CC1 d'Elements de modélisation statistique

25/11/2022

Durée : 2h
Les documents, les calculatrices et les téléphones portables ne sont pas autorisés. Vous prendrez soin à la rédaction de vos réponses et à la définition de toutes vos notations.

Exercice 1

On s'intéresse ici à un extrait du jeu de données Hitters disponible dans la librairie `ISLR`. Ces données sont composées de statistiques et des salaires de joueurs de baseball, soit 9 variables, pour décrire $n = 257$ joueurs des ligues majeures de baseball américaines :

- 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
- Salary : 1987 annual salary on opening day in thousands of dollars

```
##      Years            CAtBat         CHits            CHmRun
##  Min.   : 1.000   Min.   :  19   Min.   :   4.0   Min.   :  0.00
##  1st Qu.: 4.000   1st Qu.: 831   1st Qu.: 210.0   1st Qu.: 15.00
##  Median : 6.000   Median :1928   Median : 506.0   Median : 39.00
##  Mean   : 7.237   Mean   :2588   Mean   : 700.1   Mean   : 67.38
##  3rd Qu.:10.000   3rd Qu.:3754   3rd Qu.: 979.0   3rd Qu.: 90.00
##  Max.   :20.000   Max.   :9528   Max.   :2583.0   Max.   :548.00
##      CRuns             CRBI          CWalks       League      Salary
##  Min.   :   2.0   Min.   :   3   Min.   :   1.0   A:136   Min.   :  67.5
##  1st Qu.: 105.0   1st Qu.:  94   1st Qu.:  71.0   N:121   1st Qu.: 185.0
##  Median : 247.0   Median : 226   Median : 174.0           Median : 415.0
##  Mean   : 349.8   Mean   : 321   Mean   : 252.6           Mean   : 503.0
##  3rd Qu.: 488.0   3rd Qu.: 420   3rd Qu.: 319.0           3rd Qu.: 740.0
##  Max.   :1509.0   Max.   :1659   Max.   :1380.0           Max.   :1925.6
```

Partie 1

Q1 : Ecrivez un modèle linéaire régulier permettant d'expliquer le salaire (variable *Salary*) en fonction des variables *League* et *CRuns*. On l'appellera **mod1** par la suite.

Q2 : Ecrivez le modèle **mod1** sous forme matricielle $Y = X\theta + \varepsilon$.

Q3 : Construisez un intervalle de prédiction au niveau de confiance de 90% du salaire d'un joueur de baseball de league A avec 250 runs dans sa carrière.

Q4 : Ecrivez le modèle ajusté par la commande suivante sous R :

```
summary(lm(Salary~League*CRuns,data=Data))
```

```
##
## Call:
## lm(formula = Salary ~ League * CRuns, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -960.01 -161.67  -68.48  151.55 1019.68
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   203.79951   40.25707   5.062 7.97e-07 ***
## LeagueN        43.14185   57.82806   0.746    0.456
## CRuns           0.81174    0.08264   9.823  < 2e-16 ***
## LeagueN:CRuns  -0.03267    0.12511  -0.261    0.794
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 306.1 on 253 degrees of freedom
## Multiple R-squared:  0.3951, Adjusted R-squared:  0.388
## F-statistic: 55.09 on 3 and 253 DF,  p-value: < 2.2e-16
```

Q5 : Donnez la définition de la quantité `0.3951` de la sortie précédente.

Q6 : Construisez le test associé à la sortie suivante. Qu'en concluez-vous ?

```
## Analysis of Variance Table
##
## Model 1: Salary ~ CRuns
## Model 2: Salary ~ CRuns * League
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1    255 23780248
## 2    253 23709248  2     71000 0.3788 0.6851
```

Partie 2

Dans cette partie, on souhaite expliquer le salaire (variable *Salary*) en fonction de toutes les variables quantitatives. On considère donc un modèle de régression linéaire, supposé régulier, de la forme

$$Y = X\theta + \varepsilon \text{ avec } \varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n).$$

Q7 : Définissez le vecteur des valeurs ajustées et justifiez sa loi.

Q8 : Afin de chercher à simplifier le modèle pour gagner en interprétabilité, on met en place une méthode de sélection de variables à l'aide du code présenté avec le résultat en Figure 1.

- **Q8.a. :** Expliquez en quoi consiste cette procédure.
- **Q8.b. :** Donnez l'équation du modèle retenu par cette procédure de sélection de variables.

Q9 : On décide de mettre en place une régression régularisée avec une pénalité Lasso.

- **Q9.a. :** Ecrivez le critère minimisé par cette méthode.
- **Q9.b. :** Que représente la Figure 2 ? Comment l'obtient-on ?
- **Q9.c. :** Quel modèle est retenu pour un paramètre de régularisation de $e^{-3.2}$?

```
library(bestglm)
choix=regsubsets(Salary~.,data=Data[,-8],nbest=1,nvmax=11,method="forward")
plot(choix,scale="Cp")
```

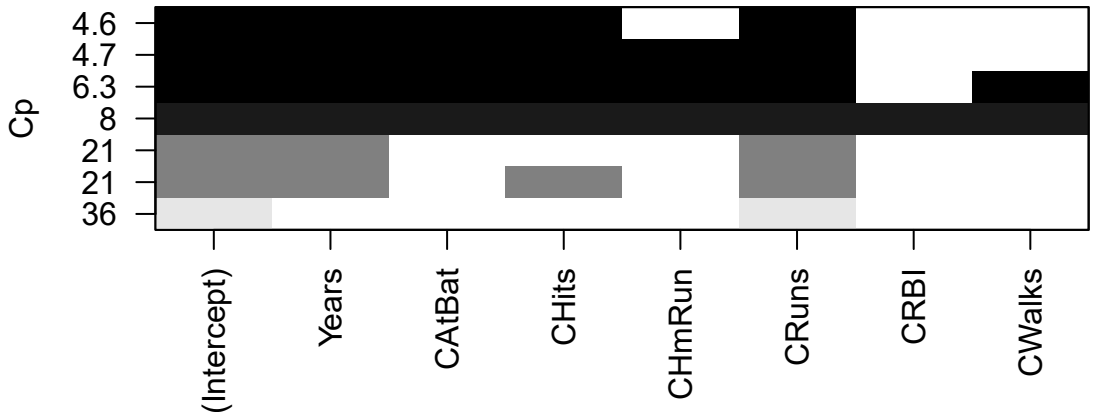

Figure 1: Figure pour la question Q8

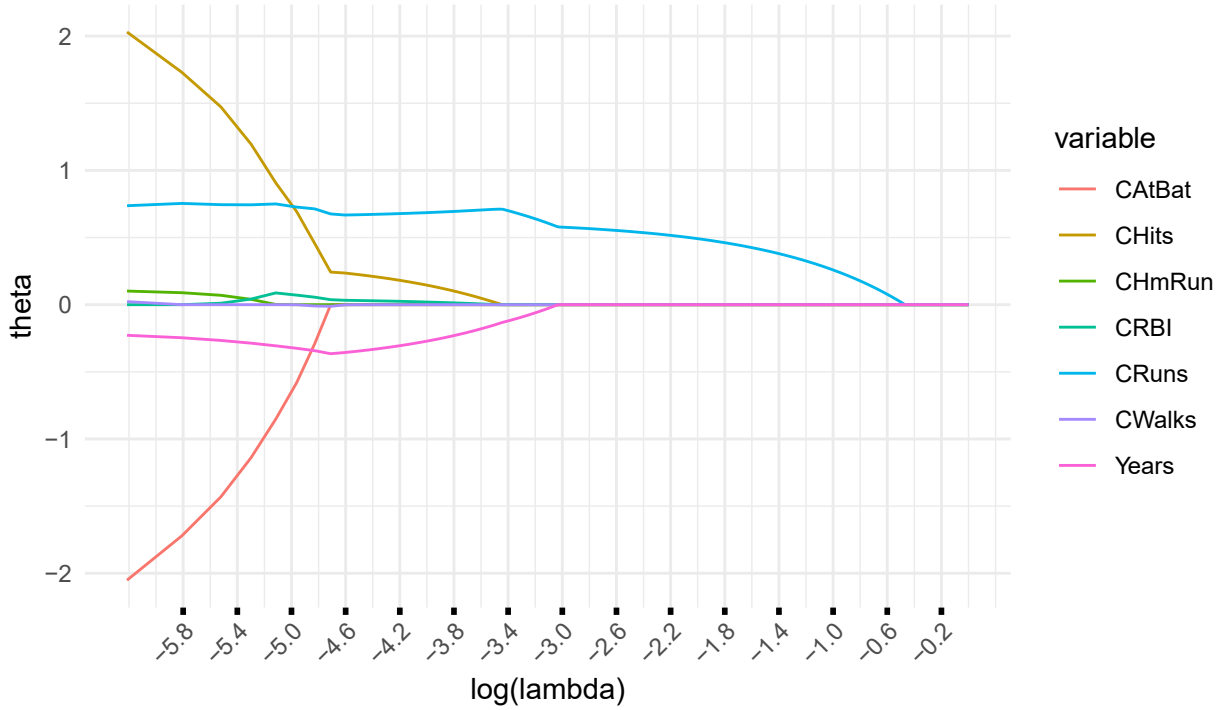

Figure 2: Figure pour la question Q9.b.

Exercice 2

On a mesuré la fréquence cardiaque de 12 femmes et 12 hommes ayant des activités de loisirs différentes (Natation, Pétanque et Pilates).

Sexe / Activite	Natation	Pilates	Pétanque	Moyenne
F	72, 69 73, 70 moy: 71	71, 73 73, 73 moy: 72.5	78, 83 79, 80 moy: 80	74.5
M	72, 67 71, 66 moy : 69	78, 77 76, 77 moy : 77	82, 81 80, 79 moy : 80.5	75.5
Moyenne	70	74.75	80.25	75

Dans la suite, on note Y_{sak} la fréquence cardiaque de la $k$ème personne de sexe s et d'activité a avec $a \in \{Natation, Pétanque, Pilates\} = \{1, 2, 3\}$, $s \in \{F, M\}$ et $k = 1, \ldots, 4$.

Partie 1

Q1 : Ecrivez le modèle linéaire ajusté ci-dessous pour expliquer les fréquences cardiaques Y_{ask} en fonction de la variable *sexe*.

```
summary(lm(freqC~Sexe,data=freqdata))
```

```
##
## Call:
## lm(formula = freqC ~ Sexe, data = freqdata)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9.50  -3.50  -0.50   3.75   8.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   74.500      1.409  52.880   <2e-16 ***
## SexeM          1.000      1.992   0.502    0.621
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.88 on 22 degrees of freedom
## Multiple R-squared:  0.01132,    Adjusted R-squared:  -0.03362
## F-statistic: 0.2519 on 1 and 22 DF,  p-value: 0.6207
```

Q2 : Testez la nullité du paramètre associé à la modalité "M" de la variable *sexe* dans la sortie ci-dessus. Concluez au risque 1%.

Partie 2

On s'intéresse maintenant à expliquer par un modèle linéaire la fréquence cardiaque en fonction des variables *sexe* et *activité*.

Q3 : Ecrivez un modèle linéaire pour répondre à cette question en prenant en compte une potentielle interaction entre les variables *sexe* et *activité*. On le notera **modF** dans la suite.

Q4 : Peut-on avoir des contraintes d'orthogonalité pour le modèle **modF** dans cette étude ? Si oui, énoncez ces contraintes d'orthogonalité.

Q5 : Sous les contraintes prises par défaut sous R, définissez les quantités suivantes à l'aide des paramètres du modèle **modF** :

- la fréquence cardiaque moyenne des femmes nageuses
- la fréquence cardiaque moyenne des hommes nageurs
- la fréquence cardiaque moyenne des femmes pratiquant la pétanque.

Q6 : Dessinez un graphique pour visualiser un potentiel effet d'interaction dans ce modèle.

Q7 : Construisez un test pour tester l'effet d'interaction dans le modèle **modF**. On a obtenu une pvaleur de 0.01, qu'en concluez-vous ?