Ecole Centrale de Lyon
Statistique appliquée aux sciences de l'ingénieur

BE1- Prise en main de R - Régression linéaire

C. Helbert

Exercice 1 (*Prise en main de R)* On s'intéresse à l'évolution de la température moyenne de surface de la planète en fonction des émissions de CO2.

1. Ouvrir le Notebook "climat.rmd" avec RStudio puis lancer chaque cellule de code les unes après les autres. Bien repérer les spécificités du code R (orienté objet).

   On souhaite maintenant compléter ce fichier avec de nouveaux modèles. On donnera les réponses aux questions directement dans le Notebook en insérant éventuellement des commandes Latex.

2. Réaliser la régression linéaire du CO2 en fonction de year.
   a Donner l'équation du modèle
   b Quelle est la part de variance expliquée par le modèle ?
   c Peut-on valider ce modèle (étude des résidus) ?
   d Proposer un autre modèle meilleur que le précédent (on cherche une transformation des variables ou une complexification du modèle). Justifier

3. On cherche à prédire la température en 2050
   a Prédire le niveau de C02 moyen attendu en 2050.
   b En déduire la prédiction de l'anomalie de température en 2050. Les 2°C de hausse seront-ils atteints ?
   c Intégrer les incertitudes de prédictions dans les 2 modèles pour donner une prévision optimiste et une prévision pessimiste de la hausse de température en 2050.

Exercice 2 On va utiliser les données contenues dans le fichier Chenilles. Ce jeu de données est issu de Tomassone et al. Il s'agit d'une étude réalisée en 1973 sur un parasite du pin bien connu : la chenille processionnaire. On désire connaître l'influence de certaines variables sur la densité de peuplement du parasite. La variable à expliquer est Y : c'est le nombre moyen de nids par arbre sur la parcelle considérée de 10 hectares. On dispose ainsi au total des résultats concernant 32 parcelles distinctes. Les différentes variables susceptibles d'avoir une influence sur Y (ou encore sur $log(Y)$, sont relatives aux différentes caractéristiques de la placette (subdivision de chaque parcelle) et sont :

— l'altitude en mètres : X1
— la pente en degrés : X2
— le nombre de pins dans la placette : X3
— la hauteur (en m) de l'arbre échantillonné au centre de la placette : X4

— le diamètre de cet arbre : X5
— la note de densité de peuplement : X6
— l'orientation de la placette (de 1=sud, à 2=autre) : X7
— la hauteur en mètres des arbres dominants : X8
— le nombre de strates de végétation : X9
— le mélange du peuplement (de 1=mélangé à 2=non mélangé) : X10

Il s'agit de données quantitatives même pour X7 et X9, la valeur dans le fichier étant en fait une moyenne sur un certain nombre de placettes échantillonnées dans chaque parcelle. L'objectif de l'étude est d'essayer de déterminer le meilleur modèle en sélectionnant les variables influentes.

1. Créer un fichier de données à partir du fichier chenilles.txt.
2. Visualiser le fichier de données.
3. Ecrire l'équation du modèle de la régression de Y en fonction des variables $(X1, .., X10)$. Combien de paramètres a-t-on à estimer ?
4. Calculer à l'aide de la commande **lm** l'équation de la régression linéaire de la variable Y en fonction des variables $(X1, .., X10)$. (On donnera un nom à cette régression.)
5. Analyser le test de non-régression (Préciser l'hypothèse, la statistique et le résultat). Conclusion.
6. Le modèle ainsi estimé convient-il ?
7. On souhaite maintenant sélectionner les variables influentes en utilisant 3 méthodes. Commencer par faire cette sélection en utilisant l'algorithme backward sur le test de Student (Expliquer le principe). Quel est le modèle obtenu ?
8. Refaire la même chose pour les critères AIC et BIC. Quels sont les modèles obtenus ?
9. Quel est pour vous le meilleur modèle ? Pourquoi ?

Exercice 3 Soient $X_1, X_2, .., X_n$ n variables aléatoires indépendantes suivant une loi binomiale $\mathcal{B}(m, p)$. Trouver l'estimateur $\hat{p}$ du maximum de vraisemblance du paramètre p. Calculer son biais et sa variance.

Exercice 4 Soient a, b deux réels tels que $0 < a < b$ et $(X_1, X_2, .., X_n)$ un n-échantillon de loi uniforme sur $[0, b]$. Soient les variables aléatoires $(Y_j)_{j=1,..,n}$ définies par

$$Y_j = sgn(X_j - a)$$

où $sgn(x) = 1$ si $x \geq 0$ et -1 sinon.

1. Calculer $E[Y_j]$ et $Var(Y_j)$.
2. Soit $\hat{a} = k \sum_{j=1}^{n} Y_j$. Existe-t-il $k \in \mathbf{R}$ tel que $\hat{a}$ soit un estimateur sans biais de a.
3. Soit $P(\epsilon_1, .., \epsilon_n; a) = P[Y_1 = \epsilon_1, ..., Y_n = \epsilon_n]$. Ecrire cette probabilité en fonction de $\frac{1}{2}\sum_{j=1}^{n}(1 + \epsilon_j)$ et de $\frac{1}{2}\sum_{j=1}^{n}(1 - \epsilon_j)$. En déduire l'estimateur $\bar{a}$ du maximum de vraisemblance de a. L'estimateur $\bar{a}$ est-il sans biais ? Calculer sa variance. Donner la limite de la variance quant n tend vers $+\infty$.