Université de Montpellier Année 2020-2021
HMMA304 - Modèles linéaires généralisés

Examen de contrôle continu
Durée 1h30 - Sans document - 14 octobre 2020

Exercice 1 (8 pts)

On considère n variables aléatoires indépendantes $y_1, \dots, y_n$ telles que

$$f(y_i; \pi_i) = \pi_i (1 - \pi_i)^{y_i - 1} \mathbf{1}_{\mathbb{N}^*}(y_i) \, .$$

Pour tout $i = 1, \dots, n$, on suppose que $\log(1 - \pi_i) = \alpha + \beta x_i$ où $x_i \in \mathbb{R}$ est supposé connu, α et β sont des paramètres inconnus.

1 (2 pts) Montrer qu'il s'agit d'un modèle linéaire généralisée. La fonction de lien canonique a-t-elle été utilisée ?

2 (3 pts) Donner la log-vraisemblance et les équations de vraisemblance. Pouvons-nous calculer les expressions analytique des estimateurs du maximum de vraisemblance de α et β ?

3 (3 pts) Calculer la matrice d'information de Fisher apportée par $(y_1, \dots, y_n)$ sur les paramètres α et β notée $I_n(\alpha, \beta)$.

Exercice 2 (4 pts)

Nous nous intéressons ici à l'étude de la pièce ayant causé l'explosion de la navette spatiale Challenger en 1986. L'étanchéité du moteur de la navette spatiale est assurée par six pièces identiques appelées "O-ring". L'explosion de la navette Challenger est due à la défaillance d'au moins l'une de ces pièces.

Au cours des 24 vols précédents d'une navette spatiale, nous disposons des données suivantes : la variable `temp` qui correspond à la température au moment du lancement et la variable `defa` qui vaut 0 si aucun des "O-ring" n'a été endommagé au cours du lancement et 1 si au moins l'un d'entre eux a été endommagé.

Proposer une méthode permettant d'estimer la probabilité de défaillance d'au moins un "O-ring" pour une température de 31 degrés Fahrenheit (température au moment du lancement de la navette challenger). Nous supposons que les données sont stockées dans un data.frame R nommé `challenger`, donner le code R associé.

Exercice 3 (2 pts)

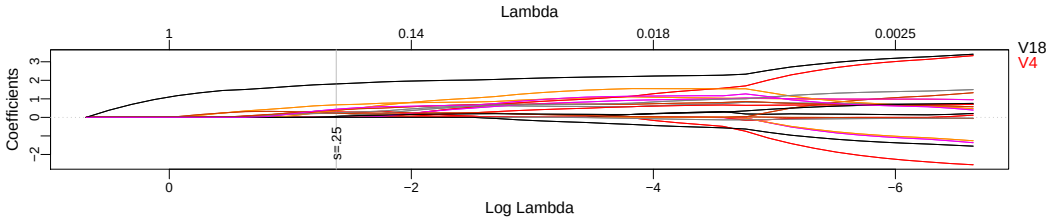
Expliquer en quoi consiste une régression ridge et donner l'expression explicite de l'estimateur ridge.

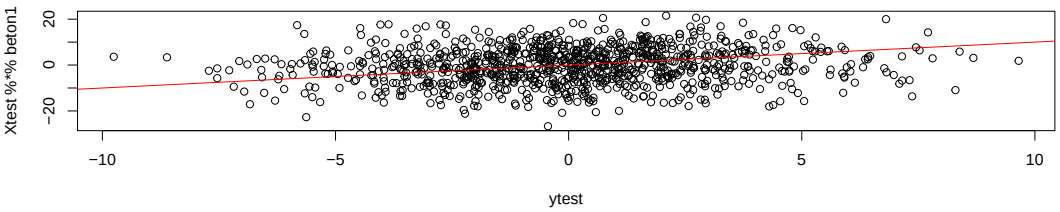
Exercice 4 (6 pts)

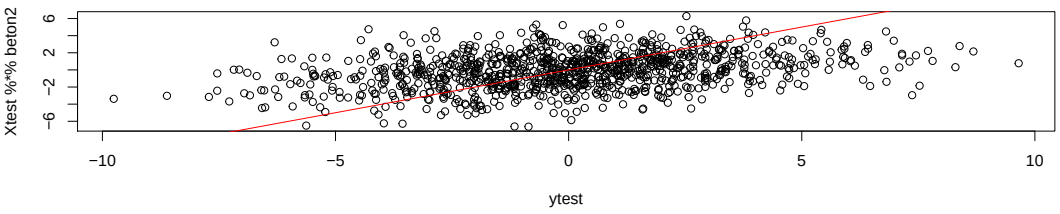
Expliquer en détails les résultats produits par le code R ci-dessous

```
library(glmnet)
library(caret)

X <- matrix(rnorm(20*20),20)
beta <- rep(0.5,20)
y <- X%*%beta+rnorm(20,sd=2)

app <- data.frame(y=y,X)
names(app)[-1] <- paste("x",1:20,sep="")

model <- lm(y~-1+.,data=app)
beton1 <- as.numeric(model$coefficients)

model <- glmnet(X,y,family="gaussian",nlambda=50,alpha=1,intercept=FALSE)
model <- train(y~.,data=app,method="glmnet",intercept=FALSE,metric="RMSE",
                trControl=trainControl(method="repeatedcv",number=20),
                tuneGrid=data.frame(alpha=1,lambda=model$lambda))

opti <- as.numeric(model$bestTune$lambda)
model$results[model$results[,2]==opti,1:3]
#    alpha    lambda     RMSE
# 29     1 0.2516718 1.873211
beton2 <- coef(model$finalModel,opti)[-1]

sum(abs(beton1))
# [1] 26.90182
sum(beton2==0)
# [1] 9
sum(abs(beton2))
# [1] 4.785677

library(plotmo)
plot_glmnet(glmnet(X,y,family="gaussian",nlambda=50,alpha=1,intercept=FALSE),
    s=as.numeric(model$bestTune$lambda),label=2)

Xtest <- matrix(rnorm(1000*20),1000)
test <- data.frame(Xtest)
```

```
names(test) <- paste("x",1:20,sep="")
ytest <- Xtest%*%beta+rnorm(1000,sd=2)

plot(ytest, Xtest%*%beton1)
abline(a=0,b=1,col="red")

plot(ytest, Xtest%*%beton2)
abline(a=0,b=1,col="red")

sqrt(mean((Xtest%*%beta-ytest)^2))
# [1] 1.975905
sqrt(mean((Xtest%*%beton1-ytest)^2))
# [1] 8.144402
sqrt(mean((Xtest%*%beton2-ytest)^2))
# [1] 2.988543
```







Correction examen de contrôle continu HMMA304 14/10/2020

Exercice 1

1) $f(y;\pi) = \pi(1-\pi)^{y-1}$ $\mathbb{1}_{\mathbb{N}^*}(y)$

$f(y;\pi) = \exp\left[y \log(1-\pi) + \log\left(\frac{\pi}{1-\pi}\right)\right]$ $\mathbb{1}_{\mathbb{N}^*}(y)$

$\theta = \log(1-\pi)$

$\Rightarrow \pi = 1 - e^{\theta}$

$b(\theta) = \log\left[\frac{1-e^{\theta}}{e^{\theta}}\right] = -\log(1-e^{\theta}) + \theta$

Famille exponentielle (2)

Oui - c'est la liaison canonique. $E(y) = \frac{1}{\Pi} = \frac{1}{1-e^{\alpha+\beta x_i}}$

2] $LV(\alpha,\beta) = \sum_{i=1}^{M} (y_i - 1)[\alpha + \beta x_i]$

$+ \sum_{i=1}^{M} \ln(1 - e^{\alpha+\beta x_i})$

$$\frac{dLV}{d\alpha}(\alpha,\beta) = \sum_{i=1}^{M} y_i - M - \sum_{i=1}^{M} \left[\frac{e^{\alpha+\beta x_i}}{1-e^{\alpha+\beta x_i}}\right]$$

$$\frac{dLV}{d\beta}(\alpha,\beta) = \sum_{i=1}^{M} x_i y_i - \sum_{i=1}^{M} x_i - \sum_{i=1}^{M} \left[\frac{x_i e^{\alpha+\beta x_i}}{1-e^{\alpha+\beta x_i}}\right]$$

Pas d'expression analytique de l'EMV.

3]

$$\frac{\partial^2 LV}{(\partial\alpha)^2}(\alpha,\beta) = -\sum_{i=1}^{M} \frac{e^{\alpha+\beta x_i}\left(1-e^{\alpha+\beta x_i}\right)+\left(e^{\alpha+\beta x_i}\right)^2}{\left(1-e^{\alpha+\beta x_i}\right)^2}$$

(3)

$$\frac{\partial^2 LV}{(\partial\alpha)^2}(\alpha,\beta) = -\sum_{i=1}^{M} \frac{e^{\alpha+\beta x_i}}{\left(1-e^{\alpha+\beta x_i}\right)^2}$$

$$\frac{\partial^2 LV}{\partial\alpha\,\partial\beta}(\alpha,\beta) = -\sum_{i=1}^{M} \frac{x_i e^{\alpha+\beta x_i}}{\left(1-e^{\alpha+\beta x_i}\right)^2}$$

$$\frac{\partial^2 LV}{(\partial\beta)^2}(\alpha,\beta) = -\sum_{i=1}^{M} \frac{x_i^2 e^{\alpha+\beta x_i}}{\left(1-e^{\alpha+\beta x_i}\right)^2}$$

Tous ces termes sont déterministes on en déduit aisément l'information de Fisher

Exercice 2

Modèle de régression logistique

→ Variable à expliquer
dégâts qui est binaire

→ Variable explicative
temp

$$\mathbb{P}(\text{dégâts} = 1 \mid \text{temp}) = \left[\frac{e^{\alpha + \beta \text{temp}}}{1 + e^{\alpha + \beta \text{temp}}}\right]$$

```
model <- glm(dégâts ~ 1 + temp,
family = "binomial")
predict(model, data.frame(temp=31))
```

Exercice 3

Modèle de régression régularisé
Moindres carrés sous contrainte

$$\hat{\beta}^R \in \arg\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

sous contraintes $\sum_{j=1}^{p} \beta_j^2 \leq \tau$

Problème dual

$$\min_{\beta} \sum_{i=1}^{n} (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ fonction de τ

$$\hat{\beta}^R = (X^T X + \lambda I_p)^{-1} X^T y$$

où $X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ matrice des variables explicatives

Exercice 4

On simule 20 données suivant le modèle :

$$y = \sum_{j=1}^{20} \beta_j x_j + u$$

où $u \sim N(0,4)$

On estime le vecteur β par la méthode des moindres carrés
$\Rightarrow$ objet R boston1

On estime le vecteur β par une technique LASSO dont la pénalisation est calibré par validation croisée à 20 ensembles (leave-one-out cross validation)
$\Rightarrow$ objet R boston2

On trouve que 9 coefficients sont estimés à 0 et que les sommes des valeurs absolues du boston 1

est largement supérieure à celle des valeurs obtenues dans le tableau 2. C'est l'effet de la régularisation LASSO.

Graphe 1 : évolution des valeurs des estimations des paramètres LASSO en fonction de la contrainte

Graphe 2 : estimation MCO non biaisé mais très variable

Graphe 3 : estimation LASSO biaisée mais peu variable

Sur les données de test, très bons résultats pour le LASSO, très proche de l'Oracle.