Statistique et Aide à la Décision 2022 - Examen 1ère session

Nom Prénom :

Tous les documents sont autorisés
Seul appareil électronique autorisé : calculatrice

Prédiction du génotype à partir de mesures biométriques

Le prix payé à une éleveuse ou un éleveur de porcs est fondée sur la teneur en viande maigre de ses animaux. Cette teneur est estimée à partir de mesures biométriques (épaiseurs de tissus gras ou maigres, poids). On sait que la connaissance du génotype de l'animal permettrait d'améliorer la précision de l'estimation de la teneur en viande maigre.

Or, il n'est pas possible aujourd'hui de disposer directement du génotype d'un animal au moment de l'évaluation de la teneur en viande maigre. En revanche, on peut penser qu'il est possible de prédire ce génotype à partir de mesures biométriques disponibles.

L'objectif de l'exercice est de **construire une règle de prédiction du génotype d'une carcasse à partir d'un profil de 16 mesures biométriques et de sexe de l'animal**.

Pour cela, on dispose de données pour 353 animaux dont les génotypes sont notés $\mathtt{P}_0$, $\mathtt{P}_{25}$ et $\mathtt{P}_{50}$:

```
'data.frame':   353 obs. of  18 variables:
 $ GGENE     : Factor w/ 3 levels "P0","P25","P50": 3 3 1 3 3 1 1 1 1 1 ...
 $ SEXE      : Factor w/ 2 levels "F","M": 2 1 1 1 2 2 1 2 1 2 ...
 $ LONGPA    : num  98.5 97 95.5 93.5 94.5 ...
 $ LONGLD    : num  82.5 81 80.5 79.5 80.5 85 88 86 86 89 ...
 $ NBCOTES   : int  15 15 14 14 16 15 15 16 14 15 ...
 $ NBCOTFL   : int  0 1 0 0 0 0 0 1 0 0 ...
 $ NBVL      : int  6 6 6 6 6 6 6 5 6 6 ...
 $ FRGRAS    : num  11.12 8.77 16.51 10.37 16.98 ...
 $ FRMUSCLE  : num  69.3 83.6 74.1 67.4 73.7 ...
 $ GR34VLFR  : num  14.1 13.4 20.4 15.1 24.8 ...
 $ GR23DCFR  : num  13.96 7.38 15.01 8.85 16.21 ...
 $ MU23DCFR  : num  58.2 71.1 60.8 62.8 61 ...
 $ GR34DCFR  : num  14.84 8.38 17.3 11.97 20.33 ...
 $ MU34DCFR  : num  56.2 68.2 56.8 57.9 54.2 ...
 $ TMUS3P    : num  81.8 87.3 79.1 84.1 76.1 ...
 $ TMUGIOSLON: num  75.3 82.2 72.5 77.6 68.2 ...
 $ TMUGIOSJAM: num  83 88.4 80 85.5 79.9 ...
 $ TMUGIOSEPA: num  81.7 88.6 81.6 85.8 75.8 ...
```

Toutes les variables du tableau sont quantitatives, sauf **SEXE** qui donne le sexe de l'animal (M pour Mâle, F pour Femelle) et (**GGENE**) qui donne son génotype.

Question 1

Donner l'expression mathématique du modèle le plus complet possible permettant d'expliquer le génotype à partir de toutes les variables explicatives.

Réponse

Question 2

Quel est le nombre de paramètres de ce modèle ?

Réponse

La table d'analyse de la déviance du modèle complet par rapport au modèle nul (celui sans aucune variable explicative) est reproduite dans le tableau suivant :

```
  Resid. df Resid. Dev   Test    Df LR stat.   Pr(Chi)
1       704    758.980                        
2       638    630.949 1 vs 2    ??       ?? 7.384e-06
```

Question 3

Quelle est l'hypothèse nulle du test mis en oeuvre dans cette analyse de la déviance ? (essayez de l'exprimer en langage non-mathématique)

Réponse

Question 4

Dans la table d'analyse de la déviance, quelles valeurs sont remplacées par des points d'interrogation ?

Réponse

Les commandes suivantes donnent la table d'analyse de la déviance de type II d'un sous-modèle du modèle complet :

```
select <- RcmdrMisc::stepwise(complet,
                              direction="forward/backward",
                              criterion="AIC",trace=0)

Direction:  forward/backward
Criterion:  AIC

car::Anova(select)

Analysis of Deviance Table (Type II tests)

Response: GGENE
            LR Chisq Df Pr(>Chisq)
MU34DCFR      31.509  2   1.44e-07 ***
LONGLD        18.848  2   8.08e-05 ***
GR34VLFR       9.081  2   0.010670 *
FRGRAS        20.817  2   3.02e-05 ***
TMUGIOSJAM    16.342  2   0.000283 ***
SEXE           4.563  2   0.102139
LONGLD:SEXE    5.179  2   0.075048 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 5

Pourquoi ce sous-modèle est-il particulièrement intéressant, par rapport à tous les autres sous-modèles possibles ?

Réponse

On donne ci-après une table d'analyse de la déviance de type I comparant le sous-modèle au modèle nul et le modèle complet au sous-modèle :

```
anova(nul,select,complet)[,-1]
```

```
  Resid. df Resid. Dev   Test    Df LR stat.     Pr(Chi)
1       704    758.980
2       690    666.669 1 vs 2    14  92.3112 1.38667e-13
3       638    630.949 2 vs 3    52  35.7197 9.58617e-01
```

Question 6

Quelle est la valeur de l'AIC du sous-modèle ? (vous pouvez vous contenter de donner l'opération permettant d'obtenir le résultat, sans donner le résultat)

Réponse

Question 7

Quelle est l'hypothèse nulle dont la p-value du test apparait à la dernière ligne, dernière colonne de la table d'analyse de la déviance ?

Réponse

Le tableau suivant permet de comparer les génotypes prédits à partir du sous-modèle et les génotypes observés :

```
predictions <- predict(select,type="class")
confusion <- table(geno$GGENE,predictions,dnn=list("Observé","Prédit"))
RcmdrMisc::rowPercents(confusion)
```

```
       Prédit
Observé   P0  P25  P50 Total Count
    P0  22.9 37.3 39.8 100.0    83
    P25  9.3 65.7 25.0 100.0   140
    P50 10.8 28.5 60.8 100.1   130
```

Question 8

Expliquez la règle de prédiction mise en oeuvre ci-dessus (comment la valeur prédite est-elle obtenue à partir d'un profil de variables explicatives ?).