Travaux Pratiques Analyse Statistique et R 01

Régressions Linéaires/Logistiques sur R

Alison PATOU
patou.alison@gmail.com
12/03/2020

Introduction

Dans ce TP nous allons mettre en pratique les régressions (linéaires et logistiques)

Exercice 1

Info: Nous travaillons sur le dataset BostonHousing qui contient 506 lignes et 14 colonnes. L'enjeu va être de réaliser une régression pour prédire le prix de vente des biens (colonne medv).

Question 1

- Charger le jeu de données grâce au package ***mlbench***
  Déterminer la structure du dataset et décrire les variables présentes.

Question 2

- Explorer et visualiser la distribution de la variable à expliquer.

Question 3

- Explorer les différentes variables explicatives et démontrer la présence de potentielles correlations entre la variable à expliquer et les variables explicatives.

Question 4

- A partir des variables explicatives significatives, essayez différentes combinaisons demodèles linéaires pour créer un modèle performant en utilisant la fonction **lm()**. (3 combinaisons)

Question 5

- Afficher l'AIC, le BIC et le R^2 pour chacun des modèles puis comparer les entre eux afin de déterminer le plus performant

Question 6

- Nous allons créer un échantillon de **test** et un echantillon **train** en prenant une extraction de notre dataset actuel. Pour cela utiliser la fonction **createDataPartition**() du package ***caret***. (25/75)

Question 7

- A l'aide des modèles créés, faire la prédiction sur l'échantillon test et tester la performance de prédiction des modèles. Comparer les erreurs de prédiction grâce à un tableau de contingence par exemple.

Question 8

- Créer d'autres modèles avec différents algorithmes glm. Comparer les performances.

Exercice 2

Info: Nous travaillons sur le dataset Ozone qui contient 1041 lignes et 20 colonnes. L'enjeu va être de réaliser une régression. La variable à prédire est **O3obs**.

Question 1

- La recherche d'une meilleure méthode de prévision suit le protocole suivant.
    - o Étape descriptive préliminaire uni et multidimensionnelle visant à repérer les incohérences,les variables non significatives ou de distribution exotique, les individus non concernés ou atypiques... et à étudier les structures des données. Ce peut être aussi la longue étape de construction de variables, attributs ou features spécifiques des données.
    - o Procéder à un tirage aléatoire d'un échantillon test qui ne sera utilisé que lors de la dernière étape de comparaison des méthodes.
    - o Créer l'échantillon d'apprentissage pour l'estimation des paramètres des modèles.
    - o Comparaison des qualités de prévision à l'aide de l'échantillon de test qui est resté à l'écart.

Aidez vous des questions précédentes pour mener à bien cette prévision.

Exercice 3

Info : Les données à étudier sont stockées dans le fichier **maladies.txt**

On a plusieurs colonnes : l'âge de l' individu
L'ID de l'individu
Sa classe d'âge AGRP
La présence d'une maladie chronique CHD

Question 1

- Importer le dataset et regarder la répartition des variables

Question 2

- Nous allons faire un peu de transformation de données.
  Convertir la valeur 1 en Vrai/0 en Faux de la colonne CHD grâce à la fonction **factor()**
  Idem pour la variable AGRP que nous allons seulement convertir en factor

Question 3

- Quel est le pourcentage de personne atteinte de maladie chronique dans ce dataset?

Question 4

- Visualiser dans un graphique l'âge en fonction de la présence de la maladie (CHD)

Question 5

- Effectuer une régression logistique ordinaire sur la variable à expliquer CHD.

Question 6

- Afficher la matrice de confusion (de contigence) pour vérifier la pertinence de votre modèle.