Modèles linéaires généralisés

Master 2 Pharmacologie 2016-2017

V. Monbet

1. Cancer de la prostate

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. Les variables considérées sont les suivantes :

- `age` âge du patient,
- `acid` niveau de "serum acid phosphatase",
- `radio` résultat "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` implication (1) ou non (0) du système lymphatique.
  On cherche un modèle permettant de bien prédire la variable `lymph`.
  Remarque : actuellement une simple échographie permet de délivrer un diagnostic avec beaucoup plus de sécurité.

packages : FactoMineR, pROC

Lecture des données

```
prostate = read.table(file="~/Dropbox/ENSEIGNEMENT/GLM_pharma/DATASETS/prostatedatr.t
summary(prostate)
```

On remarque que par défaut toutes les variables sont considérées comme des variables numériques. Il faudra donc changer le type de certaines d'entre elles.

Questions

1. Analyse descriptive et graphique
   (a) Analyse descriptive multivariée
   Ici, les variables sont quantitatives ou qualitatives binaires. On peut donc commencer par faire une analyse en composantes principales pour voir si des tendances se dégagent.

```
# install.packages(FactoMineR) # installer le package
library(FactoMineR)
acp = PCA(prostate,graph=FALSE)
par(mfrow=c(1,2))
plot(acp, choix = "var")
plot(acp, choix = "ind")
w = which(prostate$lymph==1)
points(acp$ind$coord[w,1:2],col="red",pch=20)
```

Commenter les graphiques obtenus.

On peut maintenant recoder les variables binaires

```
prostate$radio = as.factor(prostate$radio)
prostate$taille = as.factor(prostate$taille)
prostate$gravite = as.factor(prostate$gravite)
prostate$lymph = as.factor(prostate$lymph)
```

(b) Analyse descriptive de l'effet de l'âge

```
attach(prostate)
boxplot(age~lymph) # age seul
```

Une autre représentation graphique est possible qui permet de visualiser des probabilités.

```
age.m  <- matrix(0,1,ncl)
lymph.p <- matrix(0,1,ncl)
for (k in 1:ncl){
w = which(prostate$age>= q.age[k] & prostate$age< q.age[k+1])
age.m[k] = mean(prostate$age[w])
    lymph.p[k] = mean(as.numeric(prostate$lymp[w]))
}
plot(age.m,lymph.p,pch=20,xlab="age",ylab="probabilité de lymph")
lines(age.m,lymph.p)
```

Conclure.

(c) Faire le même type de graphiques pour la variable `acid`.

(d) Explorer l'effet du carré de l'âge ? de l'acide ?

(e) Analyse graphique des interactions
Le nombre de variables est peu important (attention le nombre d'observations aussi), on peut donc explorer les interactions graphiquement. En suivant l'exemple ci-dessous, illustrer les interactions entre les variables qualitatives et les variables quantitative. Interpréter les graphiques et conclure.

```
age.m  <- matrix(0,1,ncl)
lymph.p0 <- lymph.p1 <- matrix(0,1,ncl)
for (k in 1:ncl){
w = which(prostate$age>= q.age[k] & prostate$age< q.age[k+1])
age.m[k] = mean(prostate$age[w])
    lymph.p0[k] = sum((prostate$lymp[w]==1)&(prostate$radio[w]==0))/length(w)
    lymph.p1[k] = sum((prostate$lymp[w]==1)&(prostate$radio[w]==1))/length(w)
}
plot(age.m,lymph.p0,pch="0",ylim=range(lymph.p0,lymph.p1))
lines(age.m,lymph.p0)
points(age.m,lymph.p1,pch="1",col="red")
lines(age.m,lymph.p1,col="red")
```

Pour tracer les logit, il suffit d'écrire `logit(lymph.p0)` à la place de `lymph.p0` dans le plot. Conclure.

2. Modélisation, régression logistique (sans discrétisation).

   (a) On peut commencer par ajuster le modèle nul et le modèle sans interactions pour compléter l'analyse graphique et se faire une idée des variables les plus importantes.

```
models.0 <- list(
   null     = glm(lymph ~ 1, family=binomial, data=prostate),
   additif     = glm(lymph ~ ., family=binomial, data=prostate)
 )
summary(models.0$null)
summary(models.0$additif)
```

   Analyser tout d'abord les déviances. Le modèle additif apporte t'il de l'information ? Quel test permet de répondre à cette question ? Quelles sont les hypothèses nécessaires à la validité de ce test ?

   Quelle(s) variable(s) apporte(nt) le plus d'information ? Quelle(s) variable(s) apporte(nt) le moins d'information ?

   (b) Sélection de variable en arrière.
   Dans une première analyse rapide, on peut utiliser la fonction `step` de R qui permet de faire de la sélection *backward* de variables en se basant (par défaut) sur le critère AIC.

```
step(models.0$additif)
```

   Quel est le modèle retenu par cette méthode ? Comparer sa déviance et son critère AIC au modèle additif complet. Conclure.

   (c) Sélection de variables en avant.
   Une première manière de faire la sélection de variable en avant consiste encore à utiliser la fonction `step`.

```
step(models.0$additif,direction="forward")
```

Mais, il est parfois avantageux de le faire à la main. Par exemple, les commandes

```
models <- list(
   null     = glm(lymph ~ 1, family=binomial, data=prostate),
   radio     = glm(lymph ~ radio, family=binomial, data=prostate),
   taille     = glm(lymph ~ taille, family=binomial, data=prostate),
   add    = glm(lymph ~ radio+taille, family=binomial, data=prostate),
   mult     = glm(lymph ~ radio:taille, family=binomial, data=prostate)
 )
 res = matrix(0,length(models),3)
 colnames(res) <- c("Dev.","ddl","AIC" )
 rownames(res) <- c("null","radio","taille","add","mult")
 for (k in 1:length(models)){
   res[k,1] = deviance(models[[k]])
   res[k,2] = models[[k]]$df.residual
   res[k,3] = aic(models[[k]])
}
res
```

permettent de comparer les déviances et de réaliser les tests de rapport de vraisemblance pour les modèles emboités. Analyser les résultats puis ajouter la variable `acid`.

On rappelle qu'on peut obtenir les p-values des tests basés sur la déviance

```
pv = 1-pchisq(deviance(models$radio)-deviance(models$add),
    df=models$radio$df.residual-models$add$df.residual)
```

(d) Poursuivre cette analyse jusqu'à obtenir le "meilleur" modèle au sens de la déviance (et/ou) de l'AIC. Pour prendre en compte les interactions, on pourra par exemple écrire

```
mult.all     = glm(lymph ~ radio+taille+acid:radio+acid:taille,
     family=binomial, data=prostate)
```

(e) Qualité d'ajustement.
Pour les données individuelles, la déviance ne suit pas une loi du chi2. Pour tester la qualité du modèle final, on peut utiliser un test de Hosmer-Lemeshor.

3. Modélisation, régression logistique (avec discrétisation).
On peut reprendre un analyse similaire en discrétisant les deux variables continues.
4. Interprétation des effets.
Une fois qu'on a choisit le modèle, on peut interpréter les différents effets.
5. Diagnostiques.
Faire les diagnostiques de régression pour identifier d'éventuels individus qui auraient un effet trop fort en adaptant les codes introduits dans le cours.
6. Validation croisée.
Ici l'objectif est prédictif, il est donc utile de vérifier les performances du modèle par validation croisée.

(a) Estimer l'erreur de généralisation, la spécificité et la sensibilité par validation K-fold. Choisir par exemple K=5. Tracer la courbe ROC et calculer l'aire sous la courbe.

```
n = nrow(prostate)
ii = sample(n,n)
Kfold = 5
nk = floor(n/Kfold)
p.fit = NULL
Y.test = NULL
for (k in 1:Kfold){
testi = ii[((k-1)*nk+1):(k*nk)]
appri = setdiff(ii,appri)
mod   = glm(lymph ~ acid+radio+taille, family=binomial, data=prostate,
    subset=appri)
pr = predict(mod,prostate[testi,],type="response")
Y.test = c(Y.test,prostate$lymp[testi])
p.fit = c(p.fit,pr)
}
roc.res = roc(Y.test,p.fit,plot=TRUE,main=floor(roc.res$auc*100)/100)
str(roc.res)
table(Y.test,p.fit>.5)
#roc(Y.test,p.fit,plot=TRUE,smooth=TRUE,col="gray",add=TRUE)
```

Quel niveau de probabilité permet d'atteindre une sensibilité de 0.8 ? Une spécificité de 0.8 ?

```
plot(roc.res$thresholds,roc.res$sensitivities,pch=20)
grid()
points(roc.res$thresholds,roc.res$specificities,pch=20,col="red")
```

(b) Estimer l'erreur de généralisation, la spécificité et la sensibilité par validation bootstrap. Tracer la courbe ROC et calculer l'aire sous la courbe.

```
B = 100
nk = floor(n/10)
p.fit <- p.fit.all <- NULL
Y.test = NULL
for (b in 1:B){
testi = sample(1:n,nk)
appri = setdiff(ii,appri)
mod   = glm(lymph ~ acid+radio+taille, family=binomial, data=prostate,
    subset=appri)
pr = predict(mod,prostate[testi,],type="response")
Y.test = c(Y.test,prostate$lymp[testi])
p.fit = c(p.fit,pr)
}
roc.res = roc(Y.test,p.fit,plot=TRUE,main=floor(roc.res$auc*100)/100)
```

(c) Comparer avec les résultats obtenus par le modèle additif complet (celui qui entre toutes les variables explicatives). Discuter les résultats obtenus.

2. Santé Mentale

packages : MASS, VGAM

Agresti et Finlay (1997) présentent les données de rapport d'une étude de Floride examinant la relation entre santé mentale et plusieurs variables explicatives utilisant un échantillon aléatoire de 40 sujets. La variable d'intérêt est un indice de santé mentale qui incorpore des mesures d'anxiété et de dépression. Nous considérons deux variables explicatives : un score de qualité de vie qui combine le nombre et la gravité de divers événements stressants et un indice de statut socio-économique (SES).
Le fichier de données est dispobible sous le lien :

```
MH = read.table("",header=TRUE)
head(MH)
```

Questions

1. Modéle linéaire
   (a) Tracer un scatterplot des 3 variables en utilisant la function `pairs` et commentez les figures.
   (b) Ajuster un modèle de régression multiple de l'indice de santé mentale sur les variables prédictives. Interpréter les paramètres et realiser les tests de significativité adéquats.
   (c) Quel proportion de variation entre les What proportion of the variation across subjects in the index of mental health is explained by the life events index ? How is this proportion related to Pearson's correlation coefficient ?
2. Modèle linéaire généralisé - variable `ses`
   (a) Visualisation des données

```
MH$mental = factor(MH$mental,ordered=as.ordered(1:4))
                                              # ordered respons

lapply(MH[, c("mental", "ses", "life")], table)
ftable(xtabs(~ ses + life + mental, data = MH))
```

   (b) Dans un premier temps, on s'intéresse uniquement à la variable prédictive SES. Mettre les données en forme

```
Fses = ftable(xtabs(~ ses+mental , data = MH))
MHses = matrix(c(0,1,t(as.matrix(Fses))),2,5)
colnames(MHses) = c("ses","m1","m2","m3","m4")
MHses = as.data.frame(MHses)
MHses
```

Faire un graphique pour vérifier si l'hypothèse de rapport de côtes proportionnels est raisonnable.

(c) Ajuster un modèle de régression logistique multinoniale à rapport de côte proportionnels.

```
fit = vglm(cbind(m1,m2,m3,m4)~ses,data=MHses,
    family = cumulative(parallel=TRUE))
summary(fit)
```

(d) Quelle est la différence entre les deux appels de vglm suivants

```
fit = vglm(cbind(m1,m2,m3,m4)~ses,data=MHses,
    family = cumulative(parallel=TRUE))
fit1 = vglm(cbind(m1,m2,m3,m4)~ses,data=MHses,family = propodds)
fit2 = polr(mental~ses,data=MH)
```

Ecrire les modèles conrrespondants.

(e) On peut donner quelques précisions sur le modèle ajusté par la fonction `polr`. Tests sur les coefficients

```
 p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))
(ctable <- cbind(ctable, "p value" = p))
```

Test sur le modèle

```
Anova(pol)
```

(f) Que penser de l'idée d'ajuster un modèle à rapports de côte non proportionnels ?

3. Modèle linéaire généralisé - variables `ses` et `life`

(a) Ajuster un modèle pour prédire `mental` en fonction de `ses` et `life`.

3. Diabète

Le nombre de décès par diabète en New South Wales (Australie) en 2002 est fourni par par l'institut Australien de la Santé (de Jong & Heller, 200).

Gender	Age	Deaths	Population	Rate per 100 000
Male	<25	3	1141 100	0.26
	25–34	0	485 571	0.00
	35–44	12	504 312	2.38
	45–54	25	447 315	5.59
	55–64	61	330 902	18.43
	65–74	130	226 403	57.42
	75–84	192	130 527	147.10
	85+	102	29 785	342.45
Female	<25	2	1086 408	0.18
	25–34	1	489 948	0.20
	35–44	3	504 030	0.60
	45–54	11	445 763	2.47
	55–64	30	323 669	9.27
	65–74	63	241 488	26.09
	75–84	174	179 686	96.84
	85+	159	67 203	236.60

1. Analyse descriptive et modèle général.
   (a) Tracer le nombre de mort par diabètes par genre en fonction de l'âge, puis le taux de décès pour 100000 habitants et le log du taux.

   ```
   plot(age.m,diabete$Deaths[1:8],pch=20,xlab="Age",ylab="Number of deaths",ylim
   lines(age.m,diabete$Deaths[1:8])
   points(age.m,diabete$Deaths[9:16],pch=17,col="red")
   lines(age.m,diabete$Deaths[9:16],col="red")
   legend(20,180,legend=c("Male","Female"),pch=c(20,17),col=c("black","red"))
   ```

   Que peut-on en déduire sur la relation entre l'âge et le taux de décès ?
   (b) Quel modèle est adapté pour ces données ? Ecrire le modèle.
2. Modèle additif.
   (a) Age comme une variable quantitative.
   Dans un premier temps, on considère un modèle sans interaction et l'âge comme une variable continue. Les graphiques suggèrent alors de choisir une forme polynomiale pour l'effet de l'âge. Ajuster les modèles correspondant et réaliser un analyse de déviance pour tester le degré du polynôme. Puis interpréter le modèle retenu.
   (b) Age comme une variable qualitative.
   ajuster maintenant le modèle avec l'âge introduit comme une variable catégorielle. Faire l'anayse des déviances. Puis interpréter le modèle obtenu.
3. Modèle avec interactions.
   Ajuster le(s) modèle(s) avec interaction.
4. Quel est le meilleur modèle ? Dans quel sens est-il le meilleur ? Donner une interprétation de ce modèle. Tracer les courbes (prédites) du nombre de décès par genre en fonction de l'âge.
5. Estimer l'erreur de généralisation.