INSA – Génie Mathématique Année 2018-2019

Régression Non Linéaire avec Application sous R

GM5 – Feuille TP 7

Régression logistique sur variables quantitatives

L'objet de ce TP est d'expérimenter, avec la fonction `glm`, quelques techniques statistiques dans le cadre des modèles de régression logistiques binaires. Ce TP est très largement inspiré d'un TP récupéré sur le web (donc un grand merci à leurs auteurs) à l'URL
`http://www.kb.u-psud.fr/acces-etudiant/cours/biostat/html/Introduction%20aux%20biostatistiques/Introduction%20aux%20Biostatistiques.htm`
qui n'est plus active aujourd'hui ! ! !

Voici, en résumé et en anglais, la présentation de l'étude proposée par les auteurs :

The data consist of 200 subjects from a larger study on the survival of patients following admission to an adult intensive care unit (ICU). The study used logistic regression to predict the probability of survival for these patients until their discharge from the hospital. The dependent variable is the binary variable Vital Status (STA). Nineteen possible predictor variables, both discrete and continuous, were also observed.

1. Récupérer les données à l'URL
   `http://lmi2.insa-rouen.fr/~bportier/Data/dataTPUSI.csv`
   à l'aide de la commande

   ```
   nomfile = 'http://lmi2.insa-rouen.fr/~bportier/Data/dataTPUSI.csv'
   data <- read.csv2(nomfile, header = TRUE,sep=";")
   ```

   Ces données sont disponibles au format .csv, elles proviennent du site de statlib (`http://lib.stat.cmu.edu/DASL/`).

   Intitulé des variables :

`ID`	: numéro d'identification du patient
`STA`	: statu vital (0 = vivant, 1 = mort)
`AGE`	: ,ge, en années
`SEX`	: (1 = homme, 2 = femme)
`RACE`	: (1 = blanc, 2 = noir, 3 = autre)
`SER`	: service lors de l'admission en USI (0 = Medical, 1 = Chirurgical)
`CAN`	: un cancer fait-il parti du problème (0 = Non, 1 = Oui)
`IRC`	: antécédents d'insuffisance rénale chronique (0 = Non, 1 = Oui)
`INF`	: Infection probable à l'admission en USI (0 = Non, 1 = Oui)

MCE : Massage cardiaque avant l'admission en USI (0 = Non, 1 = Oui)
TAS : tension artérielle systolique à l'admission en USI (en mm Hg)
FC : fréquence cardiaque à l'admission en USI (battements/min)
ATC : antécédents d'admission en USI dans les derniers 6 mois (0 = Non, 1 = Oui)
TYP : type d'admission (0 = normal, 1 = en urgence)
FRA : fracture (0 = Non, 1 = Oui)
PO2 : PO2 de la gazométrie artérielle initiale (0 si > 60, 1 $\leq$ 60)
PH : PH de la gazométrie artérielle initiale (0 si $\geq$ 7,25, 1 < 7.25)
PCO : PCO2 de la gazométrie artérielle initiale (0 si $\leq$ 45, 1 > 45)
BIC : bicarbonate de la gazométrie artérielle initiale (0 si $\geq$ 18, 1 si $<$ 18)
CRE : créatinine du prélèvement initial (0 si $\leq 2,0$, 1 si $>$ 2,0)
CS : niveau de conscience à l'admission (0 = pas de coma ni stupeur,
1= stupeur profonde, 2 = coma)

2. Faire connaissance avec les données. On commencera par faire une étude descriptive univariée sur chacune des variables (faire des histogrammes pour les variables quantitatives). Puis, on étudiera les liens potentiels entre la variable à expliquer STA et chacune des variables explicatives. On pourra mettre en œuvre un test du chi-deux pour les variables explicatives qualitatives (fonction `chisq.test`), à l'aide des instructions suivantes :

```
# Construction de la table de contingence des variables CS et STA
tab = table(CS,STA)
# Test d'independance du Khi-deux
chisq.test(tab)
```

Quelles variables proposez-vous d'éliminer dès maintenant de l'étude ?

Les variables que vous aurez choisies de garder serviront à construire ce qu'on appellera le modèle complet du "statisticien". On pourra le noter $(M_{S,comp})$.

3. Consulter l'aide de la fonction `glm`.
4. On souhaite maintenant à l'aide d'une régression logistique déterminer le poids de chacune des variables explicatives. On lancera la commande

```
STAreglog <- glm(STA ~ AGE + etc ... , family = binomial, trace=TRUE)
```

Tester les commandes

```
> summary(STAreglog)
> anova(STAreglog)
```

Commenter les résultats obtenus.

5. A l'aide d'une méthode descendante, éliminer les variables explicatives jugées non significatives à 5%.

6. Comparer les modèles complet et réduit.
7. Etudier les résidus de Pearson et de déviance associés au modèle réduit.
8. En fait, la variable `CS`, qualitative à 3 modalités, est implicitement considérée dans le modèle comme une variable quantitative. On souhaite l'étudier comme une variable qualitative. Exécuter les instructions :

   ```
   CS <- as.factor(CS)
   summary(glm(STA ~ AGE + etc ... , family = binomial))
   ```

   Commenter les nouveaux résultats obtenus.
9. Le médecin préconise d'utiliser le modèle suivant :

   ```
   ModMedComp <- glm(STA ~ AGE + CAN + IRC + INF +
                     TAS + CS, family = binomial, trace=TRUE)
   ```

   On pourra noter ce modèle $(M_{M,comp})$. Etudier les performances de ce modèle, ainsi que celles de sa version réduite après élimination des variables jugées non significatives à 5% par une sélection de variables pas à pas descendante. On notera $(M_{M,reduit})$ le modèle réduit obtenu.
10. comparer les performances des différents modèles à l'aide de critères numériques et des tables de confusion.
11. Quelles conclusions peut-on tirer de cette étude ?
12. (Qustion bonus) Construire les courbes ROC associées aux différents modèles et conclure.