Code

Utilisation d'un modèle linéaire gaussien pour la prédiction de masse graisseuse chez l'homme

Florent VALBON, Julien MASSIP (TP B1)

23 novembre 2018

L'objectif est d'élaborer un modèle de régression linéaire afin d'expliquer le pourcentage de masse graisseuse dans le corps masculin en fonction de plusieurs facteurs. Nous avons pour cela sélectionnés 168 données sur les 252 dont nous disposons pour établir notre modèle (données d'apprentissage). Nous testerons l'efficacité de ce dernier sur le reste des valeurs (échantillon test).

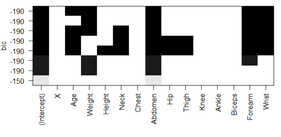
Nous disposons des données suivantes (il y en a 252 au total) :

	X <int>	bodyfat <dbl>	Age <int>	Weight <dbl>	Height <dbl>	Neck <dbl>	Chest <dbl>	Abdomen <dbl>	Hip <dbl>
1	1	12.3	23	69.87525	172.085	36.2	93.1	85.2	94.5
2	2	6.1	22	78.48225	183.515	38.5	93.6	83.0	98.7
3	3	25.3	22	69.76200	168.275	34.0	95.8	87.9	99.2
4	4	10.4	26	83.69175	183.515	37.4	101.8	86.4	101.2
5	5	28.7	24	83.46525	180.975	34.4	97.3	100.0	101.9
6	6	20.9	24	95.24325	189.865	39.0	104.5	94.4	107.8

6 rows | 1-10 of 15 columns

1 Choix du modèle

Nous allons sélectionner les variables les plus pertinentes. Le critère BIC pour tout les sous modèles possibles est donné par le graphique suivant :



On cherche à minimiser le critère BIC. On choisit donc le modèle contenant la constante, les variables Age, Weight (Poids), Abdomen, Wrist (Tour du poignet), Forearm (Tour de l'avant bras). Il semblerait que ces facteurs soient les plus pertinants pour expliquer le taux de masse graisseuse dans le corps masculin.

Nous choisirons donc le modèle linéaire avec ces variables :

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

Une régression linéaire nous permet d'estimer les différents estimateurs :

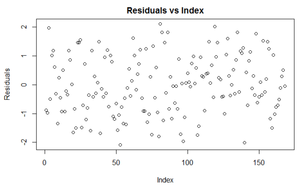
```
Call:
lm(formula = bodyfat ~ Age + Weight + Abdomen + Forearm + Wrist,
    data = bf.app)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3616 -3.1885 -0.2047  3.3955  8.3531

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -29.12038    8.07717  -3.606 0.000461 **
Age           0.07391    0.03271   2.260 0.025167 *
Weight       -0.28419    0.07581  -3.749 0.000247 ***
Abdomen       0.92796    0.07797  11.901  < 2e-16 ***
Forearm       0.89380    0.24285   3.680 0.000317 ***
Wrist        -2.36391    0.59965  -3.942 0.000120 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.104 on 162 degrees of freedom
Multiple R-squared:  0.7353,    Adjusted R-squared:  0.7272
F-statistic: 90.02 on 5 and 162 DF,  p-value: < 2.2e-16
```

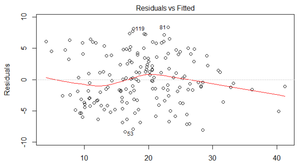
Les p-values très petites dans le test de Fisher confirment que toutes les variables sélectionnées sont significatives.

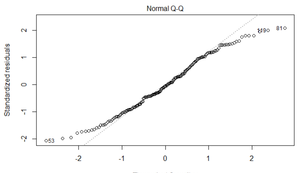
2 Validation

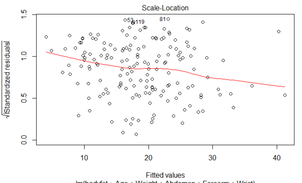
Nous allons vérifier si notre modèle est valide et le modifier le cas échéant.

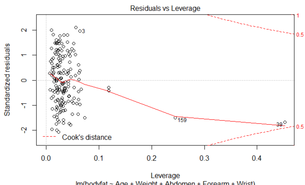
Il s'agit de vérifier les hypothèses du modèle linéaire gaussien : normalité des résidus, homogénéité des résidus et indépendance des résidus.











Pour l'homogénéité : les résidus ne présentent pas de structures particulières (premier graphique). On vérifie donc l'hypothèse d'homoscédasticité.

Pour la normalité : le deuxième graphique est un QQ-plot qui montre que la plupart des points s'alignent bien sur la première bissectrice. Nous pouvons donc accepter l'hypothèse de normalité.

Pour les valeurs aberrantes : le troisième graphique permet de comparer la racine des résidus studentisés à $(\sqrt{2}) = 1.4$: il y a 3 observations au dessus de ce seuil, donc 3 valeurs aberrantes.

Pour les points leviers (4ème graphique) :

Pour le seuil $2p/n = 0.06$: on observe une dizaine de valeurs au-dessus. pour le seuil $3p/n = 0.09$: on observe 4 valeurs au-dessus du seuil préoccupant. On en déduit donc qu'il y a 4 points leviers dans le jeu de données. Pour la distance de Cook : Il n'y a pas de point qui dépasse la bande de 0.5 (indiqué sur le quatrième graphique), donc en terme de distance de Cook, il n'y a pas d'observations suspectes.

Nous allons donc enlever les trois valeurs aberrantes du modèle : La $53^{ième}$, la $119^{ième}$, la $81^{ième}$. Les nouveaux estimateurs sont :

```
Call:
lm(formula = bodyfat ~ Age + Weight + Abdomen + Forearm + Wrist,
    data = bf.app)

Residuals:
    Min      1Q  Median      3Q     Max
-7.9546 -3.1352 -0.0753  3.1969  7.6204

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.80849    8.89316  -2.902 0.004234 **
Age           0.07411    0.03175   2.334 0.020843 *
Weight       -0.27480    0.07425  -3.701 0.000295 ***
Abdomen       0.92025    0.07617  12.082  < 2e-16 ***
Forearm       0.91455    0.23614   3.873 0.000157 ***
Wrist        -2.58485    0.58866  -4.391 2.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.979 on 159 degrees of freedom
Multiple R-squared:  0.7478,    Adjusted R-squared:  0.7399
F-statistic:  94.3 on 5 and 159 DF,  p-value: < 2.2e-16
```

3 Prédiction

Avant d'éxaminer le potentiel de notre modèle pour la prédiction, une première lecture de l'échantillon nous révèle qu'un homme a été mesuré comme possédant 0% de matière grasse. Cette valeur était hautement suspecte, car même le coeur possède un peu de graisse, cette obervation sera mis de côté.

Le modèle appliqué sur l'échantillion test des 83 données restantes permet de prédire les taux de masses graisseuses :

```
     169      170      171      172      173      174
35.51644 21.11674 12.50808 10.15325 18.31497 17.98072
```

On calcul le RMSE2 :

```
[1] 0.2824429
```

Puis le RMSE3 :

```
[1] 0.02482694
```

Et enfin la MAPE

```
[1] 0.2194296
```

4 Conclusion

Toute nos erreurs sont relativement faible, pour un modèle qui n'utilise que 6 variable explicative. Nous pouvons donc conclure que notre modèle permet, avec un coût de calcul faible, de prédire, avec une précision moindre cependant, le pourcentage de masse graisseuse d'un homme.