Régression linéaire multiple

BEASSE Joseph, **CHIMBAULT** Thomas, **GONÇALVES** Tristan

ENSC - École Nationale Supérieure de Cognitique 06 janvier 2023

jbeasse@ensc.fr | tchimbault@ensc.fr | tgoncalve002@ensc.fr

I. Introduction

Afin de mesurer finement le pourcentage de matière grasse corporelle, certaines balances utilisent des méthodes spécifiques, telles que la bio-impédance qui permet ce calcul en transmettant des signaux électriques à travers le corps.

Il serait intéressant de trouver un modèle permettant de prédire ce pourcentage en utilisant de simples outils tels qu'une balance ou un mètre-ruban. Pour ce faire, nous avons étudié un jeu de données constitué de 13 variables à étudier et d'une 14ème témoignant du pourcentage de masse graisseuse, ceci sur 250 individus.

Notre objectif est donc de soumettre un modèle pertinent pouvant inférer sur la masse graisse à l'aide de certaines des 13 variables.

II. Description du jeu de données (première observation)

Avant de se lancer dans une étude profonde de nos données, il est utile d'effectuer une première visualisation de ces dernières afin de détecter une potentielle répartition anormale ou des outliers (valeurs aberrantes).

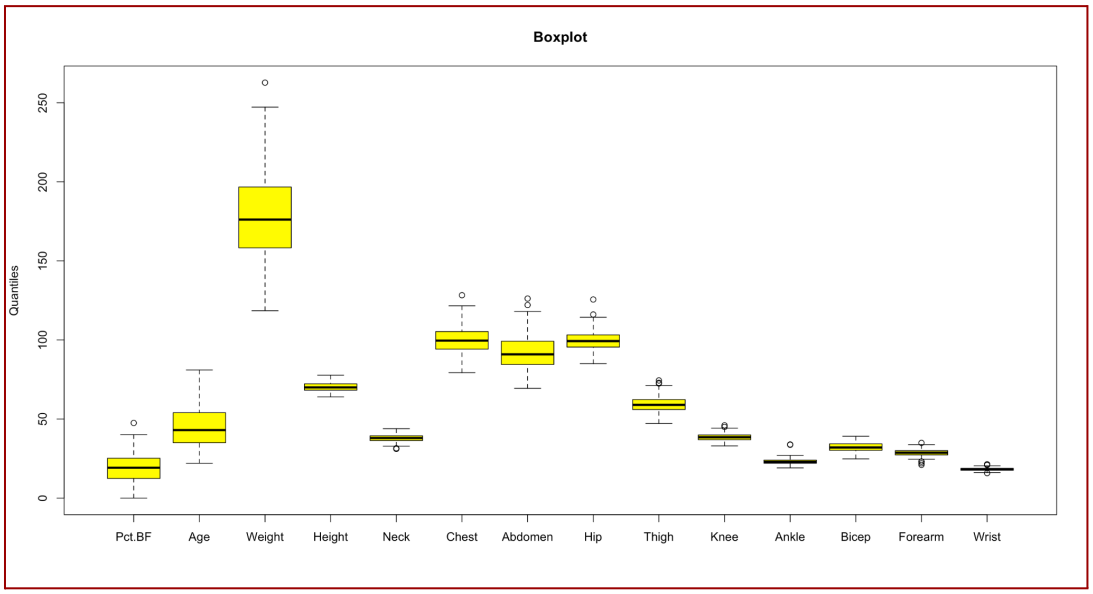
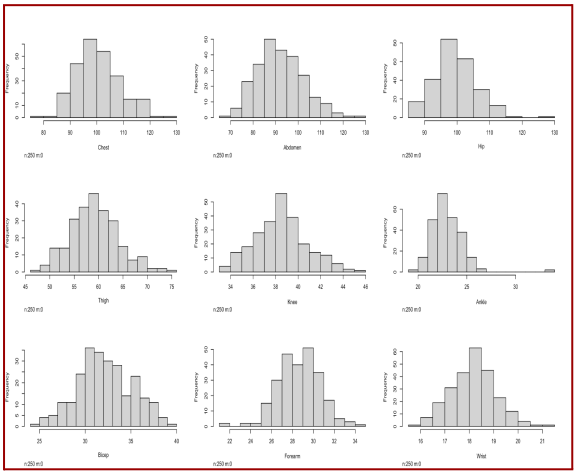


Figure 1 : Boxplot des différentes variables

L'observation de ces boxplots permet de mettre en évidence la présence de valeurs aberrantes. Certaines de ces valeurs proviennent des mêmes individus. Celles-ci peuvent être source de biais dans la génération des modèles formulés ultérieurement et méritent une attention particulière. Ces individus sont les suivants: **40** - **43** - **214** - **224**.



Dans la littérature, il est mentionné que les caractéristiques humaines comme la taille et le poids sont réparties normalement autour de sa moyenne. Observons nos échantillons sur les histogrammes ci-contre. Les variables observées sont discrètes, d'après leur forme on peut supposer qu'avec plus d'observations les données tendraient vers une répartition dite ''gaussiennes''. Ceci nous rassure sur la provenance de ces dernières. En effet, elles semblent bien provenir d'un groupe d'humains sélectionnés aléatoirement.

III. Analyse de la corrélation entre les variables

Pour réaliser l'étude en composante principale, on a choisi d'utiliser le package PCAmixdata. Dans cette première approche on représente les coordonnées des individus dans l'espace des composantes principales afin d'observer d'éventuelles valeurs aberrantes.

Sur la figure 3, on conclut que la 5e composante est décrite par 2 observations qui posent un problème. En effet, elles sont en dehors du nuage de points dans chacun des espaces.

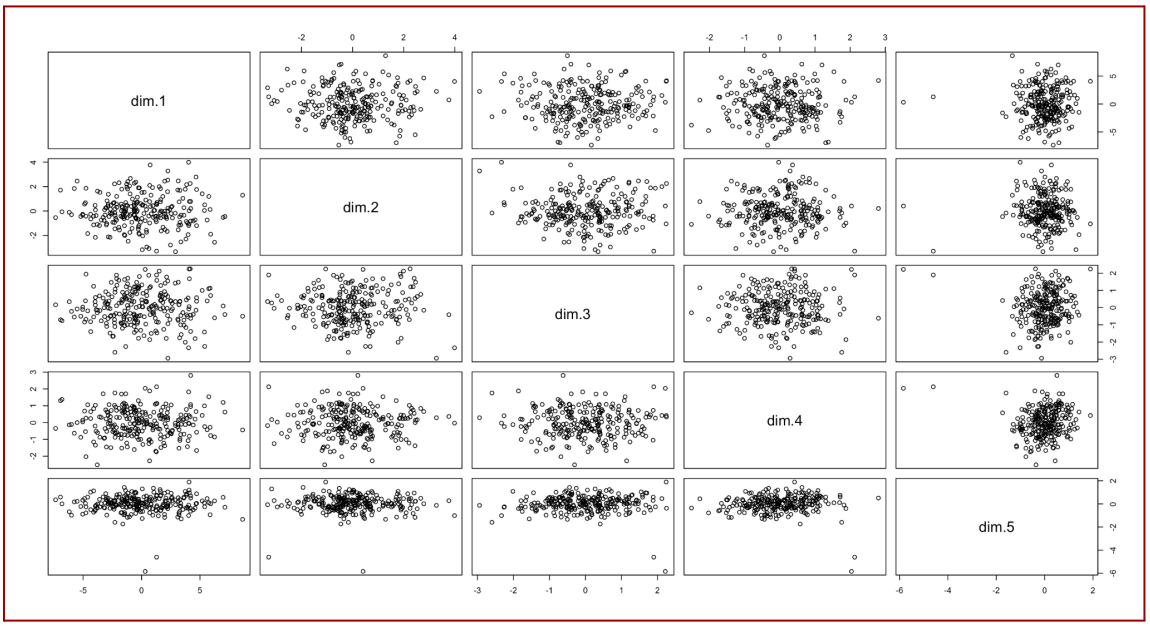
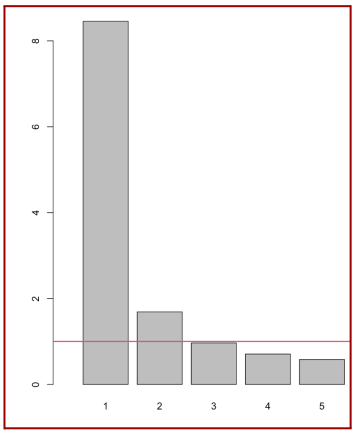


Figure 2 : Nuage de points des coordonnées des individus sur les composantes principales

Il est alors intéressant de se demander si ces valeurs ont un impact sur le modèle. Pour cela, on a regardé l'importance de la variable PCT.bf sur la 5ème dimension. On obtient la valeur suivante: **cos² = 6.81 e^{-04}**. Cette dernière est très faible, l'importance de ces outliers sur la variable de masse graisseuse est peu significative. Tout de même, les individus **31** et **84** aberrants ne sont pas à oublier, on pourrait les prendre en considération si un problème survenait.



En appliquant le critère de Kaiser, on choisit de ne garder que les deux premières dimensions. Pour rappel, le critère de Kaiser déclare que l'on ne doit choisir que les valeurs propres supérieures à 1, ou les valeurs propres dont la proportion cumulative est supérieure à 70%.
En choisissant ces dimensions, on peut expliquer **72.48**% de l'information.

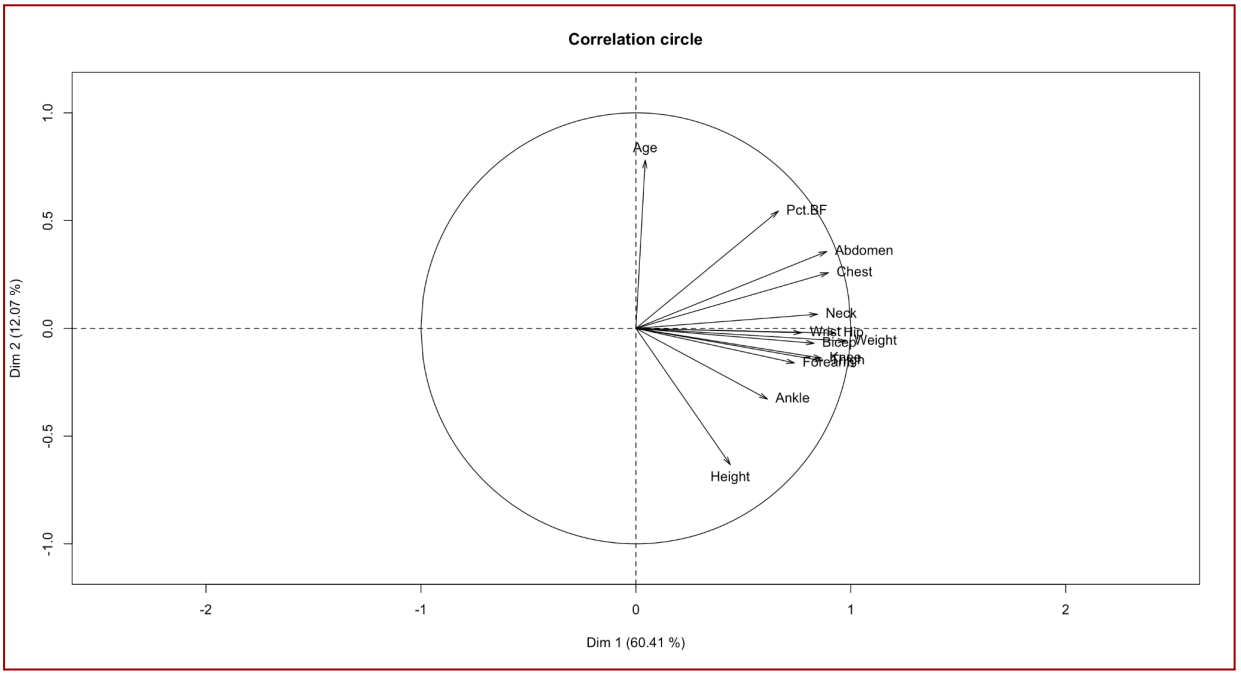


Figure 3 : Cercle des corrélations des variables dans la dimension 1 et 2

D'après le cercle des corrélations, l'âge a l'air de bien décrire la dimension 2, mais ne semble pas décrire la dimension 1. En effet, l'âge décrit à 31,5% la dimension 3 contre 0.002% pour la dimension 1.
Les deux dimensions ont l'air décrites par la taille, ainsi que par Pct.BF (pourcentage de masse graisseuse).

Pour les autres variables en revanche, elles semblent toutes plutôt bien représentées par la dimension 1, mais également par une autre dimension non visible sur ce graphique, notamment pour Wrist et Forearm, du fait de la taille réduite du vecteur.

On peut voir sur ce cercle que le pourcentage de matière grasse est totalement indépendante de la taille. De plus, l'âge et le tour de cheville ont l'air d'être peu corrélées au pourcentage de masse grasse.

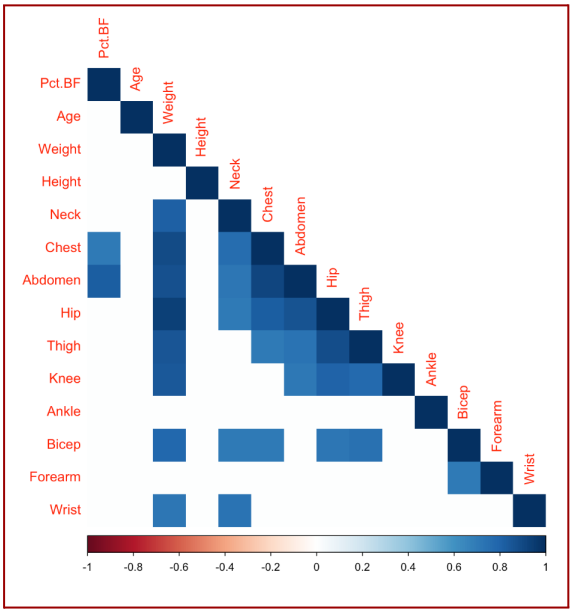


Figure 4 : Tableau de corrélation présentant les corrélations supérieures au seuil 70%

Afin de commencer à discuter de l'importance de certaines variables dans le modèle de prédiction recherché, il est important de considérer la corrélation des variables avec la variable à expliquer. En figure 4, est représenté un tableau de corrélation prenant en compte seulement les coefficients supérieurs à 0.7 (seuil de forte corrélation entre deux variables). Ici, il n'est pas intéressant de s'intéresser aux coefficients proches de -1 puisque d'après la figure 3, aucune variable n'est corrélée négativement. Le tableau de corrélation montre donc que les variables *Chest* et *Abdomen* ont un impact significatif sur la variable *Pct.BF*. Il semble donc sensé de dire que ces deux variables pourraient être utiles dans le modèle de régression. À l'inverse, d'autres variables ne semblent pas avoir d'influence significative sur la variable à expliquer et pourraient ne pas être utiles dans le modèle de régression linéaire : *Age, Weight, Height, Neck, Hip*…
Cette analyse ne représente pas une fin en soit dans la détermination d'un modèle pertinent. Elle permet seulement d'avoir une vision globale et hypothétique sur le jeu de données. Il est donc indispensable de réaliser une étude précise pour déterminer un modèle efficace et précis.

IV. Modèle de régression linéaire multiple

Dans le but de déterminer un modèle de régression linéaire cohérent, la première stratégie a été de prendre un modèle intégrant l'ensemble des 13 variables. Ce premier essai n'est pas mauvais : on obtient un Adjusted R-squared de **0.7368**. Ceci signifie que 73,68% de la variabilité des données est expliquée par notre modèle. Même si ce dernier résultat n'est pas insatisfaisant, il est important de chercher à améliorer le modèle.

On réalise alors une étude plus fine du modèle afin d'obtenir une corrélation linéaire multiple supérieure à celle du modèle disposant de toutes les variables.

Pour cela, on utilise le critère d'information d'Akaike selon 2 approches: une ascendante et une descendante. La première méthode consiste à partir du modèle sans aucune variable et d'y ajouter respectivement toutes les variables dont l'AIC est le plus faible jusqu'à ce que l'intercept ait cette valeur. On peut effectuer ce traitement à l'aide des fonctions R **step** (**ascendant**) ou **add1**. On conserve alors les variables suivantes: *Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Bicep, Forearm* et *Wrist*. L' Adjusted R-squared de ce modèle s'élève à **0.736.** Après avoir vérifié avec l'autre approche (**drop1** ou **step descendant**), on obtient un modèle et un coefficient différent.

En effet, le coefficient est de **0.7385**, une valeur supérieur au modèle précédent et le modèle est décrit par les variables suivantes que l'on va sélectionner pour la suite de l'étude: *Age*, *Height*, *Neck*, *Abdomen*, *Hip*, *Thigh*, *Forearm* et *Wrist*.

Après avoir vérifié, les résidus ne présentent pas de structure particulière et le test de Shapiro-Wilk nous indique qu'ils sont distribués normalement (p-value : **0.13**).

Ces résultats ne venant pas invalider notre modèle, on souhaite alors étudier l'écart entre la valeur par le modèle de régression et sa valeur réelle pour une observation donnée.

On réalise donc une étude des résidus standardisés à l'aide de la fonction R **rstudent():**

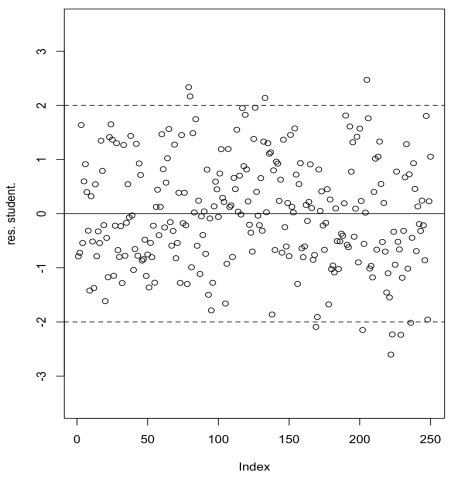


Figure 5 : Résidus de student

Pour que les hypothèses du modèle soient valides, on récupère les valeurs ''outliers'' qui sont au-delà de 2 ou en deçà de -2 sur la figure 5. On définit un nouveau modèle en enlevant ces observations dites aberrantes. Adjusted R-squared est améliorée, on obtient: **0.7745**. Cependant le test de Shapiro indique que les résidus ne sont pas distribués normalement (p-value: 0.002063). On ne peut donc pas valider ce modèle, et on revient au modèle précédent décrivant mieux notre variable PCT.bf.

V. Conclusion et perspectives

En somme, il a été décidé de garder le modèle expliqué par les variables suivantes : *Age, Height, Neck, Abdomen, Hip, Thigh, Forearm* et *Wrist*, dont l'équation est présentée plus bas. Grâce à celui-ci, il est possible d'expliquer **73,85%** de la variable *Pct.BF*. Ce résultat n'est pas mauvais compte tenu des conditions expérimentales. En effet, plusieurs facteurs viennent influencer les résultats : le peu de données présentes dans le jeu en question, la variance inter-individus… La variabilité inter-individuelle est un aspect qu'il est impossible de supprimer dans ce type d'expérience. Même si chaque donnée humaine est régie selon une Gaussienne, chaque individu possède ses propres caractéristiques qui rendent impossibles la détermination d'une règle absolue et générale.

Pour répondre à la question posée, les variables présentes dans le modèle gardé peuvent être utilisées pour prédire le pourcentage de graisse corporelle d'un individu à un certain stade. En effet, pour obtenir une estimation rapide du pourcentage à l'aide d'outils élémentaires, une méthode de prédiction par ce modèle peut être intéressante. En revanche, pour une prédiction précise, cela ne semble pas être la méthode la plus adaptée face à la technique de pesée hydrostatique.

A l'avenir, en ayant effectué les ajustements cités précédemment, il pourrait donc être intéressant d'appliquer le modèle amélioré sur de plus grands jeux de données.

Équation:

$$PCT.bf = 5.04 + 0.07*Age\ (années) - 0.27*Height\ (pouces) - 0.45*Neck\ (cm) + 0.8*Abdomen\ (cm) - 0.19*Hip\ (cm) + 0.22*Thigh\ (cm) + 0.3*Forearm\ (cm) - 1.73*Wrist\ (cm)$$

Annexe

Répartition des tâches

La répartition des tâches dans le groupe s'est faite sans accroc, nous nous connaissons bien et sommes capables de se répartir les tâches de façon à respecter une équité globale dans le projet.

Tâche	Joseph	Thomas	Tristan
Analyse de données via R			
Stats. Descriptives	25%	25%	50%
Analyse Composantes P.	30%	40%	30%
Modèle de régression M.	45%	35%	20%
Rapport			
Rédaction	34%	33%	33%

Code R

```r
#=============================================
# Installation des librairies nécessaires
#=============================================

install.packages("PCAmixdata")
install.packages("FactoMineR")
install.packages("factoextra")
install.packages("Hmisc")
library(corrplot)
library(PCAmixdata)
library(FactoMineR)
library(factoextra)
library(Hmisc)

#=============================================
# Chargement des données
#=============================================

# Pour charger le dataset donneesProjet:
# double clic sur le fichier donneesProjet2A.RData dans le dossier sous-jacent

# Première visualisation du jeu de données
#-----------------------------------
```

```r
plot(donneesProjet)

# Statistique descriptive
#-----------------------------------
head(donneesProjet)
summary(donneesProjet) # visualisation stat. de base
boxplot(donneesProjet,col = c("yellow"),main = paste("Boxplot"), ylab = "Quantiles")

hist.data.frame(donneesProjet[,2:4])
hist.data.frame(donneesProjet[,6:14])


ACP <- PCA(data.frame(donneesProjet), graph=FALSE)
round(ACP$eig,digit=2)
n = names(donneesProjet)

# Outliers
#-----------------------------------

for (i in (1:length(donneesProjet))){
  boxplot(donneesProjet[,i],xlab=n[i])
  print(n[i])
  val = min(max(donneesProjet[,i]),quantile(donneesProjet[,i],0.75)+
1.5*(quantile(donneesProjet[,i],0.75)-quantile(donneesProjet[,i],0.25)))
  val2 = max(min(donneesProjet[,i]),quantile(donneesProjet[,i],0.25)-
1.5*(quantile(donneesProjet[,i],0.75)-quantile(donneesProjet[,i],0.25)))
  print(which (donneesProjet[,i]> val))
  print(which (donneesProjet[,i]<val2))

}
# On retient l'outlier 40 - 43 - 214 - 224 car ils reviennent plusieurs fois
plot(donneesProjet$Ankle,donneesProjet$Height)
which(donneesProjet$Ankle>32)

# Analyse en Composantes Principales
#-----------------------------------

res<-PCAmix(donneesProjet)
val3 = data.frame(res$quanti$contrib.pct)
val4 = data.frame(res$ind$coord)
plot(val4)

# La 5e composante a un problème, on a 2 observations qui posent un problème
# Impactent-elles le modèle ?
# Regardons l'importance de la variable PCT.bf sur la 5e dimension
# cos2 := 6.81 e-04
# C'est très faible, donc peu d'importance de ces outliers, mais ne pas les oublier
# Ce sont les individus 31 et 84

val5 = data.frame(res$quanti$cos2)
round(res$eig,digits=2)
barplot(res$eig[,1],main="Eigenvalues",names.arg=1:nrow(res$eig))
abline(h=1,col=2,lwd=2)

plot(res,axes=c(1,2),choice="ind") # on retrouve ici le graphique des individus (plan 1-2)
```

```r
plot(res,axes=c(1,2),choice="cor") # on retrouve ici le cercle des corrélations
plot(res,axes=c(1,2),choice="sqload") # on retrouve ici le graphique des "square loadings"
(plan 1-2)
res$quanti$cos2

# Calcul de la matrice de corrélation
matCor = cor(donneesProjet)
# forte corrélation => coeff. > 0.7
matCor2 = matCor
for(i in 1:nrow(matCor)){
  for(j in 1:ncol(matCor)){
    if(matCor2[i,j]<0.7){
      matCor2[i,j]=0
    }
  }
}
corrplot(matCor,is.corr=TRUE, method="shade", type="lower")
corrplot(matCor2,is.corr=TRUE, method="shade", type="lower") # Visualisation des données
corrélées exclusivement

# On définie une étude linéaire multiple pour expliquer le modèle de Pct.BF en fonctions
des données de donnesProjet
resPct <- lm(Pct.BF~.,data=donneesProjet)
summary(resPct)
# Adjusted R-squared:  0.7368

step(resPct)

plot(resPct$fitted,resPct$residuals)
plot(resPct$fitted,donneesProjet$Pct.BF)
abline(a=0,b=1)
shapiro.test(resPct$residuals) # p value 0.09 => résidus normalement distribués

# Etude plus fine du modèle
#----------------------------------
# On enlève les variables selon le critère d'AIC - méthode drop1 ou step descendant
dataTest = donneesProjet[,c(1,2,4,5,7,8,9,13,14)]
resTest = lm(Pct.BF~.,data=dataTest)
summary(resTest)
# Adjusted R-squared:  0.7385

# resTest2 = lm(Pct.BF~Hip+Forearm+Thigh+Neck+Height+Age+Wrist+Abdomen,data=donneesProjet)
# summary(resTest2)
# Ces 2 lignes servent à montrer que l'on obtient bien le même résultat

# Avec la méthode add1 ou step ascendant
resTest3 = lm(Pct.BF~Abdomen+Weight+Wrist+Bicep+Age+Thigh,data=donneesProjet)
add1(resTest3,~Age+Weight+Height+Neck+Chest+Abdomen+Hip+Thigh+Knee+Ankle+Bicep+Forearm+Wri
st)
summary(resTest3)
# Adjusted R-squared:  0.736

# --> on obtient un meilleur Rsquared pour resTest <--
# On garde resTest pour la suite de l'étude
```

```r
plot(resTest$fitted,resTest$residuals) # pas de structure dans les résidus
abline(h=0)
plot(resTest$fitted,dataTest$Pct.BF)
abline(a=0,b=1)
shapiro.test(resTest$residuals)
# p value 0.13 => résidus distribués normalement

# Résidus de student pour trouver de potentiels valeurs aberrantes
residus.stud<-rstudent(resTest)
plot(residus.stud,ylab="res. student.",ylim=c(-3.5,3.5))
abline(h=c(-2,0,2),lty=c(2,1,2))
which (residus.stud>2)
which(residus.stud<(-2))

# Outliers
# On enlève les observations suivantes: 79 80 133 205 169 202 222 223 229 236
dataRemoved = dataTest[c(-79,-80,-133,-205,-169,-202,-222,-223,-229,-236),]
resRemoved = lm(Pct.BF~.,data=dataRemoved)
summary(resRemoved)

# Adjusted R-squared:  0.7745
residus.stud2<-rstudent(resRemoved)
plot(residus.stud2,ylab="res. student.",ylim=c(-3.5,3.5))
abline(h=c(-2,0,2),lty=c(2,1,2))
which(residus.stud2>2)
shapiro.test(resRemoved$residuals)
# p-value = 0.002063 --> Non distribués normalement
# Revenir au modèle précédent <------------------

##############################################################################
# Modèle choisi pour prédire le pourcentage de masse graisseuse:
dataTest = donneesProjet[,c(1,2,4,5,7,8,9,13,14)]
resTest = lm(Pct.BF~.,data=dataTest)
# Ainsi, pour expliquer au mieux cette indice on utilise les variables:
# Age, Height, Neck, Abdomen, Hip, Thigh, Forearm, Wrist
# Avec ces dernières on obtient 73,85 % d'explication de la masse graisseuse
##############################################################################
```