Fitting Weight to Simple Body Measurements

Target Variable

L'ensemble des données comprend le pourcentage de graisse corporelle, l'âge, la taille ainsi que dix mesures de circonférence du corps telles que l'abdomen ou le coude pour 252 hommes.

Le but de ce projet est d'ajuster le poids en fonction des différentes mesures en utilisant une multiple regression. Cela permet de fournir un moyen pratique afin d'estimer le poids chez les hommes.

Data

D'après le fichier de description, les densités corporelles des cas 48, 76 et 96 semblent comporter un chiffre erroné. Or, nous avons 252 observations pour 19 variables, donc nous pouvons nous permettre de les supprimer sans compromettre l'étude.

```
tab = tab[ -c(48, 76, 96),]
```

De plus, il y a une erreur dans les données, en effet, la personne associée au cas 42 mesure 69.5 *inches*, et non pas 29.5 *inches*, nous avons donc rectifié cette erreur.

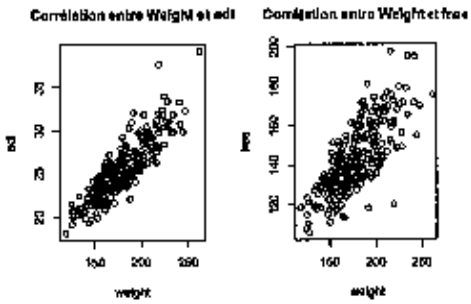
```
tab[42,]$height = 69.5
```

Ensuite, les estimations du pourcentage de graisse corporelle sont tronquées à zéro lorsqu'elles sont négatives (cas 182). Nous avons donc supprimé ce cas car il s'agit d'une valeur abérante.

```
tab = tab[ -c(182),]
```

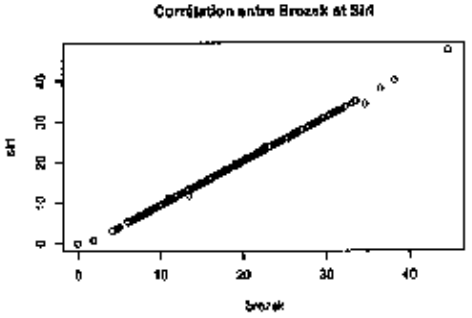
Il est nécessaire de supprimer la première colonne car elle correspond aux *indices*.

De plus, nous avons la formule suivante, $Adiposity\ index = Weight/Height^2\ (kg/m^2)$ donc cette variable est fortement liée à $Weight$. De même pour $Fat\ Free\ Weight = (1 - fraction\ of\ body\ fat) * Weight$ using Brozek's formula (lbs). Graphiquement, nous pouvons observer qu'il ne s'agit pas de fonctions linéaires. Cependant, on pourra supprimer ces variables par la suite, si le modèle de régression se base uniquement sur ces variables.



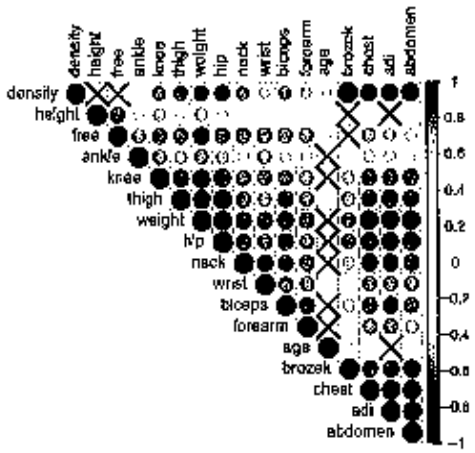
On suppose donc qu'il est possible de calculer directement l'$Adiposity$ et $Fat\ Free\ Weight$ sans utiliser les formules associées et donc le $Weight$.

Enfin, la *co – variate Percent body fat using Brozek's equation* correspond à $457/Density - 414.2$. De plus, *Percent body fat using Siri'sequation* correspond à $495/Density - 450$. Graphiquement, la regression entre les deux variables correspond à une fonction linéaire donc il s'agit de données redondantes, nous allons supprimer *Percent body fat using Brozek's equation*.




```
tab = tab[, -c(1, 3)]
```

Etude des corrélations entre les variables



Graphiquement, nous pouvons observer que la *target variable weight* n'est pas corrélée avec *age*. Cependant, cela n'est pas suffisant pour l'enlever du modèle. Il est nécessaire d'utiliser une sélection de variable afin de conserver uniquement les variables les plus significatives.

Finalement, les *co variates* du modèle sont *brozek*, *density*, *age*, *height*, *neck*, *chest*, *abdomen*, *hip*, *thigh*, *knee*, *ankle*, *biceps*, *forearm*, *adiposity*, *fat free weight* et *wrist*.

Contexte :

Étude de la relation entre le pourcentage de graisse corporelle, l'âge, le poids, la taille et des mesures de la circonférence du corps d'un homme.

Objectif :

Modéliser le poids d'un homme.

Base de données :

- Nombre de variables : **19**
- Nombre d'observations : **252**
- Source : *Dr. A. Garth Fisher, Human Performance Research Center, Brigham Young University, Provo, Utah 84 602*
- Variables (Tableau 1) :

Nom	Description	Unité	Type	Minimum	Maximum	Moyenne empirique
case	ID de l'homme étudié		Int	1,00	252,00	126,50
Brozek	Pourcentage de graisse corporelle selon l'équation de Brozek : $= \frac{457}{Density} - 414{,}12$		Numeric	0,00	45,10	18,94
Siri	Pourcentage de graisse corporelle selon l'équation de Siri : $= \frac{495}{Density} - 450{,}00$		Numeric	0,00	47,50	19,15
Density	Densité	$gm.cm^{-3}$	Numeric	0,995	1,109	1,056
Age	Âge	année	Numeric	22,00	81,00	44,88
Weight	Poids	livre (lbs)	Numeric	118,50	363,10	178,90
Height	Taille	pouce (inches)	Numeric	29,50	77,75	70,15
Adiposity	Indice d'adiposité : $= \frac{Weight}{Height^2}$	$kg.m^{-2}$	Numeric	18,10	48,90	25,44
Fat	Poids sans gras : $(1 - Brozek) \times Weight$	livre (lbs)	Numeric	105,90	240,50	143,70
Neck	Circonférence du cou	cm	Numeric	31,10	51,20	37,99
Chest	Circonférence de la poitrine	cm	Numeric	79,30	136,20	100,82
Abdomen	Circonférence de l'abdomen à l'ombilic et au niveau de la crête iliaque	cm	Numeric	69,40	148,10	92,56
Hip	Circonférence de la hanche	cm	Numeric	85,00	147,70	99,90
Thigh	Circonférence de la cuisse	cm	Numeric	47,20	87,30	59,41
Knee	Circonférence du genou	cm	Numeric	33,00	49,10	38,59
Ankle	Circonférence de la cheville	cm	Numeric	19,10	33,90	23,10
Extended	Circonférence du biceps étendu	cm	Numeric	24,80	45,00	32,27
Forearm	Circonférence de l'avant-bras	cm	Numeric	21,00	34,90	28,66
Wrist	Circonférence du poignet en aval des processus styloïdes	cm	Numeric	15,80	21,40	18,23

Tableau 1: Description des variables composant la base de données.

Nettoyage des données :

- Suppression de la variable non mesurée : ***case*** (ID)
- Suppression des variables ***linéairement dépendantes*** et ***fortement corrélées*** car calculées à partir d'autres variables (Figure 1 et Tableau 1) :

Brozek *(Density)*
Siri *(Density)*
Adiposity *(Weight, Height)*
Fat *(Brozek, Weight)*

} Coefficients de corrélation proche de *-1* ou *1*

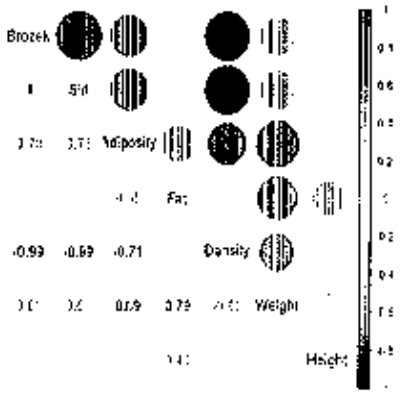


Figure 1 : Fortes corrélations et dépendances linéaires des variables Brozek, Density, Siri, Adiposity, Fat, Weight et Height.

- Modification des valeurs ***Density*** erronées des observations **[48]**, **[76]** et **[96]** à partir de l'équation de ***Brozek***:

$$Density = \frac{457}{Brozek + 414{,}2}$$

- Modification de la valeur ***Height*** erronée de l'observation **[42]** : un homme avec un poids ***(Weight)*** de 205 livres mesure 69,5 pouces plutôt que 29,5 pouces.
- Suppression de l'observation **[182]** : les valeurs de ***Brozek*** et ***Siri*** ont été tronquées à 0 car elles étaient négatives.
- Suppression de ***17*** observations ayant des ***valeurs aberrantes*** (Figures 2 et 3) vérifiant pour une ou plusieurs variables :

$$x_i < Q_1 - 1{,}5 \times IQR$$
ou
$$x_i > Q_3 + 1{,}5 \times IQR$$

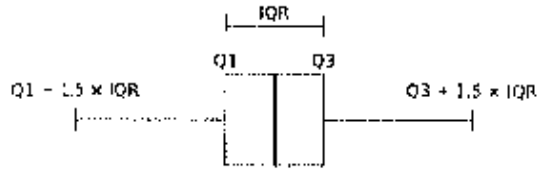


Figure 3: Boîte à moustaches indiquant Q1, Q3 et IQR.

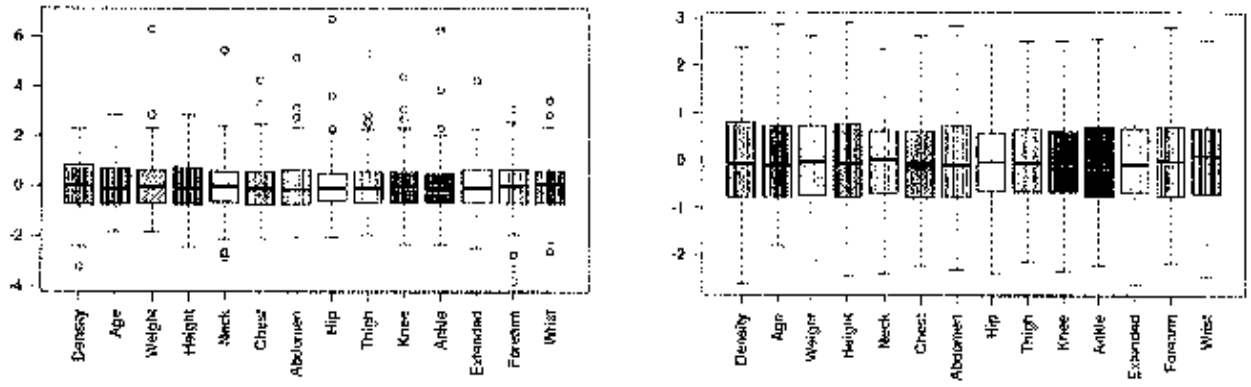


Figure 2: Boîtes à moustaches des données mises à l'échelle avant (à gauche) et après (à droite) nettoyage, supprimant les valeurs aberrantes.

Base de données de travail (après nettoyage) :

- Nombre de variables : **14**
- Nombre d'observations : **234**
- Modélisation (Tableau 2) :

Variable cible	**Variables explicatives**
Weight	*Density, Age, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Extended, Forearm, Wrist*

Tableau 2: Modèle basé sur la base de données de travail.

1 Original Dataset

The dataset we are going to study is supplied by Dr. A. Garth Fisher (Human Performance Research Center, Brigham Young University) who gave permission to use it for non-commercial purposes. We are dealing with data of medical nature which have initially been used for fitting body fat values to the other measurements using multiple regression, in order to provide a predictive equation for the determination of body fat.
This dataset consists of 252 observations and 19 variables. The investigated subjects are men aged from 22 to 81 years old and with percent body fat from 3.7 to 40.1. The variables concern the ages of the subjects and some relevant body measurements, as we can see from the brief descriptions in Table 1.
Since we have training examples and we look for a model capable of map new observations as correctly as possible, this is a supervised learning problem. This is also a case of multiple regression, indeed there are one target variable and several explanatory variables. The aim of our project is to consider the "weight" as target variable, trying to predict it referring to the other covariables.

Table 1: Description of all the variables in the original dataset, provided by Dr. A. Garth Fisher.

Variable	Description
case	Case Number
Brozek	Percent body fat using Brozek's equation, 457/Density - 414.2
Siri	Percent body fat using Siri's equation, 495/Density - 450
density	Density (gm/cm^3)
age	Age (yrs)
weight	Weight (lbs)
heigth	Height (inches)
adp	Adiposity index = Weight/Height^2 (kg/m^2)
ffw	Fat Free Weight = (1 - fraction of body fat) * Weight, using Brozek's formula (lbs)
neck	Neck circumference (cm)
chest	Chest circumference (cm)
abd	Abdomen circumference (cm) "at the umbilicus and level with the iliac crest"
hip	Hip circumference (cm)
thigh	Thigh circumference (cm)
knee	Knee circumference (cm)
ankle	Ankle circumference (cm)
biceps	Extended biceps circumference (cm)
forearm	Forearm circumference (cm)
wrist	Wrist circumference (cm) "distal to the styloid processes"

2 Adjustments

Before doing any analysis, the dataset has been adjusted and cleaned up.
First of all, we decide to delete the variable in the first column of the dataset which simply refers to the numbers of the patients, since it is useless for our purpose and, obviously, it has no explanatory power. In addition, looking at the article which firstly presented this dataset, we notice that there are few errors in

some measures. For example, body densities for cases 48, 76, and 96 seem to have one digit in error, as it can be seen computing them using the two body fat percentage values. There is also the presence of a man (case 42) over 200 pounds in weight and who is less than 3 feet tall. The height should presumably be 69.5 inches instead of 29.5 inches, and as a consequence we have adjusted this value. Moreover, the paper signals the presence of a person (case 182) whose percent body fat estimates were originally negative and have been truncated to zero. We decide not to take into account this individual in our dataset and therefore we remove this observation, since it is not a realistic measure and it could compromise our model estimation and future analysis.
Finally, looking at the same article, we become aware that the density appears not to be easily measurable, due to the existance of different techniques and to measurement errors that could happen. In this particular case, the adopetd "underwater" density measurement technique can not be totally reliable and as a result we decide to consider the possibility of excluding this variable from our dataset successively.

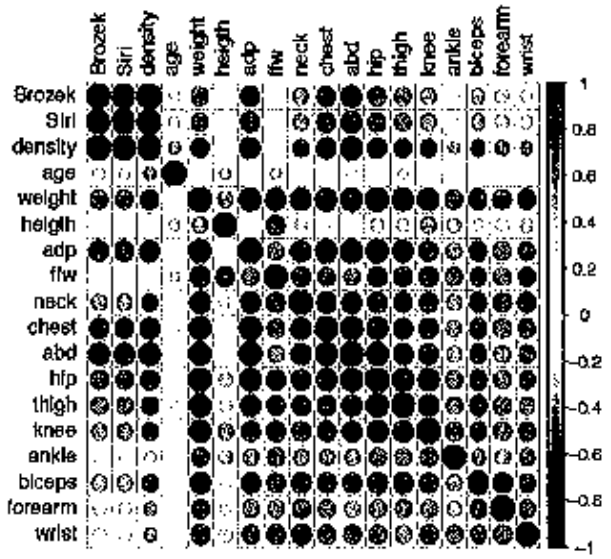


Figure 1: Correlation Plot

3 Correlation analysis

As previously mentioned, the target variable we want to model in this project is the "weight" of each man. Plotting the correlation matrix (Figure 1) we can analyse the relationships among the "weight" variable and the other explanatory variables. At a first sight we can notice that our target variable is positively correlated with all the other covariables of the dataset, except for the "density", which shows a medium size negative correlation (-0.60544129), and except for the "age", which shows a coefficient around zero (-0.01081929). Therefore, we can conclude that the men's age seems to have no linear influence on the weight of each patient. Looking at the other explanatory variables, each one exhibits a strong linear relation with the target variable, showing correlation coefficients always higher than 0.60. The only exception is the "height" variable, which has a slightly lower value, equal to +0.47674133. Always referring to the "weight" target variable, the most significant relationship which emerges from this correlation plot is the one concerning the "hip circumference (cm)". Indeed, the corresponding coefficient is equal to +0.94088412, which is really near to the maximum attainable value of 1.
In addition to this, we could even notice the presence of some relevant collinearities among the covariables of the dataset, which can generate problems in the model estimation and on the stability of the coefficients. For example, the variables in the bottom-right corner of Figure 1 show very strong positive linear relations.

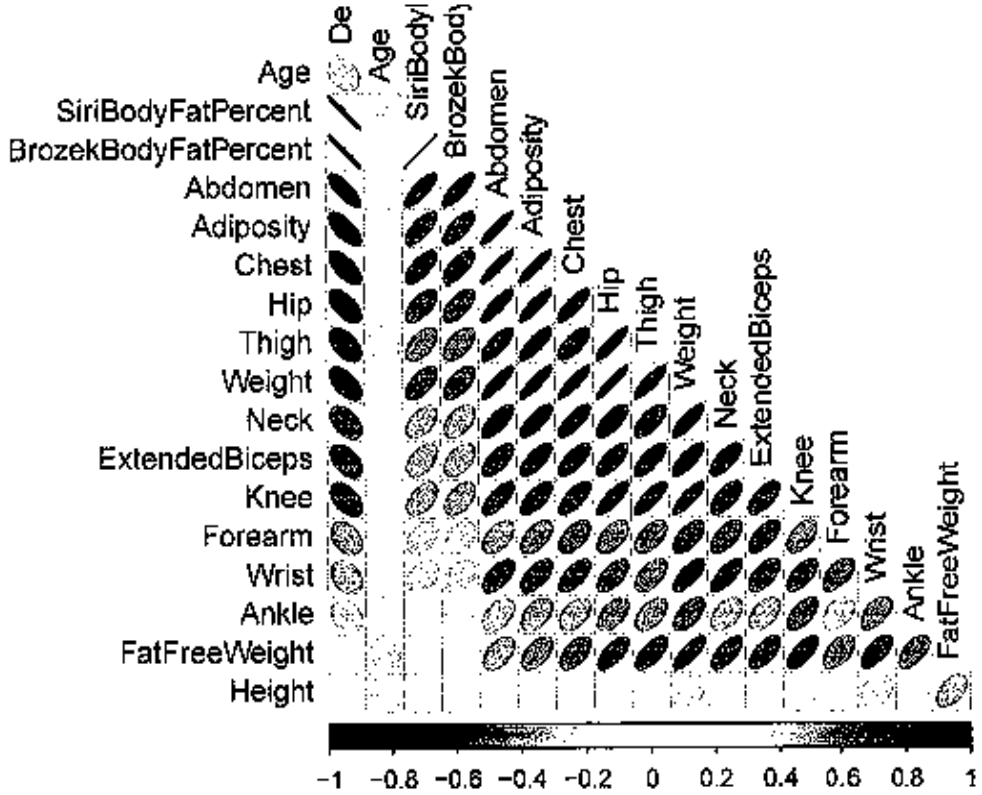
Résumé des variables

Les variables données pour expliquer le poid d'un individu sont:

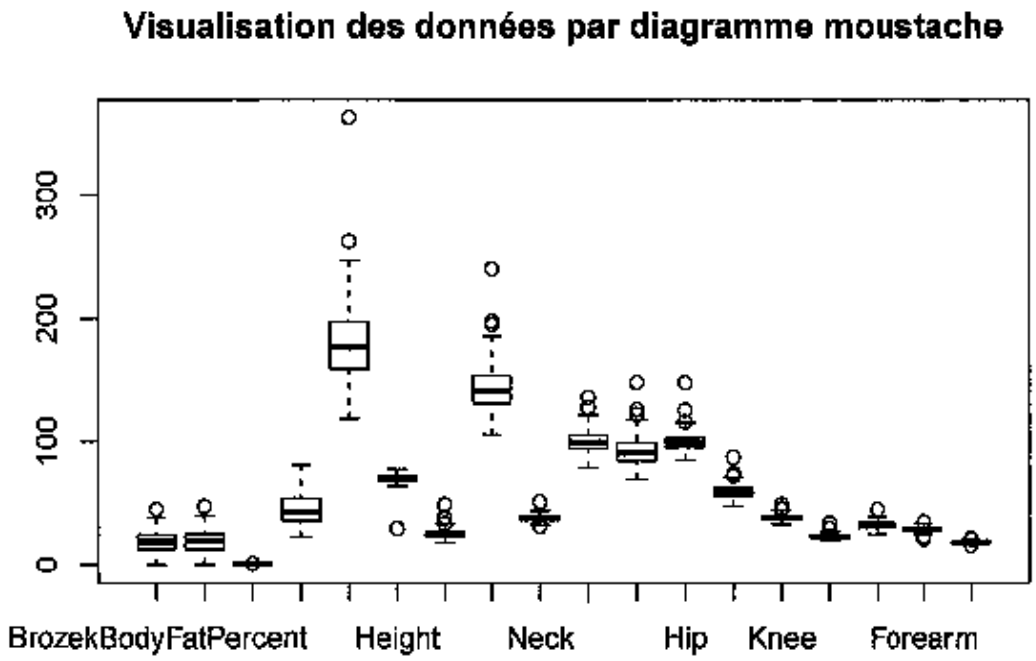
- Index : Le numéro des observations (retirées des données à l'importation)
- BrozekBodyFatPercent : Le pourcentage de graisse d'après l'équation de Brozek (457/densité-414.2)
- SiriBodyFatPercent : Le pourcentage de graisse d'après l'équation de Siri (495/densité-450)
- Density : La densité de l'individu en g par cm^3
- Age : l'age de l'individu en années
- Weight : Le poid de l'individu en livres
- Height : La taille de l'individu en pouces
- Adiposity : L'index d'adiposité en kg par m^2 (poid/taille^2)
- FatFreeWeight : Masse sèche en livres ((1 - fraction de gras selon Brozek) * poid)
- Neck : Circonférence du cou en cm
- Chest : Circonférence du torse en cm
- Abdomen : Circonférence de l'abdomon en cm
- Hip : Circonférence des hanches en cm
- Thight : Circonférence des cuisses en cm
- Knee : Circonférence des genoux en cm
- Ankle : Circonférence des chevilles en cm
- ExtendedBiceps : Circonférence des biceps au repos en cm
- Forearm : Circonférence des avants bras en cm
- Wrist : Circonférence des poignets en cm

Corrélations & Abérations

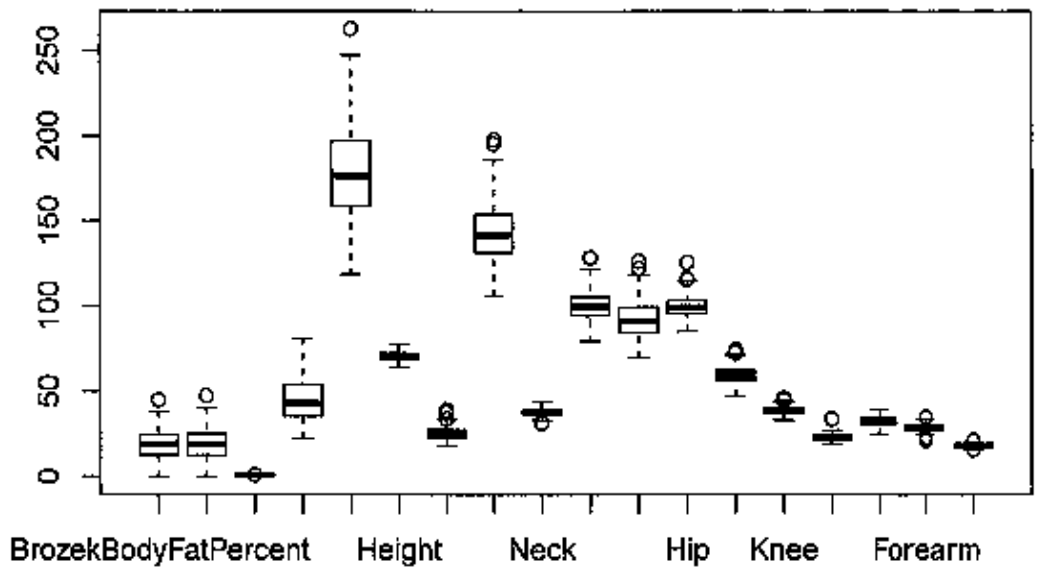
Matrice de corrélation



Le poucentage de graisse dépend linéairement de la densité et les deux méthodes de calcul sont liées de manière affine. Il sera peut être utile de ne conserver qu'une des trois variables. De plus, l'adiposité et la masse sèche sont de fonctions linéaires du poid. Leur utilisation n'est pas forcément utile pour nos modèles.



On voit apparaitre une observation abérente pour le poid. Il y a une personne avec un poid nettement supérieur aux autres, nous décidons le l'enlever. Il y a également une personne dont la taille serait fausse car trop basse selon la documentation des données. On observe qu'effectivement une personne est anormalement petite. Nous décidons de corriger sa taille selon la documentation (69.5 pouces au lieu de 29.5).



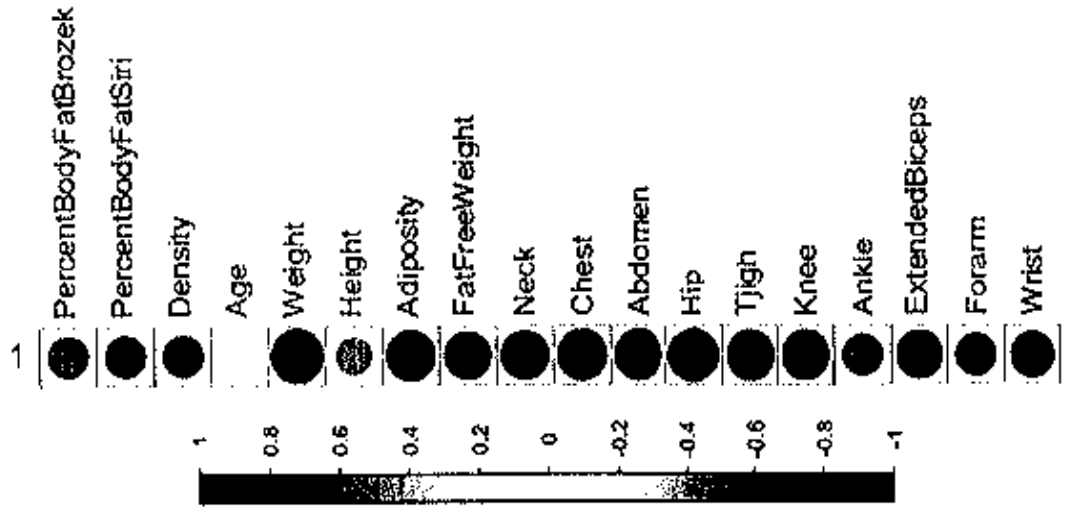
Les données sont maintenant exploitables.

Description de la base de données « Fitting percentage of body fat to simple body measurements »

La base de données est composée de 19 variables comportant chacune 252 observations. Ces 19 variables sont différentes mesures du corps humain telles que la taille, le poids ainsi que diverses circonférences des membres du corps. Ces mesures ont été faites sur 252 hommes par un docteur.

L'âge des sujets varie entre 22 ans et 81 ans avec une moyenne de 44.9 ans. Le poids varie entre 125 lbs (57kg) et 363.15 lbs (165kg) avec une moyenne de 179.16 lbs (81kg). La taille varie entre 64 (163 cm) et 77.75 inches (196cm) avec une moyenne de 70.31 (178cm)
Notons tout d'abord que les variables sont toutes quantitatives, ce qui va bien sûr, influencer la méthode de l'analyse des données. Il faut également noter que certaines erreurs avaient été relevées dans les données : les données corporelles de 3 sujets comportaient un chiffre d'erreur. Nous avons donc recalculé ces données grâce à l'équation de Brozek. Une autre erreur concernait le poids d'un sujet qui pesait plus de 200 pounds mais qui mesurait moins de 3 feet.
Nous allons à présent analyser de quelle façon est organisée cette base de données. L'objectif de notre travail est de créer un modèle ayant pour variable-cible le poids, nous allons donc réaliser plusieurs modèles puis choisir le meilleur.

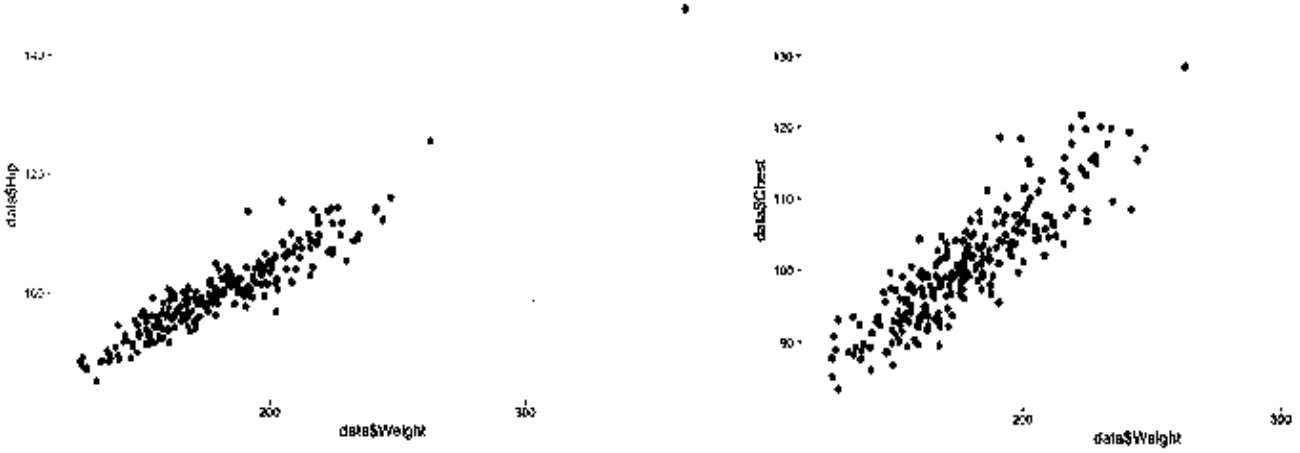
Nous pouvons, dans un premier temps, étudier la corrélation des variables avec notre variable-cible : le poids.



Les variables sont positivement corrélées lorsque le point est bleu, et négativement corrélé lorsque le point est rouge. Cela signifie que lorsque que deux variables sont positivement corrélées, si l'une d'elle augmente, alors l'autre également. Inversement, si elles sont négativement corrélées, alors si l'une augmente, l'autre va baisser.
Ici on peut remarquer que notre variable cible le poids semble être positivement corrélée à toutes les autres variables exceptée la densité où la corrélation est négative et l'âge où la corrélation est inexistante.

Nous pouvons également trouver des différences d'intensité dans toutes les variables corrélées. La variable la plus corrélée semble être la taille des hanches : en effet, c'est à cet endroit du corps que la graisse vient se loger chez la plupart des personnes (et en particulier chez les hommes) c'est donc cohérent que le poids soit fortement influencé par cette variable.

Nous allons maintenant représenter graphiquement ces deux variables pour savoir si les variables sont corrélées linéairement ou non.



Nous voyons ici pour les deux variables qui semblent être les plus corrélées à notre variable cible qu'elles le sont linéairement ce qui va nous permettre de construire un modèle linéaire. Nous pouvons de même constater que c'est le cas pour toutes les variables corrélées.

La prochaine étape de notre travail consistera en la modélisation de la variable cible « Weight » grâce aux autres variables disponibles à travers plusieurs méthodes de modélisation telles que la régression linéaire simple puis à l'aide de méthodes pénalisées telles que Ridge ou Lasso. Enfin nous testerons nos modèles afin de déterminer lequel est le meilleur.