Devoir maison 3

Aude Sportisse

24/09/2020

Ce devoir maison est à envoyer par mail aude.sportisse@upmc.fr pour le lundi 28 septembre (soir) en format .Rmd et .html.

Il comporte deux exercices, on pourra utiliser les fiches 5 et 6.

Exercice 1

A partir du jeu de données icecream, nous allons étudier la consommation de glace aux Etats-Unis sur une période de 30 semaines du 18 Mars 1950 to 11 Juillet 1953. Les variables sont la consommation (Consumption en pintes par habitant), le salaire hebdomadaire (Income en dollars), le prix des glaces (Price en dollars), la température (Temp en degré fahrenheit) et la catégorie socio-professionnelle (sc).

1. Charger le jeu de données **icecream** avec le nom des colonnes en utilisant la fonction **read.table** et en specifiant correctement les arguments sep, row.names et header. Précisez la nature des variables (qualitative ou quantitative) et faites une analyse rapide des données.
2. Créer un jeu de données comprenant seulement les variables quantitatives. Découper aléatoirement ce jeu de données en deux échantillons: un échantillon d'apprentissage (avec 70% des données) et un échantillon de test.
3. Sur le jeu de données d'apprentissage, représenter les nuages de points de **cons** en fonction de(s) variable(s) quantitative(s). Effectuer ensuite la régression linéaire de la variable **cons** en fonction de toute(s) le(s) variable(s) quantitative(s) du données. Ce modèle sera appelé **modele1**. Afficher un résumé et interpréter le résultat (donner notamment les variables significatives et en quel sens elles le sont).
4. Effectuer une procédure de sélection de variable par le critère d'information bayésien (BIC). Quelles sont les variables sélectionnées ?
5. Effectuer une nouvelle régression linéaire avec uniquement les variables retenues en question précédente, ce modèle sera appelé **modele2**. Refaire ensuite la procédure de sélection de variables.
6. Dans un vecteur, stocker les valeurs de **cons** prédites par le modèle **modele1** pour chaque individu de l'échantillon de test. Construire de même un vecteur à partir de **modele2**. Utiliser pour cela la fonction **predict**.
7. Notons Y la variable **cons**. Pour $i = 1, ..., N_t$ où N_t est le nombre d'observation de l'échantillon de test, on note $\hat{Y}_i^j$ la prévision par le modèle j du i-ème individu de l'échantillon test, et Y_i la valeur de Y observée sur le i-ème individu de l'échantillon test. Calculer

$$\text{EQM1} = \frac{1}{N_t}\sum_{i=1}^{N_t}(\hat{Y}_i^1 - Y_i)^2 \text{ et } \text{EQM2} = \frac{1}{N_t}\sum_{i=1}^{N_t}(\hat{Y}_i^2 - Y_i)^2.$$

Interpréter.

9. Représenter **cons** en fonction de(s) variable(s) qualitative(s). Effectuer une analyse de la variance en rappelant à quoi cela sert de votre compréhension. Conclure sur l'effet de(s) variable(s) qualitative(s) sur la consommation de glace.

Exercice 2

Nous considérons le modèle linéaire

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

où ε est la variable d'erreur. Rappelons que les paramètres β_0 et β_1 sont inconnus et que nous les estimons par $\hat{\beta}_0$ et $\hat{\beta}_1$. Nous notons dans la suite $\beta = \begin{pmatrix}\beta_0 & \beta_1\end{pmatrix}'$ et $\hat{\beta} = \begin{pmatrix}\hat{\beta}_0 & \hat{\beta}_1\end{pmatrix}'$. Dans cet exercice, nous voulons retrouver expérimentalement le résultat vu en cours $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$.

1. Simuler 1000 échantillons $(x_i, Y_i)_{1 \leq i \leq 100}$ suivant le modèle

$$Y_i = 1 + 4x_i + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, 1),$$

   avec les $x_i \sim \mathcal{N}(3, 1)$.

2. Pour les 1000 échantillons simulés, stocker les estimateurs des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$ renvoyés par la fonction *lm* dans une matrice de taille 1000×2. Stocker également les moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$ renvoyés par la fonction *lm* dans une autre matrice de taille 1000×2 en les calculant à la main avec la formule $\hat{\beta} = (X'X)^{-1}X'Y$. Vérifier que les valeurs sont bien les mêmes.

3. Rappeler la moyenne et l'écart type théorique de $\hat{\beta}_0$. Calculer l'écart-type. Comparer les valeurs théoriques avec la valeurs observées.

4. Superposer la densité théorique de $\hat{\beta}_0$ et un estimateur de la densité obtenu à partir de l'échantillon obtenu précédemment.

5. Reprendre la question 1 en augmentant le bruit avec $\epsilon_i \sim \mathcal{N}(0, 10)$. Comme dans la question 2, pour les 1000 échantillons simulés, stocker les estimateurs des moindres carrés $\hat{\beta}_0$ et $\hat{\beta}_1$ renvoyés par la fonction *lm* dans une matrice de taille 1000×2. Que remarque-t-on sur les valeurs estimées $\hat{\beta}$? Expliquer pourquoi (avec un graphique idéalement).