Polytech'Lille GIS 4e Année
2017-2018

Examen janvier 2018

Guillemette Marot, Vincent Vandewalle

Une seule feuille A4 recto-verso contenant des notes de cours ou TP autorisée, calculatrice autorisée

1 Question de cours

Dans quel contexte utilise-t-on une méthode d'imputation ? Quelles autres approches pourraient être utilisées pour traiter ce type de valeurs ?

2 Régression logistique

On considère un jeu de données sur la mort inattendue du nourrisson. Les 449 nourrissons inclus dans l'étude sont morts de façon inattendue. L'objectif est de retrouver des facteurs de risque de la mort par asphyxie. Les variables étudiées sont les suivantes :

- cosleeping : le nourrisson dormait dans le lit des parents
- chbParents : le lit du nourrisson était dans la chambre des parents
- tetine : le nourrisson dormait avec sa tétine
- couverture : le nourrisson dormait avec une couverture
- oreiller : le nourrisson dormait avec un oreiller
- prematurité : le nourrisson était né prématurément
- tabacMere : la mère est fumeuse

Toutes ces variables sont binaires et codées (1 : oui /0 : non).

Partie A

1. Avant d'ajuster le modèle prédictif, quelles analyses bivariées proposeriez-vous pour répondre à l'objectif de l'étude ? Justifiez.
2. On commence par étudier le lien entre les variables asphyxie (Y) et tabacMere (X). On obtient le tableau de contingence suivant :

```
          asphyxie
tabacMere   0   1
        0 239 107
        1  23  80
```

   Donnez le tableau des effectifs attendus sous l'hypothèse d'indépendance des deux variables.
3. Calculez le V de Cramer associé à cette liaison en détaillant les calculs.
4. Calculez la probabilité de mort par asphyxie sachant que la mère est fumeuse.

5. Calculez l'odd ratio associé à la variable tabacMere à l'aide du tableau de contingence.
6. On appelle $\pi(x) = P(Y = 1/X = x)$ la probabilité de mort par asphyxie sachant le tabagisme de la mère. On décide de construire un score linéaire, égal au logit de cette probabilité. Rappeler l'écriture du logit en fonction de $\pi(x)$ et donner son expression sous forme linéaire de x en utilisant les notations β_0 et β_1.
7. Calculer β_1 à partir de l'odd-ratio calculé précédemment.
8. On donne $\beta_0 = -0.80$. Calculez la valeur du score linéaire pour une mère fumeuse.
9. En déduire la probabilité de mort par asphyxie sachant que la mère est fumeuse. Comparez au résultat obtenu à partir du tableau de contingence.

Partie B :

On décide d'inclure toutes les variables dans la régression logistique sans effectuer de sélection de variables. On obtient le tableau suivant :

	OR	2,5 %	97,5 %
(Intercept)	0,2365357	0,1519389	0,3598906
cosleeping	2,9941210	1,8664112	4,8451491
chbParents	1,0877080	0,6554647	1,7910046
tetine	0,8846219	0,5503957	1,4121902
couverture	1,6767903	1,0254841	2,7432006
oreillers	1,1745373	0,7644629	1,8052349
prematurite	1,8525105	1,1512195	2,9823298
tabacMere	6,8551600	4,0411227	11,9949570

1. Quels sont les facteurs de risque ? Quels sont les facteurs protecteurs ? Justifiez les réponses.
2. Quelle commande R utiliseriez-vous pour lancer une procédure de sélection de variables pas à pas descendante ?

Partie C :

Dans cette partie, on utilise comme score la probabilité de mort par asphyxie sachant toutes les variables explicatives (et non plus le logit de la probabilité comme dans la partie A).

On choisit arbitrairement un seuil à 0,5 et on décide d'affecter tous les bébés ayant un score supérieur à ce seuil au groupe "mort par asphyxie".

1. Comment appelle-t-on le tableau de contingence renvoyé ci-dessous ?

	Ypredit	
Yreel	0	1
0	226	36
1	84	103

2. Calculer le taux de bon classement.
3. Donner le nombre de vrais positifs (VP), faux positifs (FP), vrais négatifs (VN), faux négatifs (FN)

4. Calculer sensibilité et spécificité pour ce seuil.

5. On récupère les valeurs de sensibilité et spécificité pour les seuils 0.2, 0.3, 0.4 et 0.5.

```
     seuil specificite sensibilite
[1,]   0,2        0,29        0,93
[2,]   0,3        0,53        0,75
[3,]   0,4        0,72        0,66
[4,]   0,5        0,86        0,55
```

6. Quel est le seuil qui donne le meilleur compromis entre sensibilité et spécificité ? Justifiez.

3 Analyse discriminante

1. Quelle est la différence entre une analyse discriminante linéaire et une analyse discriminante quadratique ? Quelles sont les fonctions de R permettant de lancer ces analyses ?

2. A partir des lignes de code suivantes, donner les lois de Y et de X/Y.

```
Y=rbinom(1000,1,0.5)
mean0=c(0,0)
mean1=c(2,2)
sigma0=matrix(c(1,0,0,1),ncol=2)
sigma1=matrix(c(1,0.8,0.8,1),ncol=2)
library(mvtnorm)
X<-matrix(NA,1000,2)
for (i in 1:1000) {
  if (Y[i]==0) {
    X[i,]=rmvnorm(1, mean = mean0, sigma = sigma0);
  } else {
    X[i,]=rmvnorm(1, mean = mean1, sigma = sigma1);
  }
}
```

3. On a calculé pour l'individu de coordonnées $(1, 2)$ les densités de probabilité dans chacune des classes et obtenu 0.01 pour la classe Y=0 et 0.06 pour la classe Y=1. Donnez les probabilités d'appartenance de l'individu $(1, 2)$ à la classe Y=0 d'une part et à celle Y=1 d'autre part.

4. En déduire sa classe prédite à l'aide de la règle du maximum a posteriori.