delete` du menu `edit` de la même fenêtre.
4. A 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 Ceinture de sécurité

2.1 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;
```

2.2 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 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.

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