Année 2017-2018 **M1 MINT**

Régression avancé
TD/TP 1 : Modèle logistique

Avant de commencer,

1. récupérer sur ma page le fichier le fichier `R_TDTP1.Rmd` et les données "prostate".

Les données concernent 53 patients atteints du cancer de la prostate. Pour ces patients, on veut prédire l'atteinte du système lymphatique (`lymph`, codée 1 si le système lymphatique est atteint, 0 sinon). Cette atteinte conditionne le traitement du cancer et sa mesure directe nécessite une intervention chirurgicale. On essaie donc de voir s'il est possible de la prédire à partir de données facilement mesurables. Ces données sont extraites de Collet (1991), voir aussi wikistat. Les variables explicatives considérées sont les suivantes :

- `age` âge du patient,
- `acid` niveau de "serum acid phosphatase",
- `radio` résultat d'une analyse radiographique (0 : négatif, 1 : positif),
- `taille` taille de la tumeur (0 : petite, 1 : grande),
- `gravite` résultat de la biopsie (0 : moins sérieux, 1 : sérieux).

Exercice 1

1. (cette question est corrigée dans le fichier `R_TDTP5.R`) Charger les données et vérifier le nombre d'observations et de variables ainsi que leur type. Faire les changements de type nécessaires.
2. Représenter graphiquement les liens entres la variable `lymph` et les variables `age`,`taille` et `acid`.
3. Faire un histogramme pour la variable `acid`. A partir du graphique précédent et de l'histogramme, proposer une transformation de cette variable.
4. Pour chaque variable explicative discrète, faire une table croisée de la variable et de `lymph` (`table` puis un test du χ^2 (`chisq.test`).

Exercice 2

1. Faire un premier modèle logistique, en prenant en compte toutes les variables explicatives potentielles.
2. Calculer les prédictions obtenues avec le modèle sur les individus de la table, puis calculer une matrice de confusion. Quels sont les taux de faux positifs et faux négatifs ?
3. Calculer l'AUC du modèle.
4. Faire le test de nullité simultanée des coefficients des variables explicatives.
5. Faire une recherche d'individus aberrants et isolés. Enlevez vous des individus de l'étude ?
6. A partir des p-values des tests de Wald, enlever la variable avec la plus grande p-value, recommencer jusqu'à ce que toutes les p-values soient inférieures à 0.10. Quel modèle obtenez-vous ?

Exercice 3

Recommencer l'analyse en considérant les interactions d'ordre 2 entre 2 variables explicatives discrètes et une discrète une continue. Puis comparer le modèle obtenu à celui de la question précédente.

Attention pour la dernière question concernant la sélection de variables :

- ne supprimer un effet principal qu'à la condition qu'il n'intervienne plus dans des interactions,
- ne supprimer qu'un terme à la fois.