Régression logistique

3 novembre 2021

Ce laboratoire doit être remis le **17 novembre à 17h sur Moodle**. Dans votre réponse pour chaque question, veuillez inclure une copie du code R utilisé (s'il y a lieu) et des résultats obtenus.

1. Retour sur les données de puits

Chargez le package *carData* pour avoir accès au tableau de données `Wells`, contenant les données d'une étude sur des puits à haute concentration en arsenic au Bangladesh.

```
library(carData)
head(Wells)
```

```
##   switch arsenic distance education association
## 1    yes    2.36   16.826         0          no
## 2    yes    0.71   47.322         0          no
## 3     no    2.07   20.967        10          no
## 4    yes    1.15   21.486        12          no
## 5    yes    1.10   40.874        14         yes
## 6    yes    3.90   69.518         9         yes
```

Durant le cours sur la régression logistique, nous avions modélisé la décision d'un ménage de changer de puits ou non (`switch`) en fonction de la concentration d'arsenic du puits (`arsenic`, en multiples de 100 $\mu g/L$) et de la distance au puits sûr le plus près (`distance`, en m). Le but de cet exercice est d'évaluer l'effet du prédicteur `education`, représentant le nombre d'années de scolarité de l'adulte répondant pour le ménage.

a) Estimez les paramètres d'un modèle où les trois variables (concentration d'arsenic, distance et scolarité) ont un effet additif. Interprétez *de façon qualitative* le coefficient de la variable `education`: selon le signe de ce coefficient, quel effet a le niveau de scolarité sur la décision de changer de puits?

b) Dans le contexte de cette étude, que signifierait une **interaction** entre le niveau de scolarité et la concentration d'arsenic sur la décision de changer de puits? Formulez une hypothèse sur la direction possible de cette interaction, avec une brève explication de votre choix.

c) Estimez les paramètres du modèle avec une interaction entre le niveau de scolarité et la concentration d'arsenic, puis vérifiez si le résultat correspond à votre hypothèse formulée en (b).

d) En vous basant sur les valeurs de l'AIC, comparez les modèles en (a) et (c) avec le modèle de base du cours (`switch ~ arsenic + distance`). Quel est le meilleur modèle? Est-il raisonnable d'effectuer des prédictions avec ce modèle seulement?

e) À partir d'un graphique de diagnostic approprié, vérifiez que les résidus du meilleur modèle en (d) sont distribués conformément au modèle de régression logistique.

f) Produisez un graphique de la probabilité de changer de puits (axe des y) prédite par le meilleur modèle en (d) pour des valeurs de la variable arsenic (axe des x) variant entre 0.5 et 5 (i.e. 50 à 500 $\mu g/L$), avec trois lignes correspondant aux niveaux de scolarité de 0, 5 et 10 ans. La variable `distance` n'apparaîtra pas sur le graphique, mais vous pouvez utiliser une distance constante de 50 m pour les prédictions.

2. Incidence du cancer de l'oesophage

Le tableau de données `esoph` inclus dans R contient des données d'une étude menée en France sur l'incidence du cancer de l'oesophage en fonction de l'âge, de la consommation d'alcool et de tabac.

```
head(esoph)
```

```
##   agegp     alcgp    tobgp ncases ncontrols
## 1 25-34 0-39g/day 0-9g/day      0        40
## 2 25-34 0-39g/day    10-19      0        10
## 3 25-34 0-39g/day    20-29      0         6
## 4 25-34 0-39g/day      30+      0         5
## 5 25-34     40-79 0-9g/day      0        27
## 6 25-34     40-79    10-19      0         7
```

Les données indiquent le nombre de sujets affectés (`ncases`) et non-affectés (`ncontrols`) pour chaque combinaison des catégories des trois prédicteurs.

Le tableau original utilise des facteurs ordonnés pour chaque prédicteur. Puisque l'analyse de variables ordinales n'est pas vue dans ce cours, nous convertissons ces prédicteurs en facteurs non-ordonnés (variables nominales).

```
library(dplyr)
esoph <- mutate(esoph, agegp = factor(agegp, ordered = FALSE),
                alcgp = factor(alcgp, ordered = FALSE),
                tobgp = factor(tobgp, ordered = FALSE))
```

a) Ajustez un modèle de régression logistique binomiale à ces données, en supposant un effet additif des trois prédicteurs, puis évaluez le coefficient pseudo-R^2 de McFadden du modèle.

b) Lequel des deux facteurs de risque (alcool ou tabac) augmente le plus l'incidence du cancer de l'oesophage pour cette population? Justifiez votre réponse.

c) Répondez aux questions suivantes en vous basant sur les coefficients du modèle et en utilisant la fonction `plogis`, qui permet convertir la valeur du prédicteur linéaire en probabilité.

- Comment interprétez-vous la valeur du coefficient `Intercept`?
- Quelle est la probabilité d'incidence du cancer pour une personne âgée entre 55 et 64 ans qui consomme moins de 39 g d'alcool et moins de 9 g de tabac par jour?

d) Si on inversait les nombres des cas atteints (`ncases`) et non-atteints (`ncontrols`) dans la formule du modèle: `cbind(ncontrols, ncases) ~ agegp + alcgp + tobgp`, quel serait l'effet sur les coefficients estimés? Tentez de prédire la réponse selon votre connaissance du modèle logistique, puis vérifiez avec R.