TP - REGRESSION LOGISTIQUE

Exercice 1 : Modèle Logistique Simple

On a relevé l'âge et la présence(1) ou l'absence (0) d'une maladie cardiovasculaire chez 100 individus. Les données sont stockées dans le fichier **"MCV.txt"**: sur une ligne donnée, la variable **AGE** fournit l'âge d'un individu tandis que la variable **CHD** prend la valeur 1 en cas de présence d'une maladie cardiovasculaire chez cet individu et la valeur 0 sinon. Les variables **ID** et **AGRP** donnent respectivement le numéro d'un individu et sa classe d'âge.

1. Charger et afficher les données regroupées en classe d'âge.
2. On souhaite étudier la relation entre CHD et la variable explicative **AGE**. Afficher les données à l'aide d'un nuage de points. Commenter la Figure.
3. Calculons la proportion de malades observée selon les classes d'âge définies par la variable **AGRP**. Définir un vecteur **centre** qui donne les centres de chaque classe puis représenter le nuage de points de **p** versus **centre**. Y a-t-il une liaison entre **CHD** et **AGE** ? Quelle est la forme de ce graphique ? Quel est son intérêt comparativement au graphique précédant ? Quel modèle suggérez-vous d'utiliser ?
4. Commençons pour ajuster une régression logistique de **CHD** en fonction de **AGE**. Commenter les résultats (tests de significativité, nombre de degrés de liberté).
5. Afin de mieux discerner les relations entre les différentes classes, il est demandé de représenter sur un même graphique les proportions selon la classe d'âge et la courbe logistique ajustée.
6. Ajuster de même le modèle **"probit"** puis comparer les deux modèles. Commenter les résultats.
7. Estimer, dans chacun des deux modèles la cote d'un individu âgés de 30 ans. Commenter. Estimer le rapport de cotes correspondant à la variable **AGE**.

Exercice 2 : Modèle Logistique Multiple

Nous traitons un problème de défaut bancaire. Nous cherchons à déterminer quels clients seront en défaut sur leur dette de carte de crédit (ici **defaut** = 1 si le client fait défaut sur sa dette). La variable **defaut** est la variable réponse. La base de données **Defaut** est accessible à partir du package **ISLR** que vous devez installer au préalable.

La base **Defaut** dispose d'un échantillon de tailles 10000 et 3 variables explicatives. Les variables explicatives sont les suivantes :

- **student:** variable à 2 niveaux {0,1} (**student** = 1 si le client est un étudiant).
- **balance:** montant moyen mensuel d'utilisation de la carte de crédit.
- **income:** revenu du client.

1. Charger et afficher les données.
2. Commenter les données en utilisant la fonction **summary**.
3. Afin de faciliter le traitement, vous devez transformation de la variable **default** à 0 si Non et 1 si Yes.

4. Construire un modèle de régression logistique avec la variable balance comme variable explicative qualitative. Commenter.
5. Une fois que les coefficients ont été estimés, il est simple de calculer la probabilité de défaut étant donné balance (solde moyen de carte de crédit donné). En utilisant les estimations des coefficients indiqués dans le tableau précédant, prédire la probabilité de **defaut** pour un client qui a une balance de **1000**, **1500**, **2000** et **3000** dollars respectivement. Commenter.
6. Donner le tableau de contingence des variables **default** et **student** puis estimer (avec un calcule à la main) les coefficients du modèle logistique.
7. Estimer P(default = Yes|student = Yes) et P(default = Yes|student = Non).
8. Construire un modèle de régression logistique multiple avec les 2 variables explicatives **student** et **balance**.
9. Construire un modèle de régression logistique multiple avec les 2 variables explicatives **student** et **balance**.
10. Construire un modèle de régression logistique multiple avec les 3 variables explicatives **student** et **balance** et **income**.

Exercice 3 : Modèle linéaire généralisé

Une régression logistique est habituellement utilisée lorsqu'il y a une variable de résultat dichotomique (telle que gagner ou perdre) et une variable prédictive continue qui est liée à la probabilité ou à la probabilité de la variable de résultat. Il peut également être utilisé avec des prédicteurs catégoriques et avec de prédicteurs multiples.

Supposons que nous partons d'une partie du jeu de données **" mtcars"** intégré dans R. Les données ont été extraites du magazine 1974 de Motor Trend US et comprennent la consommation de carburant et 10 aspects de la conception et de la performance automobile pour 32 automobiles (modèles 1973-74). Nous utiliserons **"vs"** comme variable de résultat, **"mpg"** comme prédicteur continu, et **"am"** comme prédicteur catégorique (dichotomique ou binaire).

1. Charger et afficher les données.
2. Créer un modèle logistique où vous considéré "**mpg"** est la variable prédictive continue et **"vs"** est la variable de résultat qualitative binaire (dichotomique).
3. Interpréter les résultats.
4. Tracer avec la fonction **plot** le graphe des données et du modèle régression logistique.
5. Refaire la même chose avec la fonction **ggplot2**.
6. Refaire les questions précédentes mais avec cette fois ci "**am"** comme variable prédictive continue et **"vs"** comme variable de résultat qualitative binaire (dichotomique).
7. Construire le modèle de régression avec "**mpg"** comme variable prédictive continue, "**am"** comme variable prédictive dichotomique et **"vs"** comme variable de résultat qualitative binaire (dichotomique).
8. Comparer les résultats avec le modèle "**probit"**.
9. Considérer maintenant que le modèle est linéaire. Refaire les mêmes questions et Conclure.

Mini-Projet à Rendre : Filtrage de Spams

George Forman, chercheur chez Hewlett-Packard, a créé un ensemble de données sur le spam. Il a pris 4601 de ses propres messages électroniques, les a étiquetés et a extrait diverses fonctionnalités. Les caractéristiques comprennent la fréquence de divers mots (par exemple, " money"), des caractères spéciaux (par exemple des signes de dollar) et l'utilisation de majuscules dans le message.

Les données spam peuvent être téléchargées à partir de : ftp://ftp.ics.uci.edu/pub/machine-learning-databases/spambase/

Il est demandé de :

1. Créer un modèle de régression logistique à partir des données spams.
2. Décrire et commenter le modèle.
3. Valider le classifieur avec un taux d'erreur acceptable.
4. Le valider par une technique de validation connue (validation croisée par exemple).
5. Documenter toutes les étapes.

Bonne Chance