Examen janvier 2017

1 Questions de cours

1. Sous quelles conditions les régles de décisions d'une analyse factorielle discriminante et d'une analyse discriminante linéaire sont-elles les mêmes ?
2. Quelle est la signification de l'expression "winsoriser une variable" ? Quelles sont les variables ciblées par cette opération ?

2 Exercice

Une enquête a été réalisée sur les résultats d'un concours il y a quelques années. Cette étude avait pour objectif de rechercher des facteurs de 'risque' ou des facteurs 'protecteurs' pour l'échec à ce concours.
Les caractéristiques suivantes ont été relevées pour 8555 candidats à ce concours :
NATIONALITE : France ou autre
BOURSE : O = oui, N= non
NBRINSP1 = nombre d'inscription(s) au concours (1, 2= redouble ou 3 = triple)
BAC = section S ou autre
MENTION : TB, B, AB ou P
SEXE = F (fille) M (garçon)
PROFPAR (profession des parents) :

- agri = agriculteur
- autr= autre profession
- cadr= cadres
- empl= employés
- interm= artisans, professions intermédiaires

SITFAM (situation familiale du candidat)

- seul
- coupenf : couple avec enfant
- couplSE : couple sans enfant

RESULTAT : ADM (admis), AJ (ajourné=échec)

Des analyses bivariées ont été réalisées entre la variable "résultats" et chacune des autres variables. Ensuite, une analyse statistique utilisant le modèle de régression logistique a été réalisée pour expliquer l'échec au concours (Y=1 si RESULTAT=AJ).

1. Quel est l'intérêt de réaliser des analyses bivariées avant le modèle de régression logistique ?
2. On a utilisé un test du χ^2 pour tester la liaison entre NATIONALITE et RESULTAT. Comment justifiez-vous ce choix ? Quelle est l'hypothèse nulle du test choisi ? On obtient une p-value $< 10^{-6}$. Comment interprétez-vous cette valeur ?
3. Expliquez en quelques mots pourquoi la régression logistique peut répondre à l'objectif de l'étude. Quelle fonction utiliseriez-vous en R pour lancer ce modèle ? On appelle `model1` le modèle complet.
   On obtient la sortie suivante :

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	4,6692357	0,58855120	7,933440	2,131572e-15
nationalitefrance	-1,0806833	0,27957801	-3,865409	1,109033e-04
bourse0	0,3554768	0,08232768	4,317828	1,575722e-05
nbrinsp1	-2,2173473	0,06650147	-33,342831	9,253867e-244
bacS	-1,2389109	0,29732431	-4,166867	3,088145e-05
mentionB	-1,3659765	0,08153160	-16,753952	5,298619e-63
mentionP	1,1790927	0,08319787	14,172149	1,362650e-45
mentionTB	-2,7200549	0,12079565	-22,517822	2,776816e-112
sexeM	-0,2884984	0,06442539	-4,478024	7,533719e-06
profparautr	1,0520769	0,32203645	3,266950	1,087130e-03
profparcadr	0,4173471	0,31468266	1,326247	1,847577e-01
profparempl	0,7654281	0,32555535	2,351146	1,871570e-02
profparinterm	0,6159180	0,31565450	1,951241	5,102837e-02
sitfamcouplSE	0,5246656	0,40451658	1,297019	1,946247e-01
sitfam seul	1,6246615	0,37018085	4,388832	1,139612e-05

4. En prenant l'exemple de la variable PROFPAR, expliquez comment sont gérées les variables qualitatives dans cette analyse.

5. Donnez l'odd-ratio associé à la variable bourse et interprétez sa valeur.

6. On lance la commande `model2=step(model1)`. Expliquez à quoi sert cette commande de manière générale. Ici, quelles variables pensez-vous retrouver dans `model2` (justifiez votre réponse) ?

7. En lançant ensuite `exp(cbind(OR=coef(model2), confint(model2)))`, on obtient :

	OR	2,5 %	97,5 %
(Intercept)	106,61622797	34,27514729	345,00560478
nationalitefrance	0,33936356	0,19174434	0,57550232
bourse0	1,42686077	1,21484812	1,67766310
nbrinsp1	0,10889760	0,09547068	0,12390984
bacS	0,28969956	0,15649930	0,50433219
mentionB	0,25513141	0,21728458	0,29912618
mentionP	3,25142286	2,76456086	3,83087435
mentionTB	0,06587113	0,05188487	0,08332105
sexeM	0,74938799	0,66039515	0,85015545
profparautr	2,86359237	1,50846231	5,34569314
profparcadr	1,51792927	0,81075420	2,79203349
profparempl	2,14991457	1,12504377	4,04208691
profparinterm	1,85135541	0,98705838	3,41206707
sitfamcouplSE	1,68989371	0,76846016	3,76367113
sitfam seul	5,07670006	2,47252614	10,59289998

   Quels sont les facteurs protecteurs ? Quels sont les facteurs de risque ? Justifiez vos réponses.

8. On lance les commandes suivantes :

```
pi_hat=predict(model1, resultat, type="response")
Y_hat=as.factor(ifelse(pi_hat>0.5, "echec", "admis"))
```

```
MatConf=table(Y_hat, resultat)
MatConf
```

```
       resultat
Y_hat    ADM   AJ
  admis 1104  532
  echec  955 5964
```

Commentez ces lignes de code et donnez le taux de bon classement.

9. On remarque que l'aire sous la courbe ROC est de 0.87 et que le test de Hosmer-Lemeshow donne une p-value de 0.32. Commentez ces valeurs.

10. Quelle approche proposeriez-vous pour obtenir le meilleur compromis entre sensibilité et spécificité ? (Décrivez la méthode sans chercher à mettre en oeuvre les calculs).

    Dans la suite de l'exercice, on s'intéresse à un étudiant qui présente les caractéristiques suivantes :

    — profession parents = intermédiaire
    — mention BAC = TB
    — situation familiale = seul
    — sexe = homme
    — bourse = oui
    — BAC = S
    — nationalité = autre

11. Quelle est la valeur du score prédictif $\ln\left(\frac{\pi}{1-\pi}\right)$? (indiquez le mode de calcul)

12. Quelle est la valeur de la probabilité d'échec prédite ?