Modèle linéaire généralisé
Feuille 1 : régression logistique

Robin Ryder

Mai 2019

1 Rapports de cotes

1. Donner la cote anglaise des événements suivants :
   a) Obtenir pile lors du lancer d'une pièce non pipée
   b) Obtenir un 6 lors du lancer d'un dé non pipé
   c) Obtenir autre chose qu'un 6 lors du lancer d'un dé non pipé
   d) L'événement qu'un nouveau-né soit un garçon, sachant qu'il naît 105 garçons pour 100 filles
   e) Des événements de probabilité respective 10^{-6}, 0.001, 0.01, 0.1, 0.25, 0.4, 0.8, 0.99, 0.999.

2. Donner le rapport des cotes dans les situations suivantes :
   a) Lors d'une épidémie de varicelle dans l'Oregon en 2002 [1], 18 des 152 enfants vaccinés ont été infectés, et 3 des 7 enfants non vaccinés ont été infectés. Quel est le rapport des cotes d'infection entre les cohortes vaccinée et non vaccinée ?
   b) Le taux de décès par cancer du poumon dans les années 1960 était de 1.30 par 1000 personnes et par an chez les fumeurs, de 0.07 chez les non-fumeurs, et de 0.67 chez les personnes ayant arrêté de fumer il y a moins de 5 ans [2]. Donner les rapports de cotes pertinents.

2 Données Titanic : exploration préalable

Nous allons analyser le jeu de données `titanic.csv`, qui contient des informations au sujet de 891 passagers du RMS *Titanic*. Nous allons chercher à expliquer la colonne `Survived`, variable binaire qui indique si le passager a survécu ou non. Les autres variables sont :

- `pclass` : classe du billet (1, 2 ou 3)
- nom
- sexe

1. Tugwell et al. *Pediatrics*, 2004
2. Doll & Hill, *Brit Med J*, 1964

- âge en années
- `sibsp` : nombre de conjoint, de frères et de sœurs à bord
- `parch` : nombre d'enfants et de parents à bord
- `ticket` : numéro du billet
- `fare` : prix du billet
- `cabin` : numéro de cabine
- `embarked` : port d'embarquement (C=Cherbourg, Q=Queenstown, S=Southampton)

1. Charger les données sous R à l'aide de la fonction `read.csv()`.
2. Explorer les données rapidement.
3. Quelles variables sont corrélées à la survie ?
4. Y a-t-il beaucoup de données manquantes ? Y a-t-il des colonnes à exclure d'emblée de notre analyse ? Y a-t-il des colonnes à recoder ?

3 Régression logistique

5. Effectuer une première régression logistique avec l'ensemble des variables qui pourraient être pertinentes. Interpréter les résultats.
6. Proposer et tester des transformations de données qui pourraient être pertinentes comme variables explicatives (effets de seuil, interaction de variables...)

4 Optionnel : algorithme de Newton

Nous allons coder l'algorithme de Newton, qui permet de trouver le minimum (ou le maximum) d'une fonction f de classe $\mathcal{C}^3$. On se donne un point initial x_0, et à l'itération n on pose

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}$$

et on itère jusqu'à ce que

$$|x_{n+1} - x_n| < \epsilon$$

où τ est une tolérance choisie à l'avance, par exemple $\epsilon = 10^{-4}$.

Si la dérivée f' n'est pas disponible, on peut l'approcher par

$$f'(x) \approx \frac{f(x+h) - f(x)}{h}$$

pour une valeur h petite, et de même pour la dérivée seconde.

Implémenter cet algorithme d'optimisation, et le mettre en pratique pour minimiser la fonction $x \mapsto x^2 \cos x + e^{-x}$ sur l'intervalle $[0, 1]$.