Devoir no. 1.2- MAT 898A-Aut 2015

Analyse de données (pratique avec "R"), 60 pt.

Problème. Analyser (modèle logistique) les données de l'étude de cohorte du fichier *data-icu-labo.txt* sur l'admission à une unité d'urgence. Soit Y la variable binaire dépendante (`DEP`=`STA`), avec $Y = 1$ si la personne a décédé aux soins intensifs. La liste complète des variables explicatives est dans le fichier *data-icu-description.txt*; trois sont de type **quantitatif** (et continues), *âge*, *HRA (heart rate)*, *SYS (systolic blood pressure)*; les autres sont de type **qualitatif**, comme *race* (1,2,3); *sexe*, *type d'admission* (en urgence ou non), ainsi que divers constats médicaux présents ou absents=(PO2, PH, PCO, etc). Les variables qualitatives doivent être entrées comme "facteurs" (les déclarer comme telles). Pour simplifier, on ne prend pas toutes les variables, et, dès le départ, on élimine de l'analyse le PO2, PH, PCO, BIC, donc on garde juste CRE dans la liste de tests sanguins. On va se mettre dans la peau d'un gérant pressé (mais pas vrai statisticien) qui veut identifier quelques variables qui peuvent lui indiquer les personnes à risque qui rentrent dans son unité et pour lesquelles il faut agir rapidement; comme il est clair qu'une personne qui est en état grave (LOC) sera traitée en priorité, au début on ne considère pas cette variable (suggestion du gérant). Donc on part avec 14 variables et on vise à réduire à 5-6 maximum (plus d'éventuelles interactions). **Note**: on peut sauvegarder des commandes "R" dans un fichier texte et après ça les télécharger chaque fois avec la commande `source(nom.fichier)`.

Questions.

i. Pour décider quelles variables garder dans l'équation je suggère de procéder ainsi (pour la commodité, l'approche est un peu artificielle) :

- on commence par faire une commande `glm()` et ensuite on applique la fonction `step()`; ainsi on garde un nombre de variables parmi les 14 proposées par le gérant;
- avec ce bloc de variables on fait des `anova()` (en changeant l'ordre des variables) pour voir pourquoi une variable qui ne semble pas significative si rentrée dans l'ordre de la liste de départ est quand même retenue dans l'équation par la fonction `step()`;
- avec les variables retenues et dans l'ordre établi précédemment, on ajoute la variable *oubliée* (LOC) afin de regarder et commenter l'effet qu'elle a (dommage la laisser de côté !);
- on ajoute une interaction, je propose essayer entre l'âge et une autre variable parmi celles retenues, cela donne l'équation finale.

N'oubliez pas d'utiliser la commande complète `anova(...,test="Chisq")` qui ajoute des *p-values* à la liste de deviances.

ii. Écrire l'équation retenue, identifier les "odds ratios" (rapports de cotes) qui correspondent à chacune des variables (en faisant attention aux interactions, si jamais elles sont présentes). Indiquer ce que les "odds ratios" représentent pour chaque variable.

iii. Après ça, comparer le "pronostic" (transformée logit de la probabilité de décéder à l'ICU) de deux personnes qui arrivent à l'urgence et dont les caractéristiques sont (on considère quatre situations): (a) 50 ans versus 70 ans; (b) arrivée en ambulance ou pour autres raisons (variable TYP, 0 ou 1); (c) femme ou homme; (d) pression arterielle 120 versus 140.

iv. Commenter brièvement l'ajustement du modèle à partir de la deviance finale, D_R; (facultatif : nuage de résidus fournis par le programme); faire aussi la courbe ROC.

v. On veut comparer l'analyse stratifiée (jusqu'à l'estimateur Mantel-Haenszel) à une régression logistique et montrer qu'on retrouve les mêmes résultats : je propose cela pour illustrer la théorie du début, où on a comparé les "odds ratios" obtenus par les deux méthodes (analyse de tableaux et régression logistique). Donc on va étudier l'association entre le décès et le type de soins (chirurgical ou médical, colonne 6, SER) selon le sexe (femme ou homme, colonne 4, SEX). On fait donc une analyse par tableaux (d'une part) et une régression logistique où les seules variables indépendantes sont SER et SEX, d'autre part. Les variables sont déjà codées ! Il faut arriver par calculer les *odds-ratios* pour les strates et retrouver la réponse dans la sortie "R". Il y a un calcul *à la main*, mais on peut faire appel à 'R" pour former les tableaux 2×2. Par exemple, une façon directe pour compter tous les décès, variable DEP=1, dont SER=1 et SEX=1, se fait avec la commande:

```
length(DEP[SER==1 & SEX==1 & DEP==1])
```

vi. Une application intéressante de la régression logistique est en classification, et on donne un aperçu au cours. Après avoir fait l'estimation par régression logistique on a les probabilités estimées $\widehat{\widehat{p}_i}$ (de la réponse d'intérêt, oui ou non) et les valeurs observées, y_i, $i = 1, \ldots, n$ (oui ou non). Cela permet de diviser les sujets selon 2 critères, un critère étant leur état (oui ou non), l'autre obtenu en comparant leur probabilité estimée avec une probabilité fixe, choisie par l'utlisateur, p_0, par exemple 50%, 25%, etc. Si $\widehat{\widehat{p}_i} > p_0$ on dit que le sujet a reçu un diagnostic positif, sinon le diagnostic est négatif. On peut former un tableau, à 4 catégories:

$$
\begin{array}{lccc}
 & B:\ \widehat{\widehat{p}} > p_0 & \overline{B}:\ \widehat{\widehat{p}} \leq p_0 & \text{Total} \\
A:\ Y=1 & n_{11} & n_{12} & r_1 \\
\overline{A}: Y=0 & n_{21} & n_{22} & r_2 \\
\text{Total} & c_1 & c_2 & n
\end{array} \tag{1}
$$

où on a mis B="le résultat au test médical est positif" ($\widehat{\widehat{p}} > p_0$), etc. La sensibilité dans ce cas est $\Pr(B|A)$ et s'estime par n_{11}/r_1. Ainsi, on peut appliquer ce "test médical" à toute personne dont le vecteur de variables explicatives est connu, car on peut calculer son $\widehat{\widehat{p}}$ et le comparer à la valeur p_0 (dire si la personne a un diagnostic positif ou négatif). Si le "test médical" a une bonne sensibilité et spécificité, on peut utiliser ce $\widehat{\widehat{p}}$ pour annoncer à la personne qu'elle est à risque ou non. Clairement la sensibilité et spécificité changent selon la valeur de p_0 et en pratique on veut choisir un p_0 afin d'optimiser les deux.

a. Expliquer comment on calcule (estime) la spécificité à partir de ce tableau.
b. Calculer effectivement la spécificité et la sensibilité dans le cas de nos données, en créant des tableaux de type (1) pour plusieurs valeurs de p_0. (Cela demande de trouver les *fitted* values.) Sinon, représenter la courbe ROC, et trouver des valeurs à partir de cette courbe.
c. Décider d'un p_0 optimal (approximatif) et aider les gestionnaires de l'unité de soins intensifs à décider dans quels cas il faut agir plus vite.

Date de remise (proposée) du Devoir 1.2: le jeudi 12 novembre.

Note sur quelles sorties d'ordinateur remettre.

- Pour le Devoir 1.2, mettre en annexe:

  A. les sorties finales avec des variables retenues (sans et avec interaction), c-à-d l'ANOVA où toutes les variables sont significatives (à 10% disons);
  B. (le graphique de résidus si c'est le cas) et la courbe ROC à la question ii);
  C. la sortie "R" pour la question (v) ;
  D. la courbe ROC à (vi).

- Pour le Devoir 1.1 il y a les résultats du test de Fisher et de la question 3)–(v).