Devoir 2

Aloui Mohamed, Jafuno Douba

3 décembre 2018

Question 1

Pour commmencer nous découpons notre jeu de données en un jeu d'apprentissage **bodyap** composé de **168** hommes et un jeu de données test **bodytest** composé de **84** hommes.Nous considérons donc le **modèle 1** linéaire suivant :

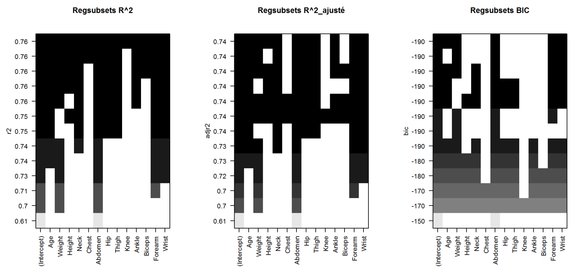
$$\text{bodyfat} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Weight} + \beta_3 \text{Height} + \beta_4 \text{Neck} + \beta_5 \text{Abdomen} + \beta_6 \text{Hip} + \beta_7 \text{Thigh} + \beta_8 \text{Knee}$$

$$\beta_9 \text{Anckle} + \beta_{10} \text{Biceps} + \beta_{11} \text{Forearms} + \beta_{12} \text{Wrist} + \varepsilon$$

Question 2

Le jeu d'apprentissage comporte beaucoup de variables explicatives. Dans le package leaps, la fonction regsubsets retourne, pour différents critères (bic, R^2, R^2_a (ajusté), Cp de Mallows, etc.), le meilleur modèle, l'analyse de ces graphiques nous permettra de choisir les variables à conserver dans le **modèle 1**.

Voici les 3 Regbusets



Pour le premier graphique on cherche donc le modèle qui maximise le coefficient de détermination R^2. On retrouve le fait que le coefficient R^2 augmente avec le nombre de variables et que pour une comparaison entre modèles emboîtés, On sélectionne toute les variables il n'est pas judicieux de considérer ce critère.

Pour le second graphique le but est de maximiser le coefficient R^2_a. Ici, c'est le modèle contenant la constante, les variables, **Age**, **Weight**, **Height**, **Neck**, **Abdomen**, **Hip**, **Tigh**, **Anckle**, **Forearm** et **Wrist**

Pour le troisième graphique on cherche à minimiser le critère BIC. On choisit donc le modèle contenant la constante, les variables **Age**, **Weight**, **Abdomen**, **Forearm** et **Wrist**.

Au vue des graphiques ci dessus la constante et les variables **Age**, **Weight**, **Abdomen**, **Forearm** et **Wrist**, nous semblent etre les plus pertinentes à l'étude, nous les choisissons donc pour la suite de l'étude.

Question 3

Etant donné un échanillion (**bodyfat**,**Age**, **Weight**, **Abdomen**, **Forearm**, **Wrist**) Le but de cette étude de notre jeu d'apprentissage cherche précisément à expliquer la variable **bodyfat** à partir des variables explicatives **Age**, **Weight**, **Abdomen**, **Forearm** et **Wrist**

Nous considérons donc le nouveau **modèle 2** de régrésion linéaire suivant:

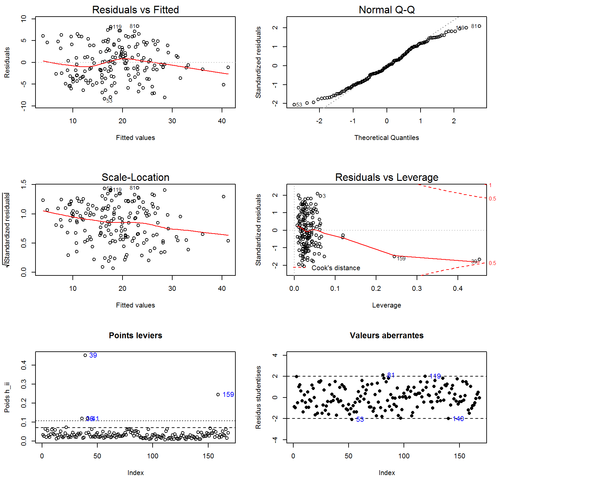
$$\text{bodyfat} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Weight} + \beta_3 \text{Abdomen} + \beta_4 \text{Forearm} + \beta_5 \text{Wrist} + \varepsilon$$

Il s'agit d'un modèle de régression multiple où ε est l'erreur du modèle qui exprime l'information manquante dans l'explication linéaire des valeurs de **bodyfat** à partir des variables explicatives (**Age**, **Weight**, **Abdomen**, **Forearm** et **Wrist**). Les coefficients β_i sont les paramètres que nous chercherons à estimer.

Question 4

Validation du modèle

Afin de pouvoir valider le modèle on effectue une analyse des résidus considérons les 4 graphiques ci dessous créé avec la fonction plot(lm)



Le premier graphique Residuals vs Fitted illustre la dispersion des résidus en fonction des valeurs prédites par le modèle de régression linéaire. Chaque point représente la distance entre la variable réponse et la réponse prédite par le modèle. Ici les résidus forment une bande horizontale approximative autour de la ligne de 0, **la variance des résidus est homogène (donc, ils sont homoscédastiques)**, le modèle est **validé**.

Le second graphique (Diagramme quantile-quantile (QQplot) compare la distribution de probabilité des résidus du modèle à une distribution de probabilité de données normales. On voit que la plus part es résidus standardisés se trouvent près de la première bissectrice on pourrait eventuellement retiré des valeurs aberrantes ou des points leviers si il y en a. **Les résidus peuvent être considérés comme normalement distribués**.

Dans le troisième graphique "Scale-location avec celui des valeurs aberantes" permet de vérifier si la dispersion des résidus augmente pour une valeur prédite donnée (i.e. si la dispersion des résidus est causée par la variable explicative). **Si la dispersion augmente, la condition de base d'homoscédasticité n'est pas respecté ici elle n'augmente pas elle diminue** de plus ce graphique permet de comparer la racine des résidus studentisés à $\sqrt{(2)} = 1.4$: il y a 4 observations au dessus de ce seuil, donc **4 valeurs aberrantes (53, 81, 119 et 140)**.

Dans les derniers graphiques Diagramme de résidus vs. influence: Distance de Cook et Points leviers Si une ou certaines observations sont aberrantes (dont, si elles ont des valeurs très différentes des autres), le modèle peut être mal ajusté en raison de leur influence exagérée sur la calculation du modèle. Si (et seulement si) ces observations correspondent à des erreurs de mesure ou à des exceptions (à la fois point levier et valeurs abérantes), elles peuvent être retirées du jeu de donnée, observons les.

- Pour les points leviers (4ème et 5ème graphique) :
  - pour le seuil $2p/n = 0.07$: on observe une valeur au-dessus (poids hii)
  - pour le seuil $3p/n = 0.10$: on voit 4 valeurs au dessus du seuil préoccupant

  on en déduit donc qu'il y a **4 points leviers (39,36,41 et 159)** dans le jeu de donnees
- Pour la distance de Cook : le seuil du cours $F_{p,n-p}(0.1) = 0.49$ est proche de 0.5 indiqué sur le 4ème graphique. Il n'y a pas de point qui dépasse la bande de 0.5, donc en terme de distance de Cook, **il n'y a pas d'observations suspectes en effet on a aucun point qui est à la fois un point levier et une valeur aberrante.**

Tout cela montre que **notre modèle est bien valide**

Question 5

Masses graisseuses prédites

	Masses graisseuses prédites pour les 10 premiers hommes du jeu de test
1er	35.61436
2ème	20.98934
3eme	12.51530
4ème	10.07434
5ème	18.25368
6ème	17.95180
7ème	16.56668
8ème	12.68862
9ème	15.67927
10ème	27.78182

Erreurs de prédiction

	Erreurs de prédiction des 10 premiers hommes du jeu de test
1er	-1.314360
2ème	-4.489340
3eme	-9.515303
4ème	-9.374345
5ème	2.246320
6ème	-1.051798
7ème	8.733325
8ème	-2.788617
9ème	-2.579270
10ème	2.118181

	Erreur quadratique moyenne de prévison
Modele 1	25.83622
Modele 2	25.02863

Le second modèle, moins lié aux données d'apprentissage, a une meilleure capacité de prédiction de la masse graisseuse que le premier modèle car son erreur quadratique moyenne de prévison est inférieure à celle du premier.