Université de Montpellier — Année 2022-2023
HAX912X - Modèles linéaires généralisés

Examen de contrôle continu

sans document - 13 octobre 2022

Exercice 1 (5 pts - QCM - Bonne réponse +1 pt - Mauvaise réponse -0.5 pt)

1 (1 pt) Nous avons effectué une régression linéaire multiple, une des variables explicatives est la constante, la somme des résidus calculés vaut :

A) 0
B) approximativement 0
C) parfois 0

2 (1 pt) Le vecteur $\hat{y}$ est-il orthogonal au vecteur des résidus estimés ?

A) oui
B) non
C) seulement si la constante fait partie des variables explicatives

3 (1 pt) Un estimateur de la variance de $\hat{\beta}$, estimateur des moindres carrés de β, est :

A) $\sigma^2(X^TX)^{-1}$
B) $\hat{\sigma}^2(X^TX)^{-1}$
C) $\hat{\sigma}^2(XX^T)^{-1}$

4 (1 pt) Une régression a été effectuée et le calcul de la SCR a donné la valeur notée SCR1. Une variable est ajoutée, le calcul de la SCR a donné une nouvelle valeur notée SCR2. Nous avons :

A) SCR1 $\leq$ SCR2
B) SCR1 $\geq$ SCR2
C) cela dépend de la variable ajoutée

5 (1 pt) Une régression a été effectuée et un estimateur de la variance résiduelle a donné la valeur notée $\hat{\sigma}_1^2$. Une variable est rajoutée et un estimateur de la variance résiduelle vaut maintenant $\hat{\sigma}_2^2$. Nous avons :

A) $\hat{\sigma}_1^2 \leq \hat{\sigma}_2^2$
B) $\hat{\sigma}_1^2 \geq \hat{\sigma}_2^2$
C) on ne peut rien dire

Exercice 2 (4 pts)

Nous ne souhaitons pas accorder la même importance à toutes les observations dans un modèle de régression linéaire multiple

$$y_i = x_i^T \beta + u_i$$

avec u_i centré. Soit $(p_1, \ldots, p_n)$ les poids respectifs que nous accordons aux différentes observations, avec $\forall i \in \{1, \ldots, n\}, \quad p_i > 0$ et $\sum_{i=1}^{n} p_i = n$.
Proposer un estimateur pour β.

Exercice 3 (2 pts)

Dans quel but utilise-t-on une méthode de validation croisée ? Comment choisir le nombre de groupes ? Donner deux exemples de mise en oeuvre de nature déférente.

Exercice 4 (4 pts)

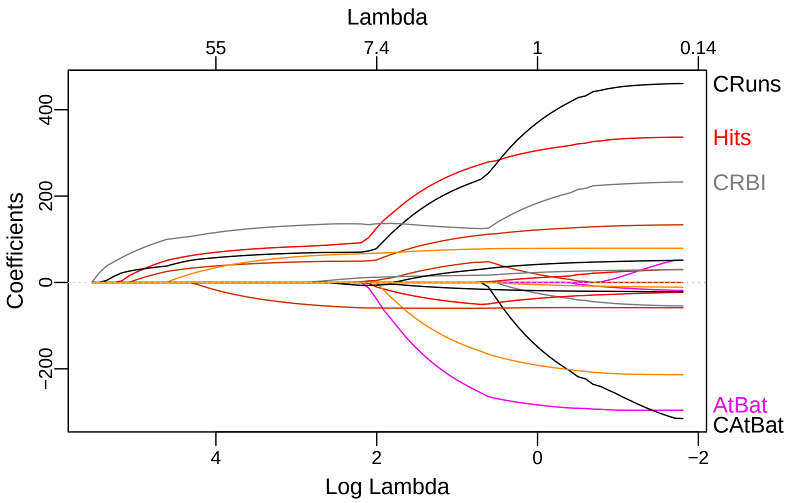
Nous considérons n variables aléatoires indépendantes $y_1, \ldots, y_n$ telles que y_i est distribuée suivant une loi binomiale de paramètre $(100, p_i)$. Pour tout $i = 1, \ldots, n$, nous supposons que $\log(-\log(1 - p_i)) = x_i^T \beta$ où $x_i \in \mathbb{R}^p$ est un vecteur supposé connu et β un vecteur de paramètres inconnus. Montrer qu'il s'agit d'un modèle linéaire généralisé. La fonction de lien canonique a-t-elle été utilisée ?

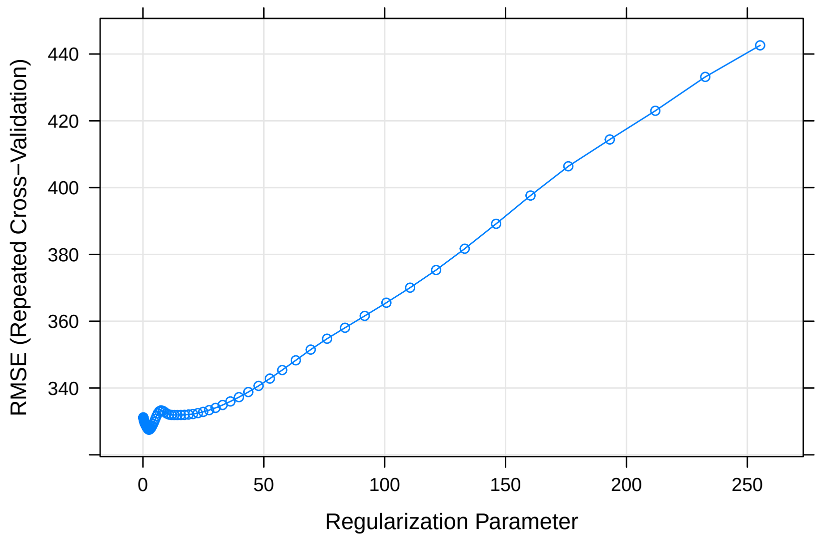
Exercice 5 (5 pts)

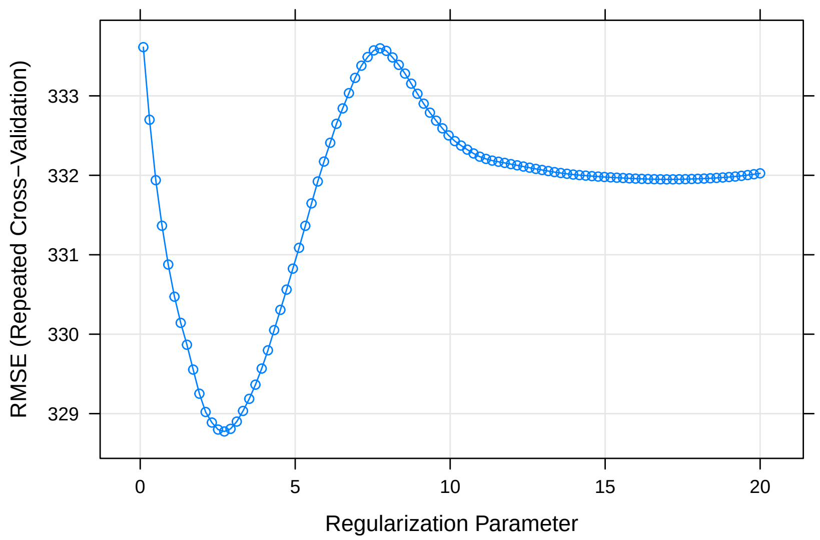
Expliquer en détails les procédures mises en oeuvre par l'intermédiaire du code R ci-dessous et expliciter clairement ce que montrent les résultats. Une description du jeu de donnés est fournie en annexe.

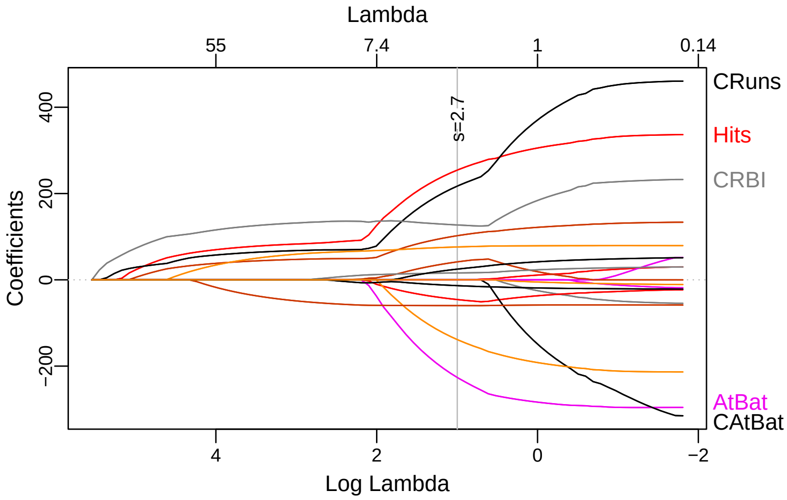
```
library(ISLR)
Hitters <- na.omit(Hitters)
dim(Hitters)

# [1] 263  20

names(Hitters)

#  [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"
#  [6] "Walks"     "Years"     "CAtBat"    "CHits"     "CHmRun"
# [11] "CRuns"     "CRBI"      "CWalks"    "League"    "Division"
# [16] "PutOuts"   "Assists"   "Errors"    "Salary"    "NewLeague"

Hitters.X <- model.matrix(Salary~.,data=Hitters)
Hitters.X <- scale(Hitters.X)

library(glmnet)
library(plotmo)

model1 <- glmnet(Hitters.X,Hitters$Salary,alpha=1)
plot_glmnet(model1,label=5)

library(caret)
model2 <- train(Hitters.X,Hitters$Salary,method="glmnet",metric="RMSE",
        trControl=trainControl(method="repeatedcv",number=10,repeats=10),
        tuneGrid=data.frame(alpha=1,lambda=model1$lambda))
plot(model2)

model3 <- train(Hitters.X,Hitters$Salary,method="glmnet",metric="RMSE",
                trControl=trainControl(method="repeatedcv",number=10,repeats=10),
                tuneGrid=data.frame(alpha=1,lambda=seq(0.1,20,length=100)))
plot(model3)

lambda_opt_lasso <- as.numeric(model3$bestTune[2])
model3$results[model2$results[,2]==lambda_opt_lasso,]
#  alpha   lambda     RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
#      1 2.713131 328.7756 0.4885506 232.9141 81.10729  0.1731684 42.08644
```

```
glmnet(Hitters.X,Hitters$Salary,lambda=lambda_opt_lasso,alpha=1)$beta

# (Intercept)        .
# AtBat       -229.39762
# Hits         256.58174
# HmRun              .
# Runs               .
# RBI                .
# Walks        102.98202
# Years        -44.62761
# CAtBat             .
# CHits              .
# CHmRun        45.53399
# CRuns        222.54955
# CRBI         120.14755
# CWalks      -141.46130
# LeagueN       16.17401
# DivisionW    -59.59561
# PutOuts       76.43519
# Assists       25.15668
# Errors       -13.29525
# NewLeagueN         .

plot_glmnet(model1,label=5,s=lambda_opt_lasso)
```




RMSE (Repeated Cross−Validation)
440
420
400
380
360
340
0
50
100
150
200
250
Regularization Parameter


RMSE (Repeated Cross−Validation)
333
332
331
330
329
0
5
10
15
20
Regularization Parameter



Annexe

```
Hitters (ISLR) R Documentation Baseball Data

Description
Major League Baseball Data from the 1986 and 1987 seasons
A data frame with 322 observations of major league players
on the following 20 variables

AtBat
Number of times at bat in 1986

Hits
Number of hits in 1986

HmRun
Number of home runs in 1986

Runs
Number of runs in 1986

RBI
Number of runs batted in in 1986

Walks
Number of walks in 1986

Years
Number of years in the major leagues

CAtBat
```

Number of times at bat during his career

CHits
Number of hits during his career

CHmRun
Number of home runs during his career

CRuns
Number of runs during his career

CRBI
Number of runs batted in during his career

CWalks
Number of walks during his career

League
A factor with levels A and N indicating player's league at the end of 1986

Division
A factor with levels E and W indicating player's division at the end of 1986

PutOuts
Number of put outs in 1986

Assists
Number of assists in 1986

Errors
Number of errors in 1986

Salary
1987 annual salary on opening day in thousands of dollars

NewLeague
A factor with levels A and N indicating player's league at the beginning of 1987

Source
This dataset was taken from the StatLib library which is maintained at Carnegie Mellon University. This is part of the data that was used in the 1988 ASA Graphics Section Poster Session. The salary data were originally from Sports Illustrated, April 20, 1987.

The 1986 and career statistics were obtained from The 1987 Baseball Encyclopedia Update published by Collier Books, Macmillan Publishing Company, New York.

References

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, https://www.statlearning.com, Springer-Verlag, New York

1

Correction examen
partiel HAX 912X
13/10/2022

Exercice 1

1 - A)

2 - A)

3 - B)

4 - B)

5 - C)

Exercice 2

Toutes les observations n'ont pas le même poids dans un modèle de régression linéaire multiple.

L'observation (y_i, x_i) est ②
associée au poids p_i.
Clairement plus ce dernier
est important plus on
souhaite donner de l'importance
à l'observation associée.
On peut alors utiliser l'estimateur
des moindres carrés généralisés
avec comme matrice de
variance-covariance
la matrice diagonale
associée au vecteur

$$\left(\frac{1}{p_1}, ----, \frac{1}{p_n}\right).$$

$$\Omega = \begin{bmatrix} 1/\rho_1 & & 0 \\ & \ddots & \\ 0 & & 1/\rho_m \end{bmatrix}$$

$$\hat{\beta} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y$$

Exercice 3

On utilise la validation croisée pour estimer l'erreur associée à une stratégie (un modèle) de prédiction.

Le nombre de groupes dépend du nombre d'observations dont on dispose.

4

Si nous avons beaucoup d'observations en égard à la complexité du problème considéré alors on prend 2 groupes.

S'il est faible, on prend autant de groupes que d'observations.

S'il est entre les deux cas évoqués ci-dessus, on prend 5 ou 10 groupes.

On peut utiliser la validation croisée pour fixer le paramètre de régularisation d'une regression lasso.

On peut également utiliser la validation croisée pour évaluer l'erreur associée à un modèle de régression logistique.

Exercice 4

La loi binomiale appartient à la famille exponentielle 1-valeur. En effet

$$f(y|p) = C_{100}^{y}\, p^{y}(1-p)^{100-y}\, \mathbb{1}_{\{0,\dots,100\}}(y)$$

$$f(y|p) = \exp\left\{ y \log\left(\frac{p}{1-p}\right) + 100 \log(1-p) \right\} r(dy)$$

$$\theta = \log\left(\frac{p}{1-p}\right) \Rightarrow p = \frac{e^{\theta}}{1+e^{\theta}}$$

6

$$b(\theta) = -100 \log\left(1 - \frac{e^\theta}{1+e^\theta}\right)$$

$$b(\theta) = 100 \log(1+e^\theta)$$

$$b'(\theta) = 100 \frac{e^\theta}{1+e^\theta} \quad (= 100p)$$

Par ailleurs,

$$\log(-\log(1-p)) = x^T\beta$$

$$-\log(1-p) = e^{x^T\beta}$$

$$1-p = e^{-e^{x^T\beta}}$$

$$p = 1 - e^{-e^{x^T\beta}}$$

$$\Rightarrow \mathbb{E}(y) = g(x^T\beta)$$

Il s'agit donc bien
d'un modèle linéaire généralisé.

Par ailleurs, comme $h'(\cdot) \neq g(\cdot)$
ce n'est pas le lien canonique
qui a été utilisé.

Exercice 5

Mise en œuvre d'un modèle
de régression avec régularisation
de type lasso sur des données
collectées entre 1986 et 1987
où l'on essaie d'expliquer le
salaire de joueurs de Baseball
à partir de 19 covariables
décrivant les performances des joueurs.

Le paramètre de régularisation est déterminé à partir d'une stratégie de validation croisée à 10 ensembles répétés 10 fois. Pour affiner la précision de la valeur de régularisation à choisir, on raffine la grille initiale. En effet, le graphe RMSE pour model2 montre qu'il convient de se focaliser sur les valeurs de λ comprises entre 0 et 20.

9

Le modèle lasso finalement choisi contient 23 variables actives.