

BE2 - Regression logistique - A RENDRE

MOD - Statistique appliquée aux sciences de l'ingénieur.

Exercice : Régression logistique (11h30 - 12h15, dépôt sur Moodle)

On considère le fichier "Tuyns.txt" dont les variables sont les suivantes :

- TAB : en lien avec le nombre de cigarettes consommées par jour $nbcig$.

$$\begin{cases} 1 & \text{si } nbcig \text{ est inférieur à 7} \\ 2 & \text{si } 8 \leq nbcig \leq 15 \\ 3 & \text{si } 16 \leq nbcig \leq 25 \\ 4 & \text{sinon} \end{cases} \quad (1)$$

- ALC : en lien avec la quantité d'alcool (g) consommée par jour $malc$

$$\begin{cases} 1 & \text{si } malc \text{ est inférieur à 40} \\ 2 & \text{si } 41 \leq malc \leq 80 \\ 3 & \text{si } 81 \leq malc \leq 120 \\ 4 & \text{sinon} \end{cases} \quad (2)$$

- cancer : 1 si présence d'un cancer du larynx ou de l'hypopharynx, 0 sinon.
1. On considère le modèle de régression logistique en l'absence de covariables (i.e. avec intercept uniquement) $\log\left(\frac{p_0}{1-p_0}\right) = \alpha$.
 - A quoi correspond le coefficient α ?
 - Comment s'interprète-t-il ?
 - Mettre en oeuvre et interpréter le résultat (signe du coefficient et p_valeur).
 2. Etudier l'effet de la variable TAB sur la survenue du cancer.
 - Donner l'équation du modèle et interpréter l'intercept dans ce cas.
 - Mettre en oeuvre le modèle sur R.
 - Interpréter le résultat de la routine *Anova*.
 - Interpréter le résultat du *summary*, on regardera si tous les niveaux d'exposition ont une influence significative.
 3. Recommencer les étapes précédentes en étudiant l'effet de la variable ALC sur la survenue du cancer. Quelle est la variable la plus influente ALC ou TAB ?
 4. Proposer un modèle complet avec interaction. Interpréter les résultats de ce modèle. Simplifier éventuellement le modèle et calculer l'Odds pour une population ALC = 1 TAB = 2. Donner la matrice de confusion de ce modèle.