Projet R

Objectif général : *Estimer plus facilement le pourcentage de masse grasse d'un individu*

ANDRÉ-ROURE Titouan,

RENAUD Thomas,

RICHERT Noah

TABLE DES MATIÈRES

1. Introduction

Une mesure précise de la masse graisseuse d'un individu est peu pratique et nécessite des moyens importants. Il serait intéressant de trouver une méthode simplifiée à cette estimation qui serait moins complexe/coûteuse.

2. Description du jeu de donnés utilisé

Les variables présentes dans ce jeu de donnés sont :

- Pct.BF : Le pourcentage de masse grasse
- Age (en années)
- Weight (en lbs)
- Height (en inches)
- Neck circonférence (en cm)
- Chest circonférence (en cm)
- Abdomen 2 circonférence (en cm)
- Hip circonférence (en cm)
- Thigh circonférence (en cm)
- Knee circonférence (en cm)
- Ankle circonférence (en cm)
- Biceps (extended) circonférence (en cm)
- Forearm circonférence (en cm)
- Wrist circonférence (en cm)

2. Statistiques descriptives

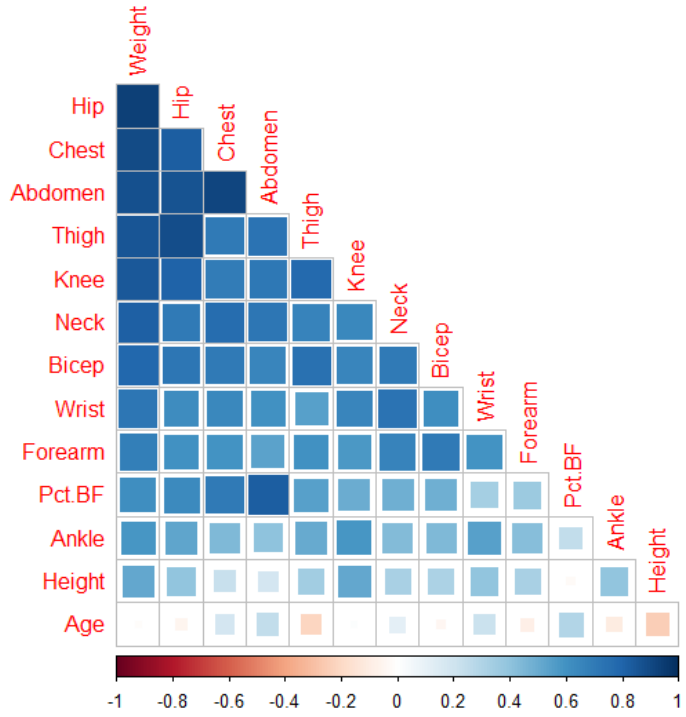

Fig. 1 : Corrélation des variables entre elles

Cette représentation en *corrplot*, nous permet d'inférer sur la corrélation de certaines variables entre elles. Ici, on se place dans le cadre d'une étude univariée, la figure 1, nous permet alors d'inférer sur la corrélation du Pct.BF (pourcentage de masse graisseuse) avec les autres variables. On observe que les variables telles que : le poids (Weight), les hanches (Hip), le torse (Chest), ainsi que l'abdomen (Abdomen) semblent très corrélées au Pct.BF, spécialement l'abdomen.

Notre analyse étant univariée, nous aurons besoin de préciser cette corrélation plus tard.

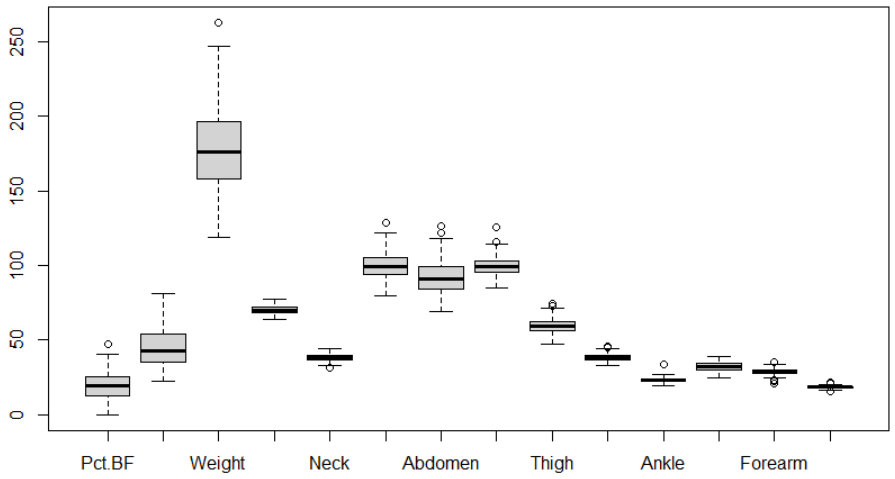

Fig. 2 : Mise en évidence des valeurs extrêmes par variables

La figure 2 nous permet de repérer d'éventuelles observations hors normes ou aberrantes, comme en témoignent les valeurs à l'extérieur des moustaches représentées par des points. Pour autant, on ne peut pas dire que ce sont forcément des observations aberrantes. Par contre cela indique qu'il faut étudier plus en détail ces observations.

3. ACP

3.1. Nettoyage des données

Nous ne sommes pas des spécialistes en biologie. Ainsi l'extraction de valeurs aberrantes en regardant uniquement leur valeurs est à proscrire. On va suivre une démarche analytique en étudiant les valeurs aberrantes dans leur globalité. Pour cela nous allons vérifier la normalité de chacune des dimensions propres de l'ACP. Dans notre cas, on s'intéresse à la normalité des 5 premières dimensions propres de l'ACP. On récupère alors la répartition des individus suivant les dimensions propres considérées pour réaliser un test de normalité de celle-ci.

Rappel test de Shapiro : H_0 : distribution normale de l'échantillon. H_1 : non H_0. Si la p-value du test est inférieur au risque de première espèce (5%) on rejette H0 et inversement.

Avec ce rappel nous remarquons que les individus suivant la dimension 2 et la dimension 5 ne sont pas distribués suivant une gaussienne. Il y a donc parmi les variables bien représentées sur ces dimensions des valeurs qui peuvent être aberrantes.

Dimension 2 :

	dim 1	dim 2	dim 3	dim 4	dim 5
Pct.BF	0.439	0.297	0.079	0.037	0.001
Age	0.002	0.609	0.314	0.007	0.004
Weight	0.957	0.003	0.001	0.005	0.007
Height	0.193	0.400	0.094	0.085	0.170
Neck	0.714	0.004	0.023	0.064	0.007
Chest	0.803	0.067	0.003	0.000	0.000
Abdomen	0.789	0.127	0.015	0.018	0.002
Hip	0.860	0.001	0.036	0.019	0.002
Thigh	0.754	0.023	0.101	0.001	0.000
Knee	0.746	0.019	0.005	0.049	0.001
Ankle	0.374	0.108	0.053	0.058	0.381
Bicep	0.688	0.005	0.006	0.096	0.000
Forearm	0.545	0.026	0.001	0.249	0.001
Wrist	0.595	0.000	0.233	0.019	0.003

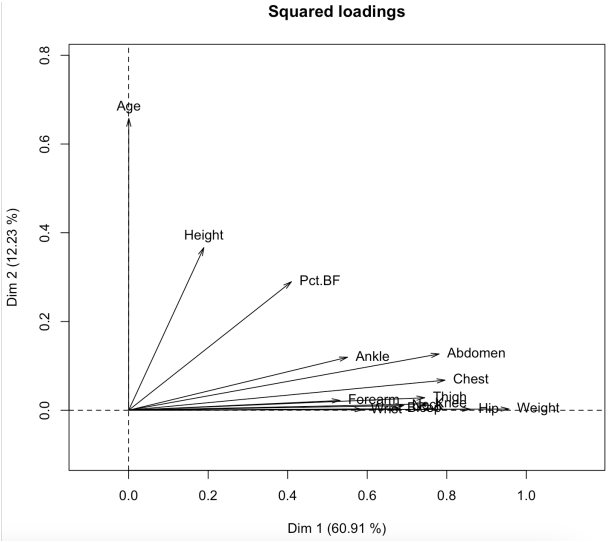

Fig. 3 : Représentation de la qualité de représentation des variables sur le plan principal 1-2

Les 3 variables les mieux représentées sur la dimension 2 sont l'âge, la taille et le pourcentage de masse graisseuse.

Représentons les histogrammes des individus suivant ces variables :

Fig. 4 : Histogrammes des individus suivant les 3 variables les mieux représentées suivant la dimension 2.

Nous supprimons les valeurs venant fausser la distribution gaussienne de ces histogrammes (les remontées aux extrémités).

Valeurs finalement conservées :

- PcT.BF : [3,35]
- Âge : [20,75]
- Height : [65,76]

Dimension 5 :

	dim 1	dim 2	dim 3	dim 4	dim 5
Pct.BF	0.409	0.290	0.117	0.027	0.003
Age	0.000	0.658	0.252	0.020	0.001
Weight	0.956	0.003	0.001	0.008	0.005
Height	0.189	0.366	0.099	0.168	0.159
Neck	0.692	0.010	0.034	0.052	0.020
Chest	0.795	0.068	0.002	0.000	0.003
Abdomen	0.781	0.127	0.019	0.019	0.001
Hip	0.859	0.002	0.033	0.018	0.000
Thigh	0.745	0.028	0.108	0.000	0.009
Knee	0.750	0.014	0.002	0.056	0.014
Ankle	0.550	0.119	0.027	0.010	0.207
Bicep	0.680	0.003	0.003	0.099	0.026
Forearm	0.532	0.022	0.010	0.245	0.010
Wrist	0.589	0.001	0.245	0.012	0.034

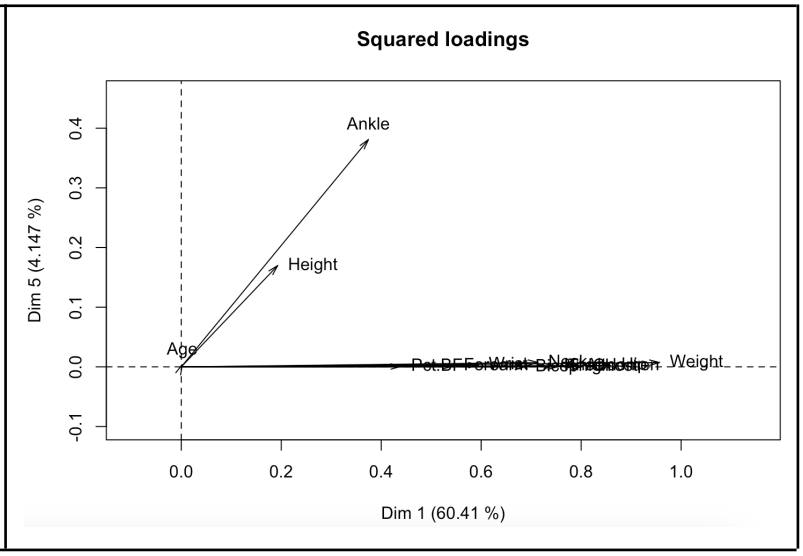

Fig. 5 : Représentation de la qualité de représentation des variables sur le plan principal 1-2

Les 2 variables les mieux représentées sur la dimension 4 sont la taille et la cheville.

Représentons uniquement l'histogramme des individus suivant la variable cheville au vu du travail de nettoyage précédent :

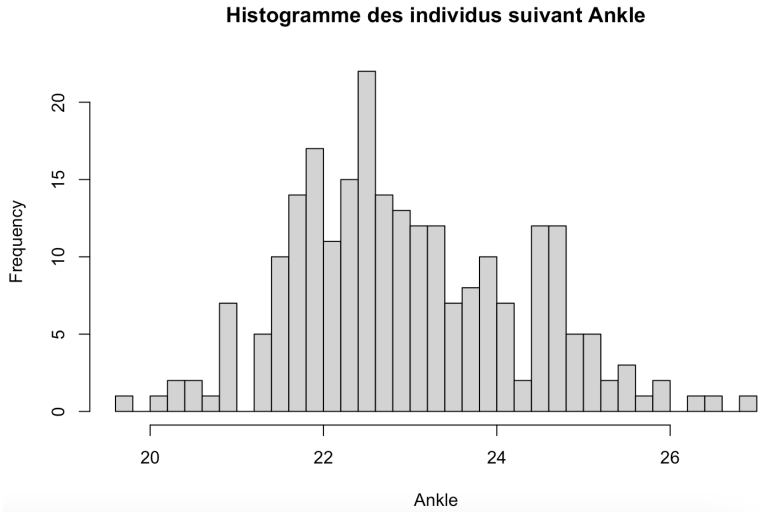

Fig. 6 : Histogramme des individus suivant la variable la mieux représentée suivant la dimension 5.

Valeurs finalement conservées :

- Cheville : [0,30]

Maintenant que le jeu de données est totalement nettoyé, nous allons pouvoir de nouveau tester la normalité de nos dimensions propres. Cette fois ci aucun problème à signaler nous pouvons commencer l'analyse multidimensionnelle de ce jeu de données.

3.2. Analyse

Notre analyse à composante principale (ACP) s'est portée sur la liaison entre le pourcentage de masse grasse et un certain nombre de variables sélectionnées. Par défaut, c'est une ACP centrée réduite qui est faite avec le même poids (1/n) pour tous les individus de l'étude. Cette méthode se voit adaptée à ce jeu de données dont les variables ont des unités ainsi qu'un variance bien différente. L'espace des individus est R^{14} car p=14 variables quantitatives ont été mesurées sur chaque individu.

La métrique associée est $M=D_{1/s2}$ pour une ACP centrée. L'espace des variables est R^{250} car chaque variable a été mesurée sur n=250 individus. La métrique associée est $D=(1/250)*I_{250}$ car tous les individus ont le même poids.

Dans un premier temps, nous allons mettre en évidence les valeurs propres et les pourcentages de variances expliquées par chaque axe :

	Eigenvalue	Proportion	Cumulative
dim 1	8.46	60.41	60.41
dim 2	1.69	12.07	72.48
dim 3	0.97	6.90	79.38
dim 4	0.71	5.06	84.44
dim 5	0.58	4.15	88.59
dim 6	0.35	2.52	91.10
dim 7	0.30	2.17	93.27
dim 8	0.27	1.89	95.17
dim 9	0.20	1.44	96.61
dim 10	0.19	1.39	98.00
dim 11	0.14	0.99	98.99
dim 12	0.08	0.55	99.54
dim 13	0.05	0.33	99.87
dim 14	0.02	0.13	100.00

Fig. 7 : Valeurs propres et pourcentages de variances expliquées par chaque axe

On est dans le cas d'une ACP centrée réduite, on peut donc utiliser le critère de Kaiser qui dit que seuls les axes factoriels ayant une valeur propre (eigenvalue) supérieure à 1 sont des axes intéressants à retenir. On a donc les deux premières valeurs propres

supérieures à 1. Selon le critère de Kaiser, il convient donc de retenir les 2 premiers axes factoriels (cf figure 7). Le premier axe factoriel permet d'expliquer 61% de l'inertie (ou de la variance), le second axe factoriel permet d'expliquer 12% d'inertie supplémentaire. Ainsi, en considérant le plan principal 1-2, on récupère 73% de l'information du jeu de données.

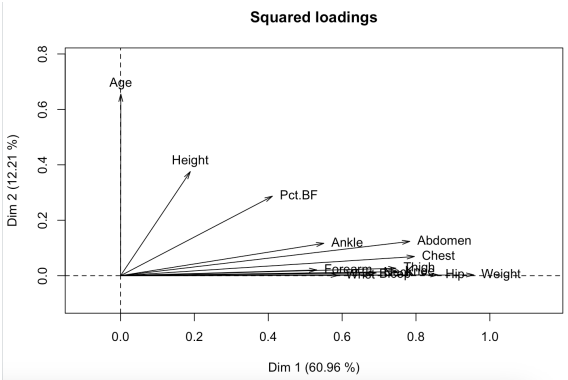

Fig. 8 : Représentation de la qualité de représentation des variables sur le plan principal 1-2

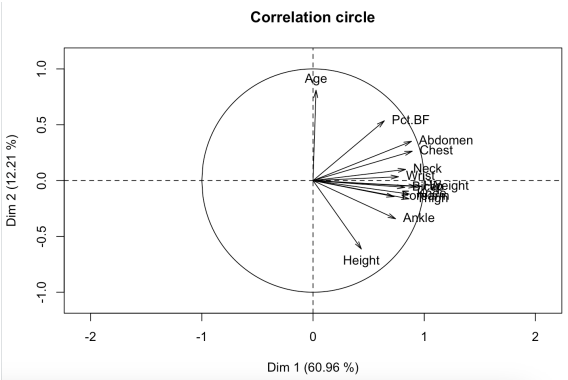

Fig. 9 : Cercle des corrélations et mise en évidence de la densité et du pourcentage de masse grasse sur le plan 1-2

Interprétation des différentes sorties de l'ACP :

Tout d'abord, d'après le critère de Kaiser, seules 2 valeurs propres ont des valeurs supérieures à 1, réduisant ainsi notre étude dans le plan 1-2 avec des variables pas toujours bien représentées. Cependant, tout n'est pas à jeter dans cette étude et plusieurs choses peuvent être analysées et mises en évidence par les différents graphiques fournis par R.

Sur la figure 8, on peut voir que plusieurs variables sont bien représentées sur l'axe 1, elles sont très proches en norme de la valeur 1 (c'est notamment le cas des cuisses, de la hanche, de l'abdomen ou encore de la poitrine). En revanche, d'autres variables comme la cheville, le poignet ou les biceps sont nettement moins bien représentées sur ce même axe. Étant limité dans le plan 1-2 par le critère de Kaiser, on ne pourra trouver un axe permettant de mieux représenter ces variables dans cette étude, comme le pourcentage de masse graisseuse qui est assez mal représenté.

Enfin, de nombreuses corrélations sont à observer entre plusieurs variables. Ces premières sont mises en évidence par de faibles angles entre 2 variables données. Ce

constat permet d'émettre quelques hypothèses quant à la réponse à donner à la problématique de ce sujet. En effet, il apparaît potentiellement possible de déterminer le taux de masse grasse global d'un individu en fonction d'un faible nombre de mensurations. Ceci faisant alors gagner du temps à quiconque souhaiterait avoir un ordre de grandeur de son taux de masse graisseuse.
En revanche, l'angle quasi droit formé entre l'âge et les différentes variables liées aux taux de graisse ou entre la taille des individus et ces mêmes variables mettent en évidence une indépendance entre elles, ce qui peut sembler étonnant.

La mise en place d'une méthode d'évaluation du taux de masse graisseuse semble pertinente. L'ACP seule ne permet pas d'aboutir à un véritable modèle. Nous allons alors compléter ce projet en réalisant une régression linéaire permettant d'obtenir un modèle complet d'évaluation de Pct.BF.

4. Régression linéaire multiple

Complétons notre analyse, avec la régression linéaire multiple de ces données. On utilise pas l'entièreté des données pour en garder quelques-unes pour tester notre modèle final.

```
> data<-data.frame(bodyfat[1:220,])
> mod1<-lm(Pct.BF~.,data)
> summary(mod1)
```

On commence par créer un premier modèle.

A. Modèle 1

Afin de déterminer la pertinence de notre modèle nous discuterons des résultats obtenus lors du *summary*.

1. Fisher

On observe une p-value inférieure à alpha (5%), on peut alors conclure sur la nullité des coefficients. L'hypothèse H_0 ("tous les coefficients sont nulles/le modèle est inutile") est donc invalidée. Le modèle est donc utile pour expliquer la variable Pct.BF.

2. R-squared

Le modèle permet d'expliquer 70 % de la variabilité de Pct.BF, 30% de la variabilité reste cependant inexpliquée. Plus ce pourcentage est proche de 100 %, plus le modèle est en adéquation avec les données. Ici, on conclut alors que le modèle est bien en adéquation avec nos données.

3. Student

On observe les valeurs qui semblent intéressantes (celles marquées par des étoiles).

- On a une p-value pour le test de Student sur *Abdomen* et *Wrist* inférieure à 5%, ce qui signifie que l'on rejette H0 et que la variable est potentiellement utile pour expliquer la variable Pct.BF.
- On a une p-value pour le test de Student pour les autres variables supérieure à 5%, ce qui signifie que l'on ne rejette pas H0 et que la variable est potentiellement inutile pour expliquer la variable Pct.BF.

Nos résultats s'expliquent par la non simplification de notre modèle. On peut en déduire qu'il va falloir créer un nouveau modèle simplifié. Il n'est donc pas nécessaire d'étudier la normalité des résidus sur ce premier modèle.

B. Sélection des variables utiles

En utilisant *step*, on a pu sélectionner les variables utiles suivant le critère AIC :
Age, Height, Neck, Abdomen, Hip, Thigh, Forearm, Wrist. Nous construirons alors le second modèle avec ces variables.

C. Modèle 2

1. Fisher

On observe une p-value inférieure à alpha (5%), on peut alors conclure sur la nullité des coefficients. L'hypothèse H0 ("tous les coefficients sont nulles/le modèle est inutile") est donc invalidée. Le modèle est donc utile pour expliquer la variable Pct.BF.

2. R-squared

Le modèle permet d'expliquer environ 70 % de la variabilité de Pct.BF, 30% de la variabilité reste cependant inexpliquée. Plus ce pourcentage est proche de 100 %, plus le modèle est en adéquation avec les données. Ici, on conclut alors que le modèle est bien en adéquation avec nos données. Cependant, il n'y a pas d'amélioration par rapport au premier modèle.

3. Student

On observe les valeurs qui semblent intéressantes (celles marquées par des étoiles).

- Les p-values pour le test de Student sur les variables *Wrist, Abdomen, Neck* sont inférieures à 5%, ce qui signifie que l'on rejette H_0 et que ces variables sont potentiellement utiles et pertinentes pour expliquer la variable Pct.BF.
- Les variables *Age, Height* et *Thigh* ont une p-values supérieures mais proches de 5%. Elles peuvent aussi être sélectionnées pour l'explication de la variable Pct.BF.
- Les variables *Hip* et *Forearm* ont des p-values entre 10% et 15% ce qui montre qu'elles sont peu utiles pour décrire le modèle. Cependant, la fonction *step* trouve le

modèle le plus efficace en utilisant le critère AIC. Lorsque l'on compare des modèles ajustés par le maximum de vraisemblance aux mêmes données, plus l'AIC est petit, plus l'ajustement est bon.

4. Etude des résidus

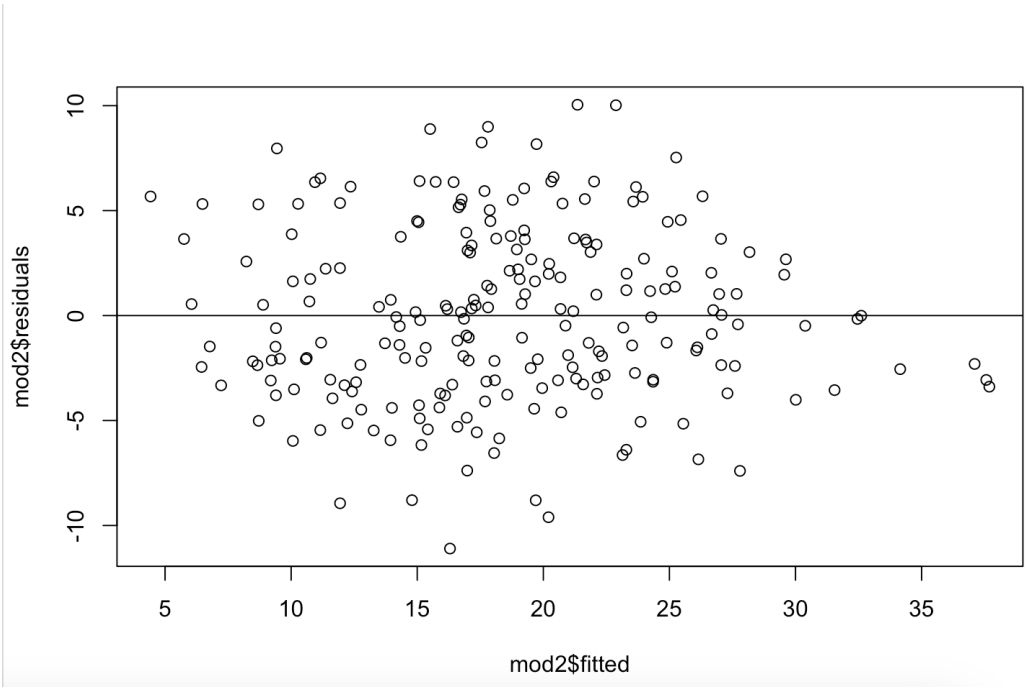

Fig. 10 : Représentation de l'organisation des résidus en fonction de la valeur prédite

Le graphique (cf figure 10) des résidus ne révèle aucune structure particulière. Il est bien le reflet d'un comportement aléatoire (les résidus pouvant être vus comme les réalisations du terme d'erreur aléatoire ε du modèle). En conclusion, ce graphique nous permet de dire qu'il n'y a donc pas d'informations dans les résidus qui auraient dû être prises en compte dans le modèle. Le test de normalité des résidus (test de Shapiro-Wilk) permet de valider l'hypothèse H_0 de normalité des résidus avec une p-value de 12% supérieure à $\alpha = 5\%$.

Avec les données et les sorties disponibles, il ne paraît pas utile de poursuivre d'autres pistes pour améliorer le modèle 2. Seule la mise à disposition de nouvelles variables explicatives pourrait être intéressante.

5. Prédiction

Nous récupérons maintenant la partie des données réelles non exploitées. 200 valeurs ont donc été sélectionnées afin de créer notre modèle "entraîné", dont nous pourrons observer les comportements sur la prédiction des 39 autres valeurs.

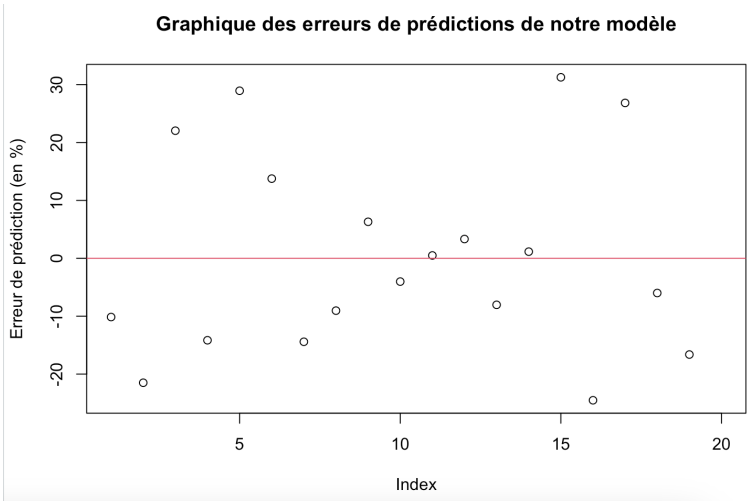

Fig. 11 : Erreur de prédiction de notre modèle.

En conclusion nous avons un modèle qui se trompe en valeur absolue de 13,8%, ce qui est convenable mais à améliorer. Comme nous pouvions le prédire, estimer le pourcentage de masse grasse avec uniquement des mensurations s'avère compliqué. Un contre exemple symbolisant est un bodybuilder avec un pourcentage de masse graisseuse très faible mais des mensurations très importantes venant déjouer notre modèle.