Nous travaillons sous R. Tous les tests sont à 10%. Sauf demande explicite du test de rapport de vraisemblance, vous utilisez les tests de Wald.

1. Supports

Le site de notre cours est http://eric.univ-lyon2.fr/~ricco/cours/cours_regression_logistique.html

Plus spécifiquement pour cette séance, nous nous référerons à :

TUTO 1 – http://eric.univ-lyon2.fr/~ricco/cours/slides/regression_logistique.pdf

TUTO 2 – http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf

TUTO 3 - http://tutoriels-data-mining.blogspot.com/2012/03/introduction-r-regression-logistique.html

2. Données

On cherche à déterminer **les facteurs déterminant le ronflement** à partir des caractéristiques des personnes (âge, poids, taille, alcool [niveau de consommation], sexe [femme = 1], tabac [fumeur = 1]).

3. Exercices – Etude du ronflement

0. Importez le fichier « **etude_ronflement.xlsx** » (**TUTO 3**, page 7). Affichez la liste (str) et le résumé (summary) des variables. De combien d'observations et variables disposons-nous ? (100 x 7). Quelle est la proportion des personnes qui ronflent dans le fichier ? (35 %)
1. Réalisez la régression logistique sur l'ensemble des variables (**TUTO 3**, page 8, glm). Quelles sont les variables explicatives significatives au sens du test de Wald ? (summary) (à 10% : âge, alcool, tabac).
2. Est-ce que l'on peut supprimer le POIDS de la régression ? Utilisez le test du rapport de vraisemblance (**TUTO 1**, pages 26 et 28 ; comparez les déviances des modèles) (oui, les données ne contredisent pas la nullité du coefficient associé : KHI-2 = 0.2157, DDL = 1, p-value = 0.6422).
3. Est-ce que l'on peut supprimer simultanément le POIDS et la TAILLE de la régression ? Utilisez le test du rapport de vraisemblance (**TUTO 1**, page 29) (oui, test des coefficients simultanément nuls, KHI-2 = 0.304, DDL = 2, p-value = 0.8589).
4. Le SEXE pris isolément (comme seule variable explicative) est un facteur de différenciation. Est-ce vrai ? Comment interpréter le coefficient fourni par la régression logistique ? (SEXE = 1 ➔ « femme » ; coefficient négatif ; on a moins de chances de ronfler lorsque l'on est une femme).
5. Pris isolément, le fait même de consommer de l'alcool (oui/non), indépendamment du niveau, est un facteur de risque. Est-ce vrai ? Comment avez vous procédé ? (oui, après recodage de la variable ALCOOL).

6. L'interaction consommation alcool (oui/non) et tabac (oui/non) est elle significative ? (**TUTO 1**, page 44 ; attention, le modèle doit être hiérarchiquement bien formulé).
7. Parmi les consommateurs d'alcool, on décide de créer des méta-niveaux de consommation avec A : niveaux 1 à 5 ; B : niveaux 6 à 10 ; C : niveaux 11 et +. Peut-on mesurer le surcroît de risque lors d'un passage d'un méta niveau à l'autre ? (**TUTO 1**, page 43 ; **TUTO 2**, section 5.2.4) (Attention au recodage de la variable ALCOOL. Conclusion : boire fait ronfler, boire plus n'augmente pas les risques de ronflement).
8. On veut analyser maintenant le rôle de AGE (numérique) et ALCOOL (numérique).
    a. Réalisez la régression RONFLE ~ AGE + ALCOOL, quelles sont les variables significatives au sens de Wald ? Comment interprétez-vous les coefficients ?
    b. Testez la significativité simultanée des coefficients des deux variables au sens du test du rapport de vraisemblance (puisqu'il n'y a que ces variables, cela revient à tester globalement la régression, et elle est significative à 10%).
    c. Introduisez le terme d'interaction AGE * ALCOOL, est-il significatif ? (non)