Devoir surveillé de classification supervisée

Durée : 2h, trois feuilles recto-verso manuscrites autorisées, tout type de calculatrice autorisé

Nous disposons de données sur des patients atteints d'une hépatite ($n = 143$ patients, $d = 15$ variables). Dans cette étude nous nous intéresserons à prédire l'état des patients, variable `class` (`die` ou `live`), à partir des autres variables à disposition.

age	bilirubin	sgot	class	sex	steriod	antivirals	fatigue
Min. : 7.0	Min. :0.300	Min. : 14.00	die : 29	female : 15	no :69	no : 22	no :94
1st Qu. :32.0	1st Qu. :0.700	1st Qu. : 31.00	live :114	male :128	yes :74	yes :121	yes :49
Median :39.0	Median :1.000	Median : 58.00					
Mean :40.8	Mean :1.415	Mean : 82.85					
3rd Qu. :50.0	3rd Qu. :1.500	3rd Qu. :100.50					
Max. :78.0	Max. :8.000	Max. :648.00					

TABLE 1: Résumé des variables 1 à 8

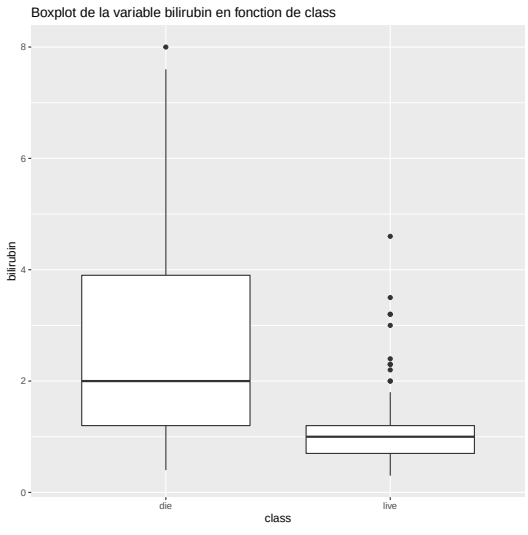
malaise	anorexia	spleen_palpable	spiders	asites	varices	histology
no :57	no : 29	no : 28	no :48	no : 20	no : 18	no :76
yes :86	yes :114	yes :115	yes :95	yes :123	yes :125	yes :67

TABLE 2: Résumé des variables 9 à 15

Analyses préliminaires

1. Le jeu de données de l'étude résulte en fait d'un jeu de données plus grand duquel on a supprimé les individus avec des données manquantes. Pourquoi ne peut-on pas prendre en compte le individus avec des valeurs manquantes lors de l'ajustement d'une régression logistique ?

On décide ici d'ajuster un modèle prédictif de la variable `class` en fonction de la variable `bilirubin`. Mais avant on réalise un graphique descriptif, ainsi qu'un test de l'ANOVA de l'effet de la variable `class` sur la variable `bilirubin` :




```
> summary(lm(bilirubin ~ class, data = d))

Call:
lm(formula = bilirubin ~ class, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-2.1724 -0.4202 -0.1202  0.1798  5.4276

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.5724     0.1963  13.105  < 2e-16 ***
classlive    -1.4522     0.2198  -6.606 7.49e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.057 on 141 degrees of freedom
Multiple R-squared:  0.2363,Adjusted R-squared:  0.2309
F-statistic: 43.64 on 1 and 141 DF,  p-value: 7.495e-10
```

2. Quel est la part de variance de la variable `bilirubin` expliquée par la variable `class` ?
3. Ici quelle est l'hypothèse nulle testée par le test de l'ANOVA ?
4. La variable `class` a-t-elle un effet significatif sur la variable `bilirubin` au niveau $\alpha = 0,05$?
5. La probabilité critique pourrait cependant être erronée du fait que toutes les hypothèses de l'ANOVA ne semblent par réunies ici. Quelle est l'hypothèse de l'ANOVA qui semble sûrement violée ici ?
   On lance `var.test(bilirubin  class, data = d)`, et on obtient :
   ```
   F test to compare two ???
   F = 8.907, num df = 28, denom df = 113, p-value < 2.2e-16
   ```

6. Que compare le F test ? Qu'en conclure ?

Ajustement d'un modèle de régression logistique

On décide maintenant d'ajuster la régression logistique de la variable `class` en fonction de la variable `bilirubin`. On lance `glm(formula = class  bilirubin, family = "binomial", data = d)` :

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   2.9953     0.4424   6.770 1.28e-11 ***
bilirubin    -1.0247     0.2337  -4.385 1.16e-05 ***

    Null deviance: 144.22  on 142  degrees of freedom
Residual deviance: 114.42  on 141  degrees of freedom
AIC: 118.42
```

7. A partir de l'ensemble des éléments à votre disposition, la modalité prédite pour la variable `class`, est-elle la modalité `dead` ou `live` ?
8. Quel est l'utilité du critère AIC ?
9. Quel est la probabilité de survie d'un patient avec une valeur de $0,75$ pour la `bilirubine` ?
10. Comment peut-on mesurer les performances du modèle ajusté ?

On décide maintenant d'ajuster le modèle avec l'ensemble des variables explicatives, puis de réaliser une sélection de variables pas à pas.

Les coefficients du modèle final sont les suivants :

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   18.6848  1476.1356   0.013  0.98990
bilirubin     -1.0764     0.3136  -3.432  0.00060 ***
sexmale      -16.8908  1476.1352  -0.011  0.99087
malaiseyes     1.7556     0.7035   2.495  0.01258 *
anorexiayes   -3.0382     0.9825  -3.092  0.00199 **
spidersyes     1.7582     0.6540   2.688  0.00718 **
asitesyes      2.2301     0.7186   3.103  0.00191 **
---

    Null deviance: 144.217  on 142  degrees of freedom
Residual deviance:  73.379  on 136  degrees of freedom
AIC: 87.379
```

11. Quel critère la sélection pas à pas cherche-t-elle à minimiser ?
12. Quel intérêt peut-il y avoir à effectuer une étape de sélection de variables ?
13. Quel est la probabilité de survie pour le patient suivant :

```
   bilirubin  sex malaise anorexia spiders asites
16         2 male      no       no     yes     no
```

14. Commenter la ligne correspondant au coefficient `sexmale` ?

En effectuant le croisement `table(d$sex, d$class)`, on obtient :

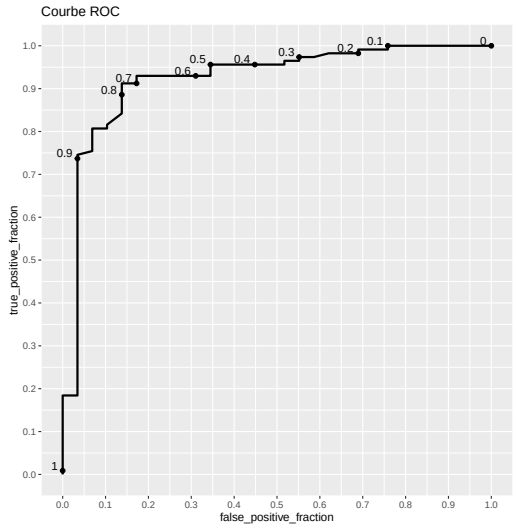
```
       die live
female   0   15
male    29   99
```

15. Cela explique-t-il le résultat obtenu ?

Soit β_0 et β_{male} les coefficients du modèle de la régression logistique de la variable `class` en fonction de la variable `sex`.

16. Exprimer $P(\text{class} = \text{live}|\text{sex} = \text{female})$ en fonction de β_0 et/ou β_{male}.
17. Pour coller aux données, quelle devrait être la valeur β_0 ? Que doit-il se passer dans le logiciel à votre avis ?

On trace la courbe ROC associée au modèle final, les points affichés représentent des valeurs particulières pour le seuil sur la probabilité.



18. Quelle quantité est représentée sur l'axe de x ? Sur l'axe des y ?
19. A combien faut-il fixer le seuil si on veut retrouver 95% des patients survivants ? Quel est le taux de faux positifs qui en découle ?
20. Approximativement quelle valeur optimale en terme de compromis sensibilité/spécificité suggère le graphique ? Quelles sont dans ce cas les valeurs de la sensibilité et de la spécificité qui en découlent ?

Questions sur l'analyse discriminante probabiliste

18. Rappeler le principe de l'analyse discriminante probabiliste.
19. Quelle est la différence entre analyse discriminante linéaire (LDA) et quadratique (QDA) ?
20. Est-il possible ici d'ajuster une LDA ou QDA à partir de l'ensemble des variables à votre disposition ? Pour quelle autre solution d'analyse discriminante probabiliste pourriez-vous opter ?