

# Estimation de la Condition Physique Liée à la Santé à l'Aide de la Régression Linéaire Multiple

**Analyse de l'Article Préparée par :**

Mohamed Khalil Kamessi

Farah Baklouti

Mohamed Ali Jabri

Ines Bendhifallah

Najeh Benrebeh

<b>Encadrant Académique :</b>	Youssef Mejri
<b>Matière :</b>	Statistiques
<b>Type de Projet :</b>	Analyse d'un Article Scientifique
<b>Année Universitaire :</b>	2025-2026

# Table des matières

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	Contexte du projet . . . . .	2
1.2	Objectif du projet . . . . .	2
1.3	Description des données . . . . .	3
<b>2</b>	<b>Importation et Préparation des Données</b>	<b>3</b>
2.1	Chargement des packages . . . . .	3
2.2	Importation et exploration initiale . . . . .	5
2.3	Vérification de la densité des variables . . . . .	7
2.4	Détection des valeurs aberrantes . . . . .	14
2.5	Vérification des corrélations . . . . .	15
<b>3</b>	<b>Analyse Statistique Avancée</b>	<b>29</b>
3.1	Vérification de la densité des variables . . . . .	29
3.2	Convergence vers la normalité des moyennes (Bootstrap) . . . . .	31
3.3	Test ANOVA . . . . .	33
3.3.1	Création des groupes . . . . .	33
3.3.2	ANOVA et tests de Bartlett . . . . .	34
3.3.3	Conclusion . . . . .	38
3.4	Comparaison selon le sexe : analyse des moyennes et des variances . . . . .	38
3.4.1	Indice de masse corporelle (BMI) . . . . .	39
3.4.2	Flexibilité (Sit and Reach) . . . . .	40
3.4.3	Capacité cardiorespiratoire ( $VO_2$ estimé) . . . . .	41
3.4.4	Âge . . . . .	41
3.4.5	Pourcentage de masse grasse . . . . .	42
3.4.6	Force de préhension (Hand Grip Strength) . . . . .	43
3.4.7	Endurance musculaire (Nombre de sit-ups) . . . . .	43
3.4.8	Conclusion . . . . .	44
3.5	Analyse comparative selon les groupes d'âge . . . . .	44
3.6	Test de Proportions . . . . .	48
3.6.1	Sexe (Homme vs Femme) . . . . .	49
3.6.2	Âge (Moins de 35 ans vs 35 ans et plus) . . . . .	50
3.6.3	IMC (Inférieur à 25 vs Supérieur ou égal à 25) . . . . .	50

3.7	Modèle pour la force de préhension (Hand Grip Strength) . . . . .	50
3.8	Modèle pour la $\text{VO}_2$ estimée ( $\text{VO}_2$ max) . . . . .	55
3.9	Synthèse et Comparaison des Modèles . . . . .	60
<b>4</b>	<b>Conclusion Générale</b>	<b>61</b>

# 1 Introduction

## 1.1 Contexte du projet

Dans le domaine des sciences du sport, de la santé publique et de la biométrie, l'analyse statistique des données physiologiques et anthropométriques joue un rôle fondamental dans la compréhension des déterminants de la condition physique humaine. Les indicateurs tels que l'âge, l'indice de masse corporelle (IMC), le pourcentage de masse grasse, la force musculaire, l'endurance et la capacité cardiovasculaire sont largement utilisés pour évaluer l'état de santé et le niveau de forme physique des individus.

Cependant, ces variables présentent souvent des distributions complexes, influencées par des facteurs biologiques, sociaux et comportementaux, notamment le sexe. L'étude statistique rigoureuse de ces données nécessite donc l'utilisation conjointe de méthodes descriptives, de tests de normalité et de tests paramétriques et non paramétriques permettant d'identifier les relations significatives entre les variables.

Dans ce contexte, le présent projet s'inscrit dans une démarche d'analyse statistique approfondie d'un jeu de données réel contenant des informations démographiques, anthropométriques et des mesures de performance physique. Ce travail vise à appliquer différentes techniques statistiques afin de valider des hypothèses, d'évaluer la structure des données et de comparer les groupes selon des variables pertinentes.

Ce projet constitue également une application pratique des concepts théoriques abordés dans le cadre du cours de statistique, en mettant l'accent sur la rigueur méthodologique, l'interprétation correcte des résultats et la justification du choix des tests statistiques utilisés.

## 1.2 Objectif du projet

L'objectif principal de ce projet est de réaliser une analyse statistique complète et structurée d'un ensemble de données physiologiques et anthropométriques, en vue de mieux comprendre les relations entre les caractéristiques individuelles et les performances physiques.

Plus précisément, ce projet vise à :

- Explorer et décrire les données à l'aide de statistiques descriptives afin d'identifier les tendances centrales, la dispersion et la structure globale des variables étudiées.
- Vérifier les hypothèses de normalité des variables quantitatives à l'aide de tests statistiques appropriés, notamment le test de Shapiro-Wilk, afin de justifier le choix des méthodes paramétriques ou non paramétriques.
- Détecter et analyser les valeurs aberrantes susceptibles d'influencer les résultats statistiques.
- Comparer les groupes définis par la variable *sexe* à l'aide de tests paramétriques (tests t, ANOVA, tests de proportion) et non paramétriques (tests de Wilcoxon, tests des signes), selon la nature et la distribution des données.
- Étudier les relations entre les variables à travers l'analyse des corrélations et l'évaluation de leur significativité statistique.

À travers ces objectifs, ce projet cherche à démontrer la capacité des outils statistiques à fournir des résultats fiables et interprétables, tout en mettant en évidence l'importance du choix méthodologique dans l'analyse des données réelles.

### 1.3 Description des données

Les données utilisées dans ce projet proviennent d'un ensemble de mesures physiologiques, anthropométriques et de performance physique collectées auprès d'un échantillon d'individus. Chaque observation correspond à un individu et regroupe des informations démographiques, des caractéristiques corporelles ainsi que des indicateurs de condition physique.

Les variables étudiées ont été sélectionnées afin de couvrir différents aspects de la condition physique, notamment la composition corporelle, la force musculaire, l'endurance musculaire, la flexibilité et la capacité cardiorespiratoire. La présence de variables qualitatives et quantitatives permet de mener des analyses statistiques variées, incluant des comparaisons de groupes et des tests d'hypothèses.

La variable *sexe* est utilisée comme variable de stratification afin d'analyser les différences potentielles entre les groupes masculins et féminins, tandis que les variables quantitatives sont exploitées pour évaluer les relations statistiques et les performances physiques.

Table 1: Description des variables du jeu de données

Variable	Type	Description
sex	Qualitative	Sexe de l'individu (M : masculin, F : féminin)
age	Quantitative	Âge de l'individu (en années)
bmi	Quantitative	Indice de masse corporelle ( $\text{kg}/\text{m}^2$ )
percent_body_fat	Quantitative	Pourcentage de masse grasse corporelle (%)
hand_grip_strength_kg	Quantitative	Force de préhension manuelle (kg)
sit_ups_count	Quantitative	Nombre de redressements assis réalisés
sit_and_reach_cm	Quantitative	Mesure de flexibilité (cm)
vo2_estimate_ml_per_kg_min	Quantitative	Estimation du $\text{VO}_2\text{max}$ ( $\text{ml}/\text{kg}/\text{min}$ )

## 2 Importation et Préparation des Données

Cette section a pour objectif de présenter les étapes préliminaires nécessaires à l'analyse statistique des données. Elle comprend le chargement des bibliothèques requises, l'importation du jeu de données, une exploration initiale des variables ainsi que l'étude de la distribution des variables quantitatives. Ces étapes sont essentielles pour assurer la qualité des analyses statistiques réalisées par la suite et pour orienter le choix des méthodes appropriées.

### 2.1 Chargement des packages

L'analyse des données a été réalisée à l'aide du logiciel R, en utilisant plusieurs bibliothèques spécialisées. Ces packages permettent respectivement la lecture des fichiers de données, la manipulation des données, la visualisation graphique, l'analyse statistique et la modélisation.

- `openxlsx` et `readxl` : importation de fichiers Excel ;
- `dplyr` et `tidyr` : manipulation et transformation des données ;
- `ggplot2` et `gridExtra` : visualisation graphique ;
- `psych` : statistiques descriptives ;
- `corrplot` : visualisation des corrélations ;
- `MASS` : méthodes statistiques avancées, notamment la régression pas à pas ;

- `naniar` : détection et visualisation des valeurs manquantes.

```
library(openxlsx)
```

```
## Warning: package 'openxlsx' was built under R version 4.5.2
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.5.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.5.2
```

```
##
```

```
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
```

```
##      filter, lag
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.5.2
```

```
##
```

```
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
```

```
##
```

```
##      %+%, alpha
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.5.2
```

```
## corrplot 0.95 loaded
```

```
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.5.2
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##      select

library(naniar)

## Warning: package 'naniar' was built under R version 4.5.2
```

```
library(tidyr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##      combine
```

## 2.2 Importation et exploration initiale

Le jeu de données est importé à partir d'un fichier Excel à l'aide de la fonction `read.xlsx`. Une fois les données chargées, une exploration initiale est réalisée afin de visualiser les premières observations, de résumer les principales caractéristiques statistiques et d'examiner la structure des variables.

Les fonctions `head()`, `summary()` et `str()` sont utilisées pour obtenir respectivement un aperçu des premières lignes du jeu de données, des statistiques descriptives générales et des informations sur le type de chaque variable.

```
## participant_id sex age measurement_date bmi percent_body_fat
## 1             1  M  46      2019-01-19 26.5             45.0
## 2             2  M  25      2018-03-16 28.8             45.0
## 3             3  M  53      2017-06-02 25.7             45.0
## 4             4  M  46      2018-02-11 28.0             45.0
## 5             5  F  32      2018-01-25 25.0             57.7
## 6             6  F  28      2018-11-06 24.7             60.0
## hand_grip_strength_kg sit_and_reach_cm sit_ups_count
## 1             47.0             13.4             13
## 2             28.0             -1.2             19
## 3             37.7              4.5             18
## 4             48.9              9.0             20
## 5             20.0             14.8             23
## 6             21.2             13.1             27
## vo2_estimate_ml_per_kg_min
## 1             29.8
## 2             46.5
## 3             37.2
## 4             52.5
## 5             24.6
## 6             30.1
```

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
if("participant_id" %in% colnames(data)) {  
  data <- data %>% dplyr::select(-participant_id)  
}  
head(data)
```

```
##   sex age measurement_date  bmi percent_body_fat hand_grip_strength_kg  
## 1  M  46      2019-01-19 26.5          45.0          47.0  
## 2  M  25      2018-03-16 28.8          45.0          28.0  
## 3  M  53      2017-06-02 25.7          45.0          37.7  
## 4  M  46      2018-02-11 28.0          45.0          48.9  
## 5  F  32      2018-01-25 25.0          57.7          20.0  
## 6  F  28      2018-11-06 24.7          60.0          21.2  
##   sit_and_reach_cm sit_ups_count vo2_estimate_ml_per_kg_min  
## 1              13.4             13              29.8  
## 2              -1.2             19              46.5  
## 3               4.5             18              37.2  
## 4               9.0             20              52.5  
## 5              14.8             23              24.6  
## 6              13.1             27              30.1
```

```
summary(data)
```

```
##      sex           age      measurement_date      bmi  
## Length:2000      Min.   :19.00      Length:2000      Min.   :15.00  
## Class :character  1st Qu.:30.00      Class :character  1st Qu.:21.80  
## Mode  :character  Median :41.00      Mode  :character  Median :24.00  
##                               Mean  :41.54              Mean  :23.98  
##                               3rd Qu.:53.00              3rd Qu.:26.10  
##                               Max.   :64.00              Max.   :37.10  
## percent_body_fat hand_grip_strength_kg sit_and_reach_cm sit_ups_count  
## Min.   :26.50      Min.   : 7.40      Min.   : -14.900      Min.   : 5.00  
## 1st Qu.:45.00      1st Qu.:27.88      1st Qu.:  3.700      1st Qu.:15.00  
## Median :45.00      Median :36.40      Median :  7.900      Median :17.00  
## Mean   :47.80      Mean   :35.96      Mean   :  7.887      Mean   :17.62  
## 3rd Qu.:53.33      3rd Qu.:43.70      3rd Qu.:11.900      3rd Qu.:20.00  
## Max.   :60.00      Max.   :68.00      Max.   :34.400      Max.   :33.00  
## vo2_estimate_ml_per_kg_min  
## Min.   :10.00  
## 1st Qu.:31.50  
## Median :36.40  
## Mean   :36.28  
## 3rd Qu.:41.20  
## Max.   :60.10
```

```
str(data)
```

```
## 'data.frame':   2000 obs. of  9 variables:  
## $ sex           : chr  "M" "M" "M" "M" ...  
## $ age           : num  46 25 53 46 32 28 34 34 49 60 ...
```



```
## $ measurement_date      : chr  "2019-01-19" "2018-03-16" "2017-06-02" "2018-02-11" ...
## $ bmi                   : num  26.5 28.8 25.7 28 25 24.7 24.7 27.8 25 24.6 ...
## $ percent_body_fat      : num  45 45 45 45 57.7 60 60 45 45 45 ...
## $ hand_grip_strength_kg : num  47 28 37.7 48.9 20 21.2 28.9 38.2 41.9 34.1 ...
## $ sit_and_reach_cm      : num  13.4 -1.2 4.5 9 14.8 13.1 3.9 17.1 9.9 14.6 ...
## $ sit_ups_count         : num  13 19 18 20 23 27 17 12 12 15 ...
## $ vo2_estimate_ml_per_kg_min: num  29.8 46.5 37.2 52.5 24.6 30.1 30 45.7 39.5 28.8 ...
```

```
num_vars <- c("bmi", "percent_body_fat", "hand_grip_strength_kg",
              "sit_and_reach_cm", "sit_ups_count", "vo2_estimate_ml_per_kg_min")
```

## 2.3 Vérification de la densité des variables

L'étude de la distribution des variables quantitatives constitue une étape clé de l'analyse exploratoire. Elle permet d'évaluer visuellement la forme des distributions, d'identifier d'éventuelles asymétries et d'anticiper le choix des tests statistiques appropriés.

Pour chaque variable numérique, un histogramme accompagné de sa courbe de densité est tracé. Cette représentation graphique facilite l'appréciation de la normalité ou de la non-normalité des distributions et met en évidence la présence éventuelle de valeurs extrêmes.

```
numeric_vars <- data %>% dplyr:: select(where(is.numeric))

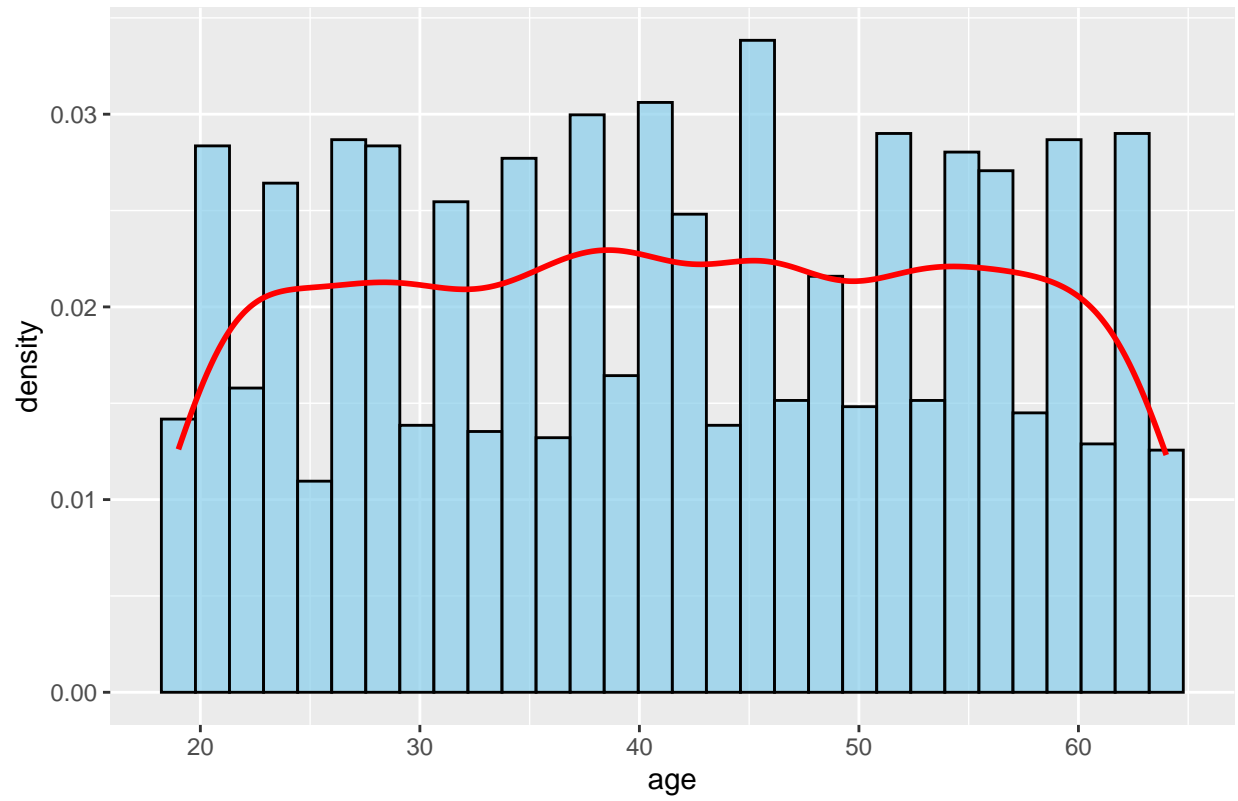
for (col in names(numeric_vars)) {
  print(
    ggplot(data, aes_string(col)) +
      geom_histogram(aes(y = ..density..), bins = 30,
                     fill = "skyblue", color = "black", alpha = 0.7) +
      geom_density(color = "red", size = 1) +
      ggtitle(paste("Histogramme + Densité :", col))
  )
}

## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()``.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

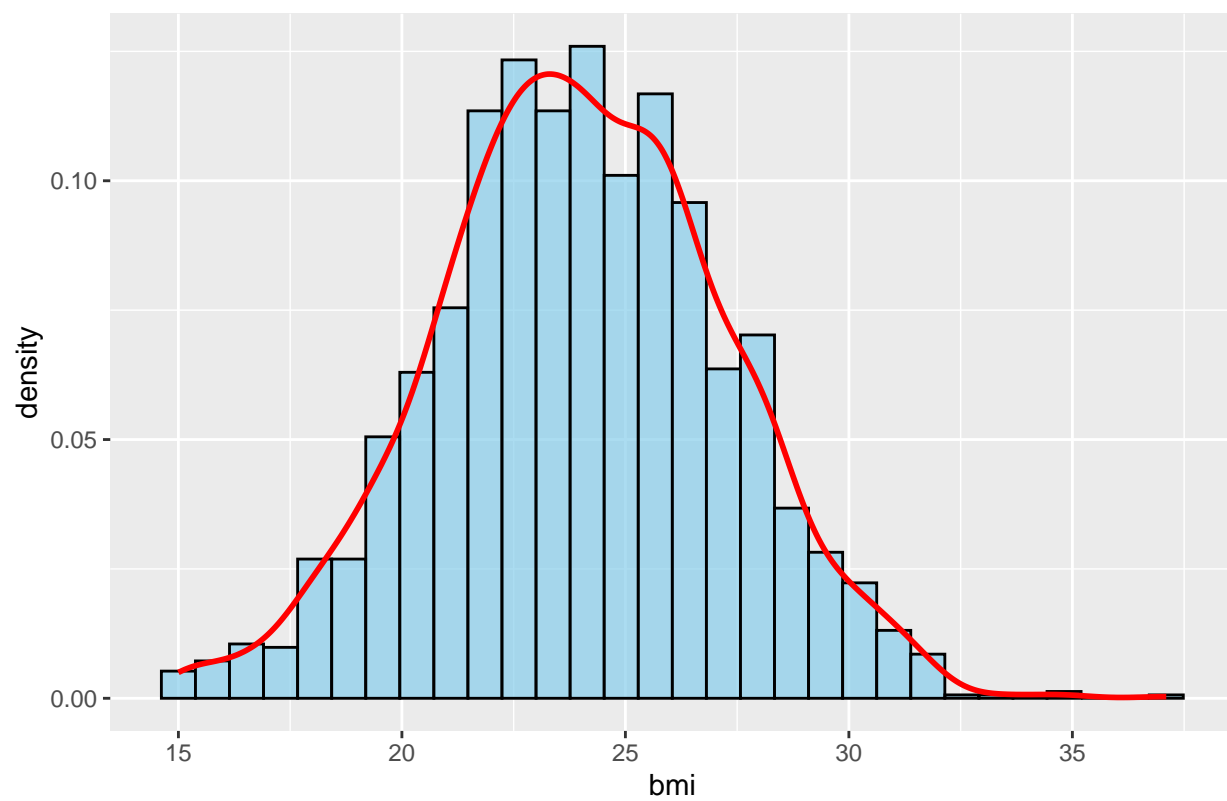
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## Warning: The dot-dot notation (`..density..`) was deprecated in ggplot2 3.4.0.
## i Please use `after_stat(density)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

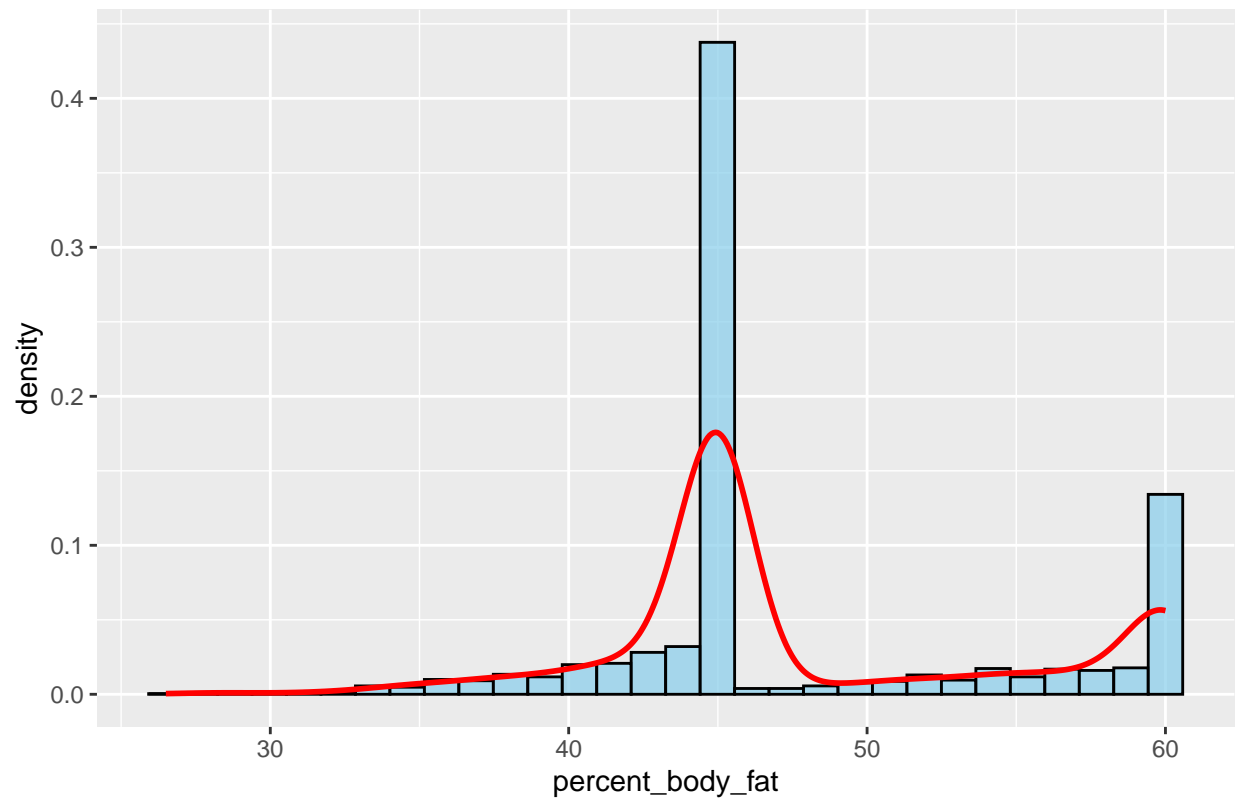
Histogramme + Densité : age



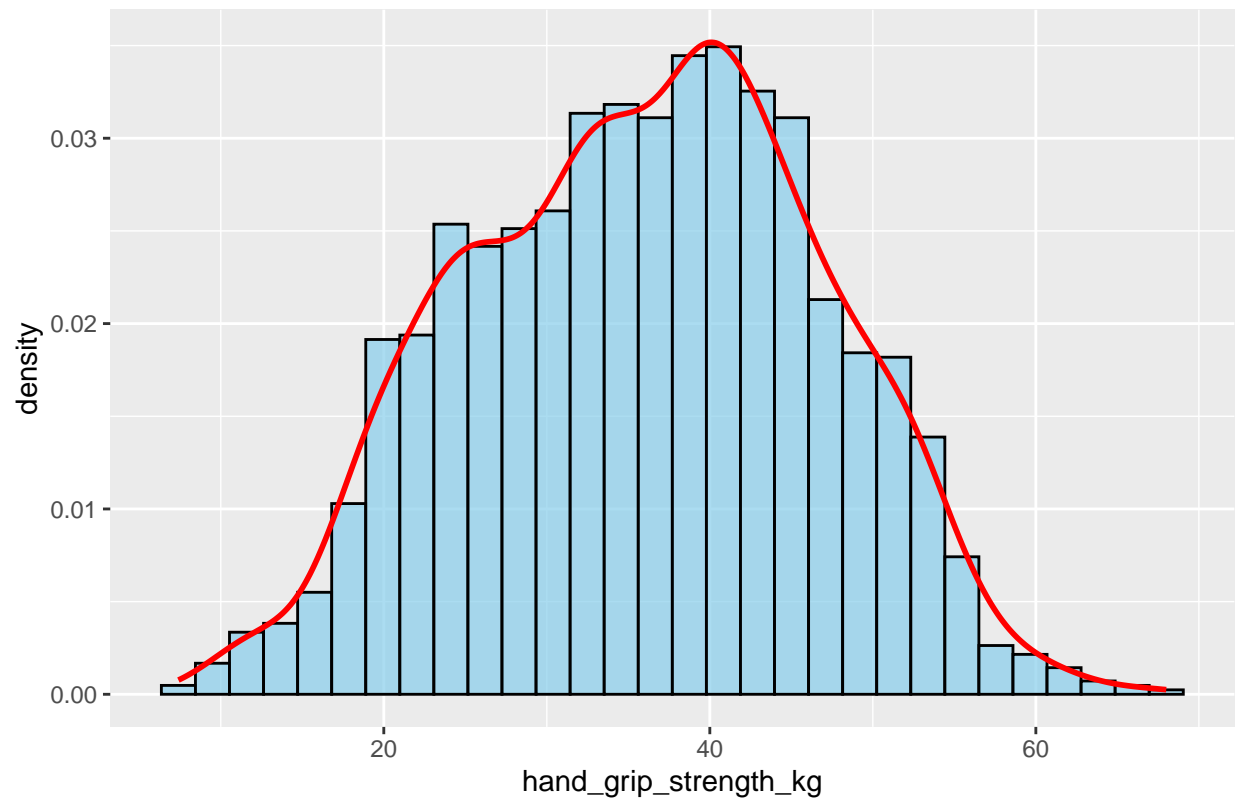
Histogramme + Densité : bmi



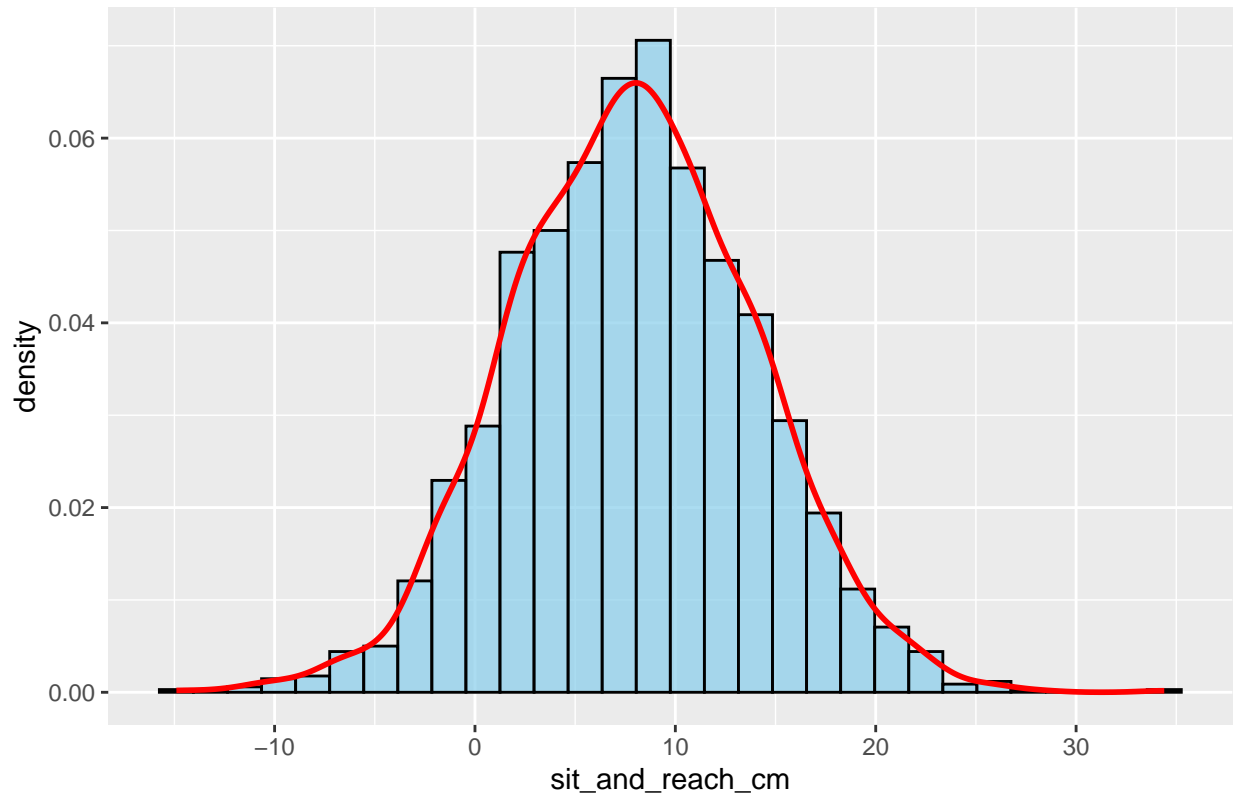
Histogramme + Densité : percent\_body\_fat



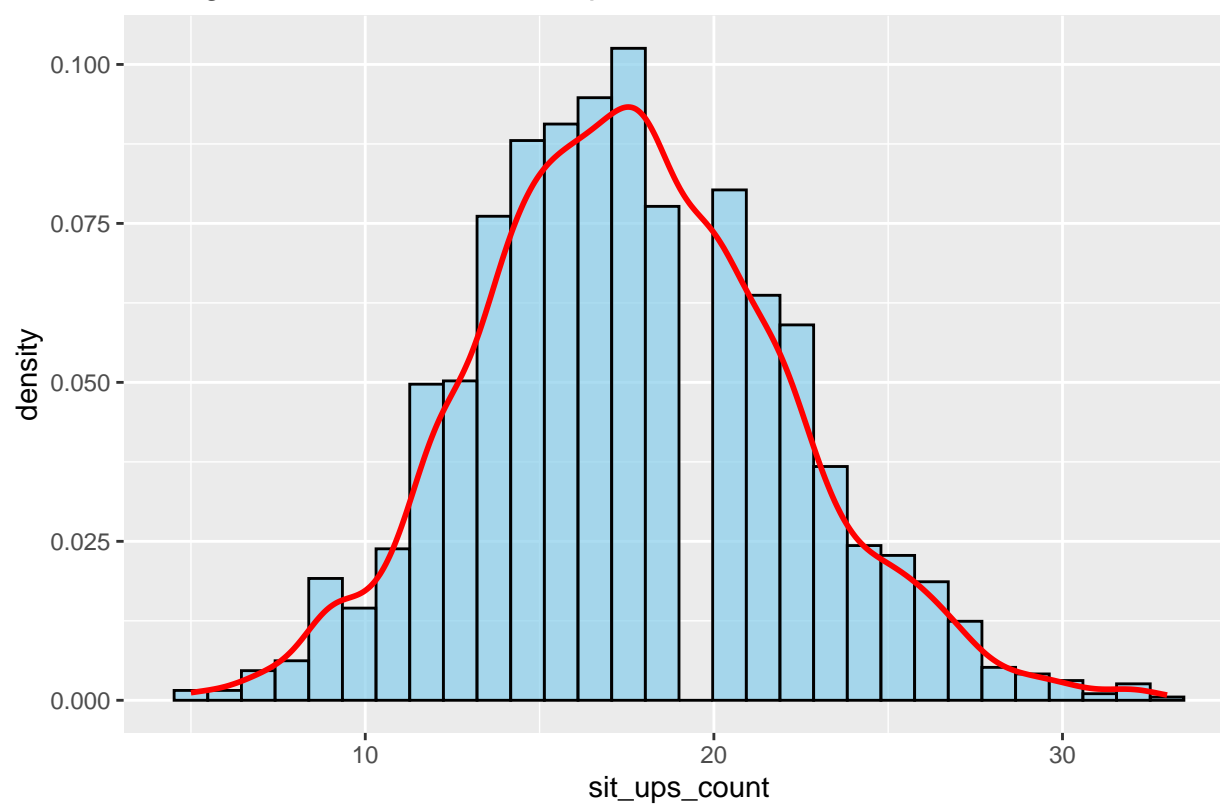
Histogramme + Densité : hand\_grip\_strength\_kg

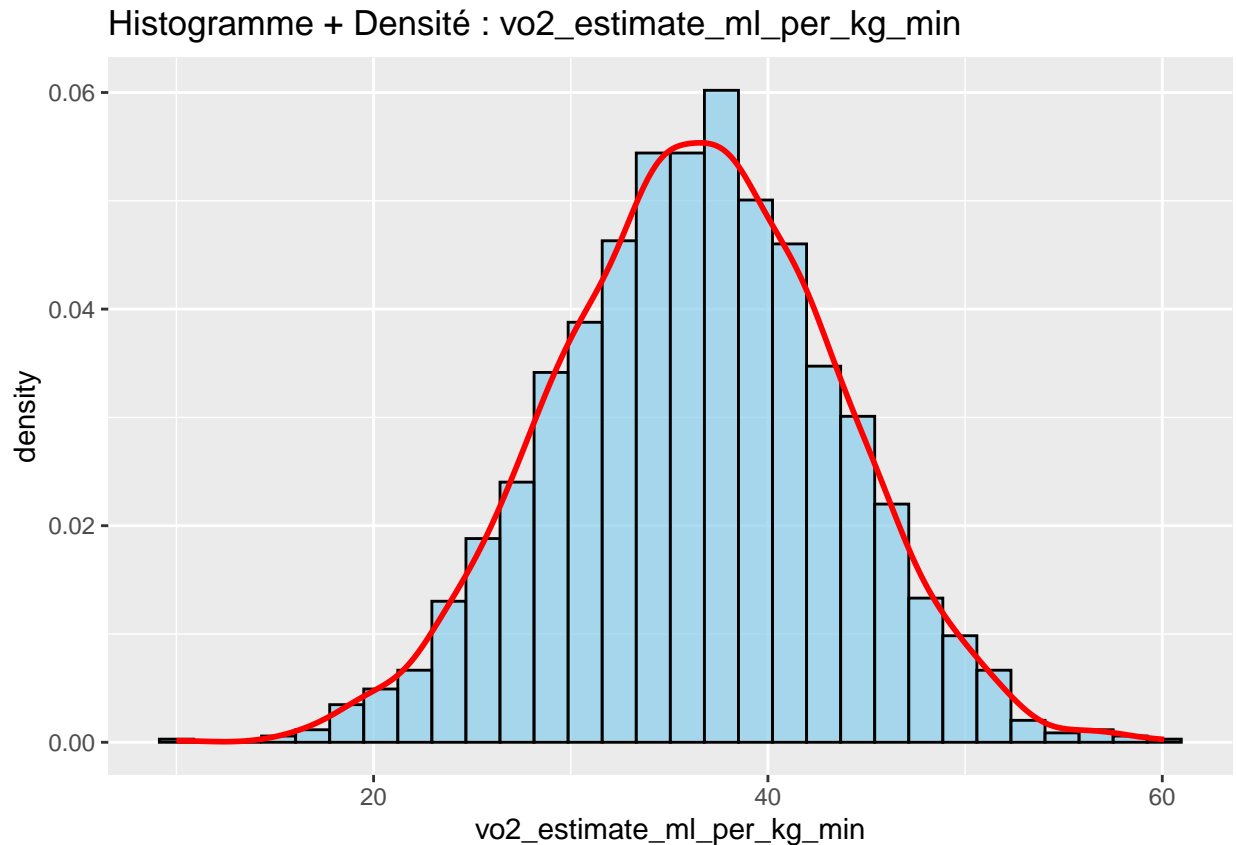


Histogramme + Densité : sit\_and\_reach\_cm



Histogramme + Densité : sit\_ups\_count





## 2.4 Détection des valeurs aberrantes

La détection des valeurs aberrantes constitue une étape importante de l'analyse exploratoire des données, car la présence d'observations extrêmes peut influencer les statistiques descriptives, les tests d'hypothèses et les modèles statistiques ultérieurs.

Dans cette étude, les valeurs aberrantes sont identifiées à l'aide d'une approche graphique basée sur les *boxplots*. Afin de faciliter la comparaison simultanée des différentes variables numériques, les données sont d'abord transformées au format long, chaque observation étant associée à une variable et à une valeur correspondante.

Les boxplots représentent la distribution des données à l'aide de la médiane, des quartiles et des étendues interquartiles. Les observations situées en dehors des bornes définies par l'intervalle  $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$  sont considérées comme des valeurs aberrantes et sont mises en évidence graphiquement.

Cette visualisation permet d'identifier rapidement les variables présentant une forte dispersion ou un nombre important de valeurs extrêmes.

```
data_long_before <- data %>%
  dplyr::select(all_of(num_vars)) %>%
  pivot_longer(cols = everything(),
               names_to = "variable",
               values_to = "value")

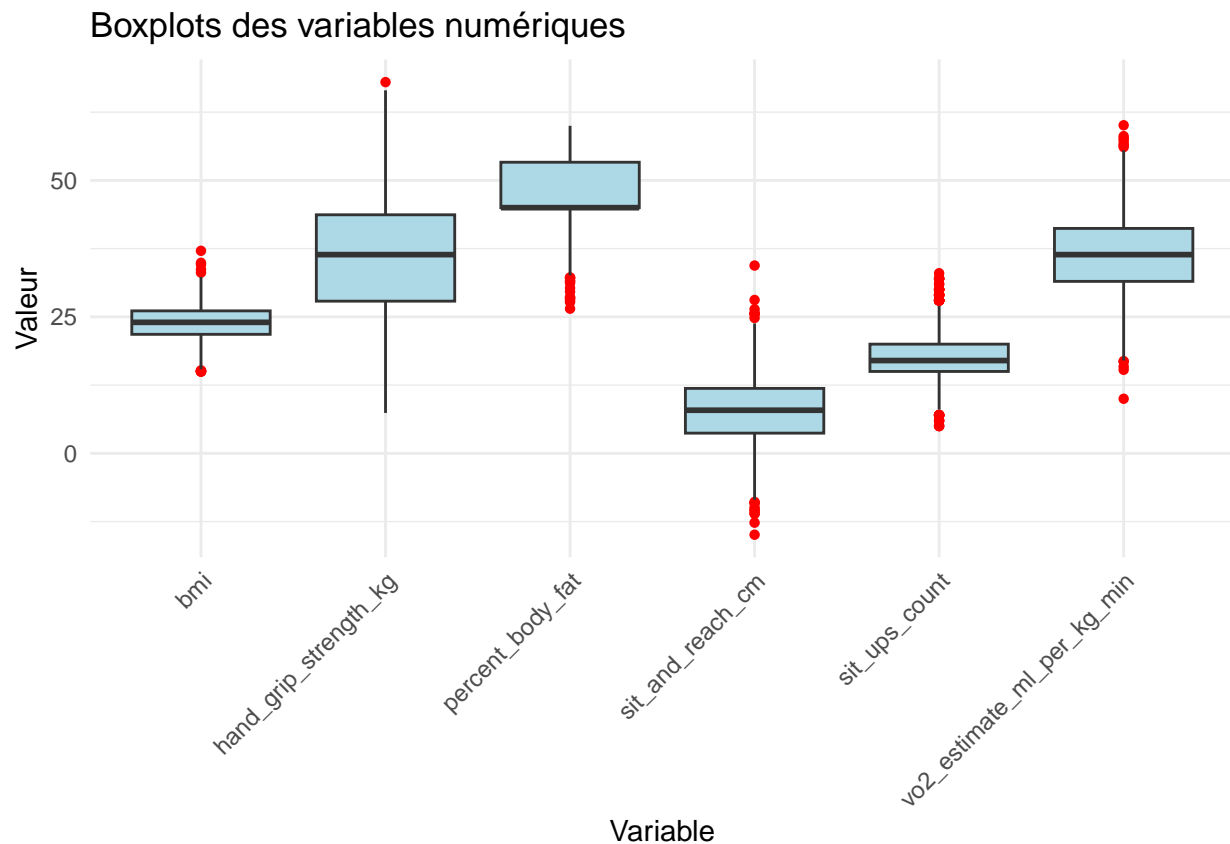
boxplot_before <- ggplot(data_long_before, aes(x = variable, y = value)) +
  geom_boxplot(fill = "lightblue", outlier.color = "red", outlier.shape = 16) +
```



```

theme_minimal() +
labs(title = "Boxplots des variables numériques",
     x = "Variable",
     y = "Valeur") +
theme(axis.text.x = element_text(angle = 45, hjust = 1))
print(boxplot_before)

```



## 2.5 Vérification des corrélations

% 2.6 Vérification des corrélations

```

y_vars <- c(
  "sit_and_reach_cm",
  "vo2_estimate_ml_per_kg_min",
  "sit_ups_count",
  "hand_grip_strength_kg"
)

x_vars <- c("age", "bmi", "percent_body_fat")

```

```

library(ggplot2)

for (y in y_vars) {
  for (x in x_vars) {

```

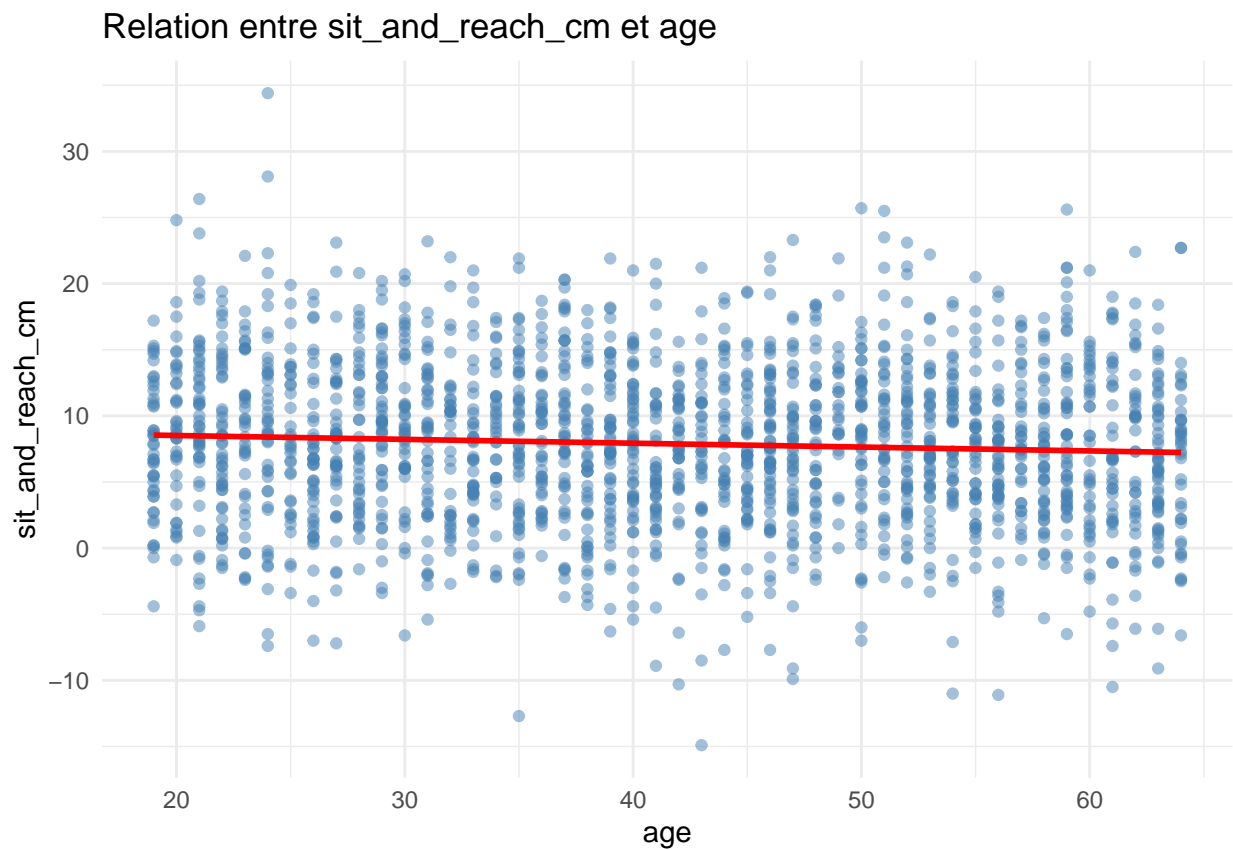
```

p <- ggplot(data, aes_string(x = x, y = y)) +
  geom_point(alpha = 0.5, color = "steelblue") +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  theme_minimal() +
  labs(
    title = paste("Relation entre", y, "et", x),
    x = x,
    y = y
  )

print(p)
}
}

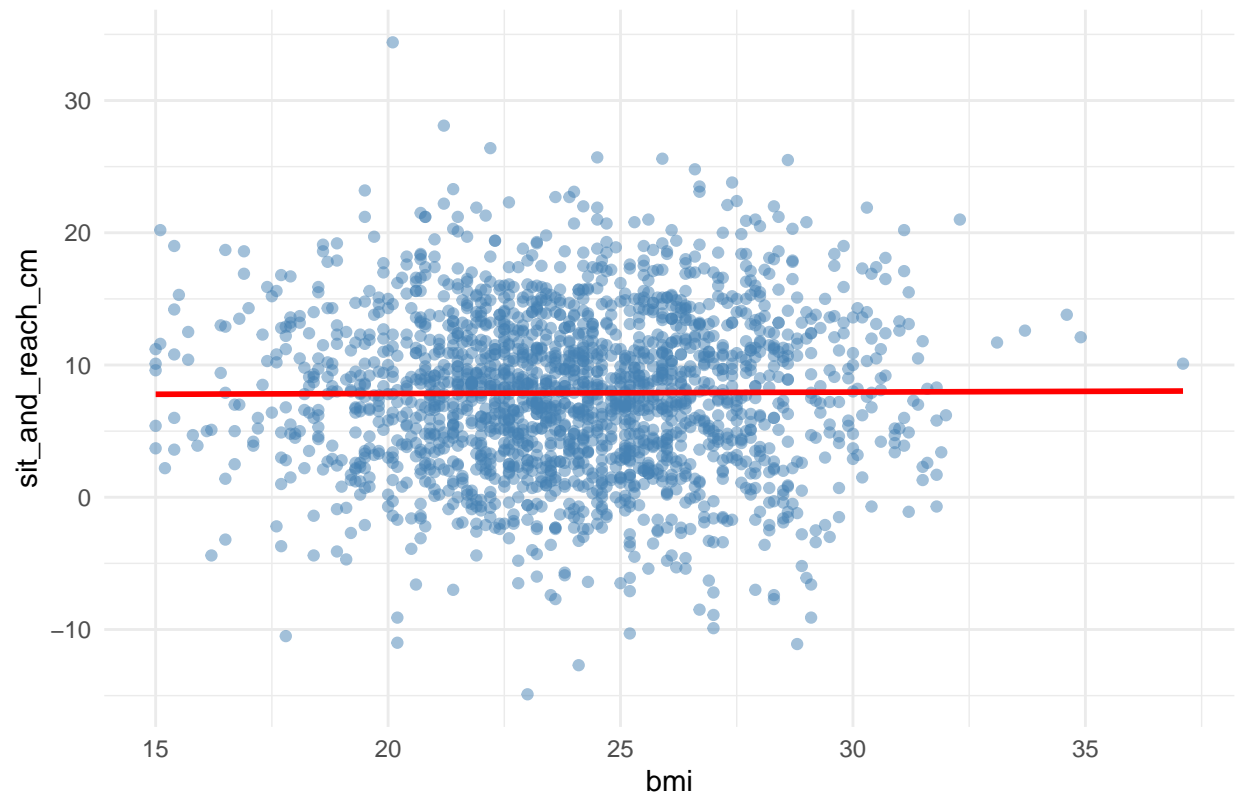
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



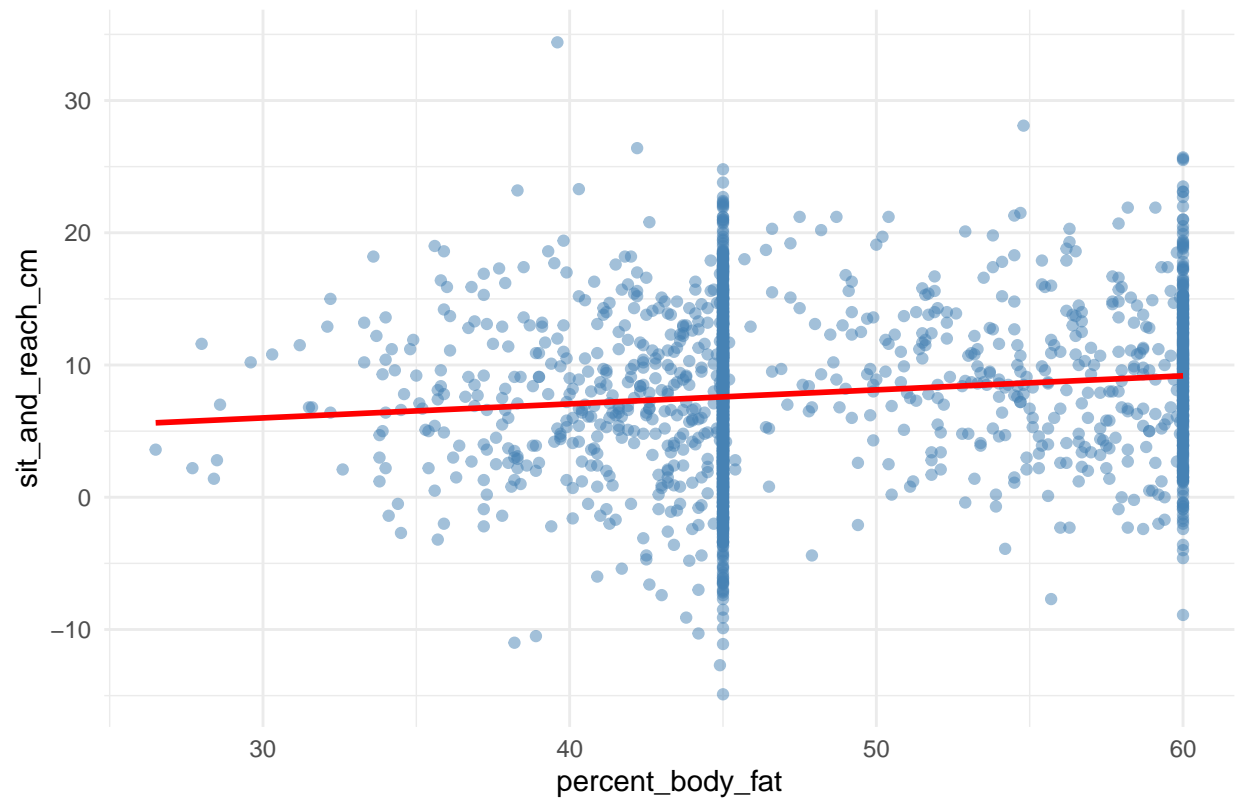
```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre sit\_and\_reach\_cm et bmi



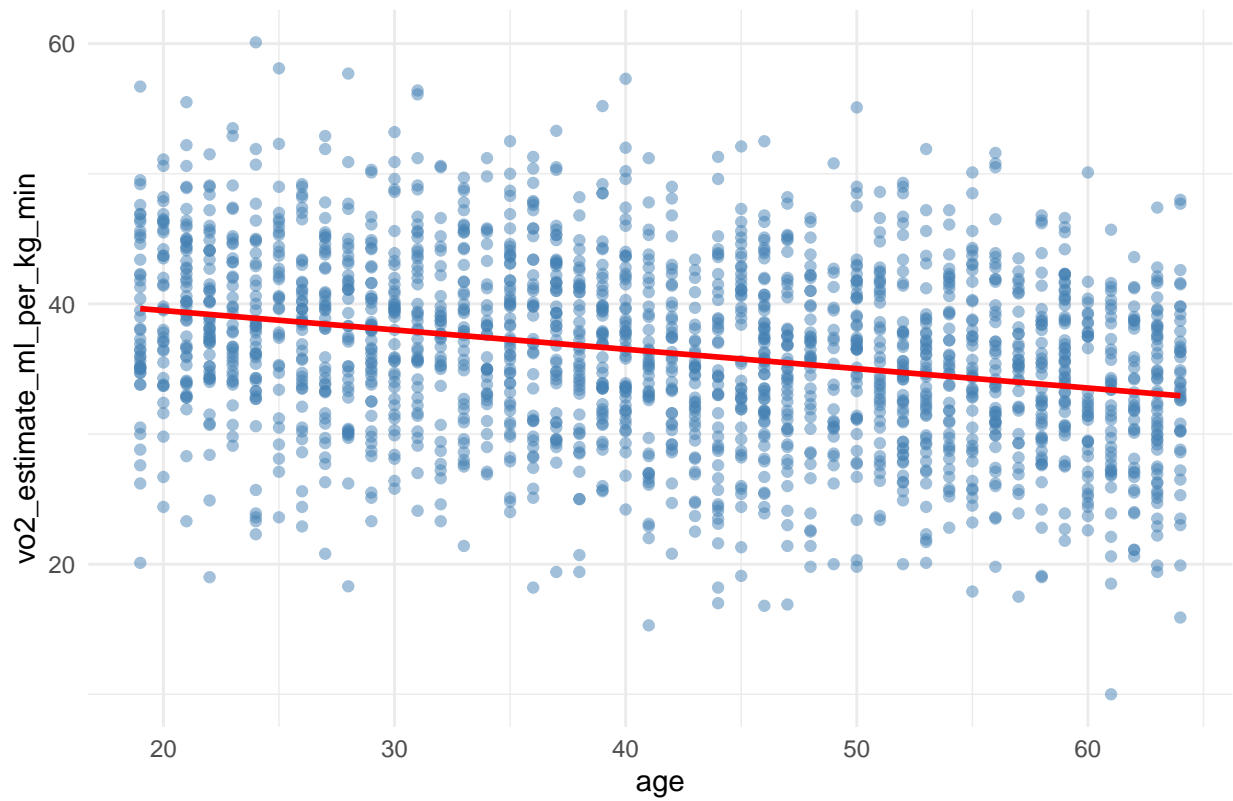
```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre sit\_and\_reach\_cm et percent\_body\_fat



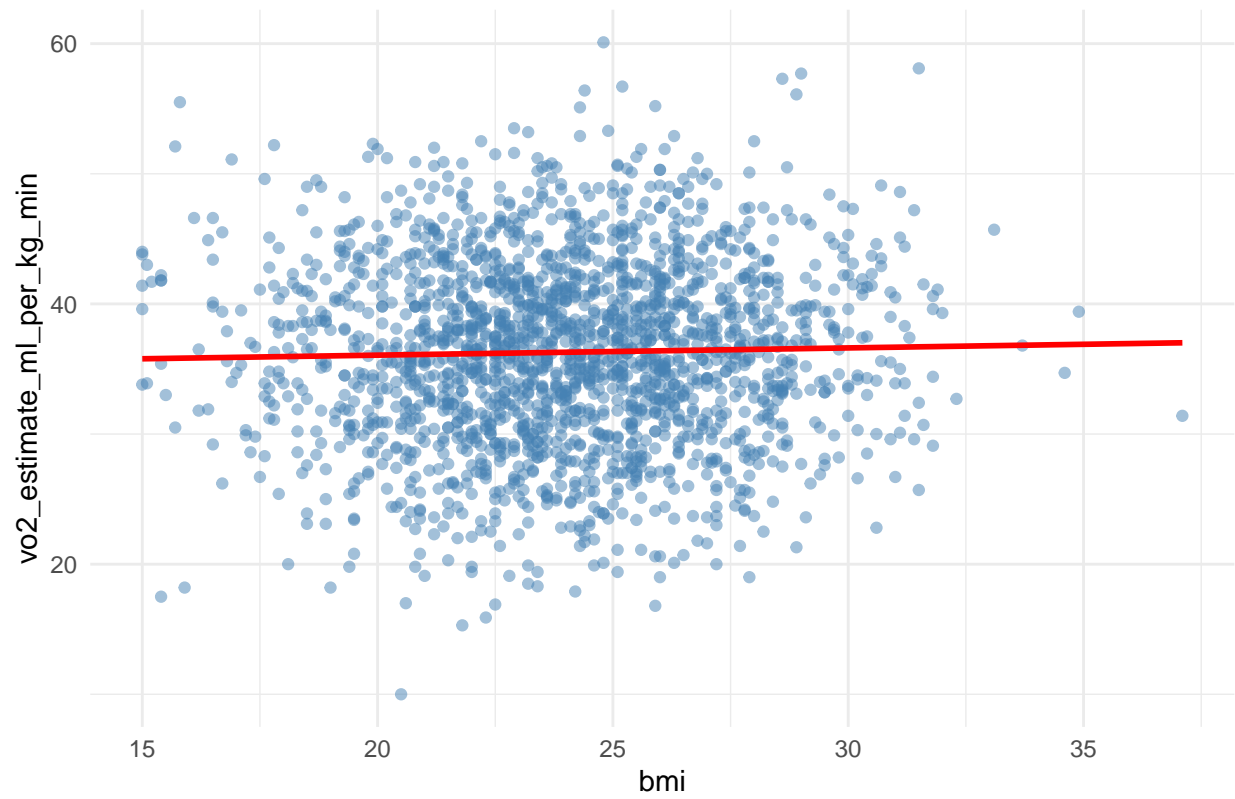
```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre vo2\_estimate\_ml\_per\_kg\_min et age



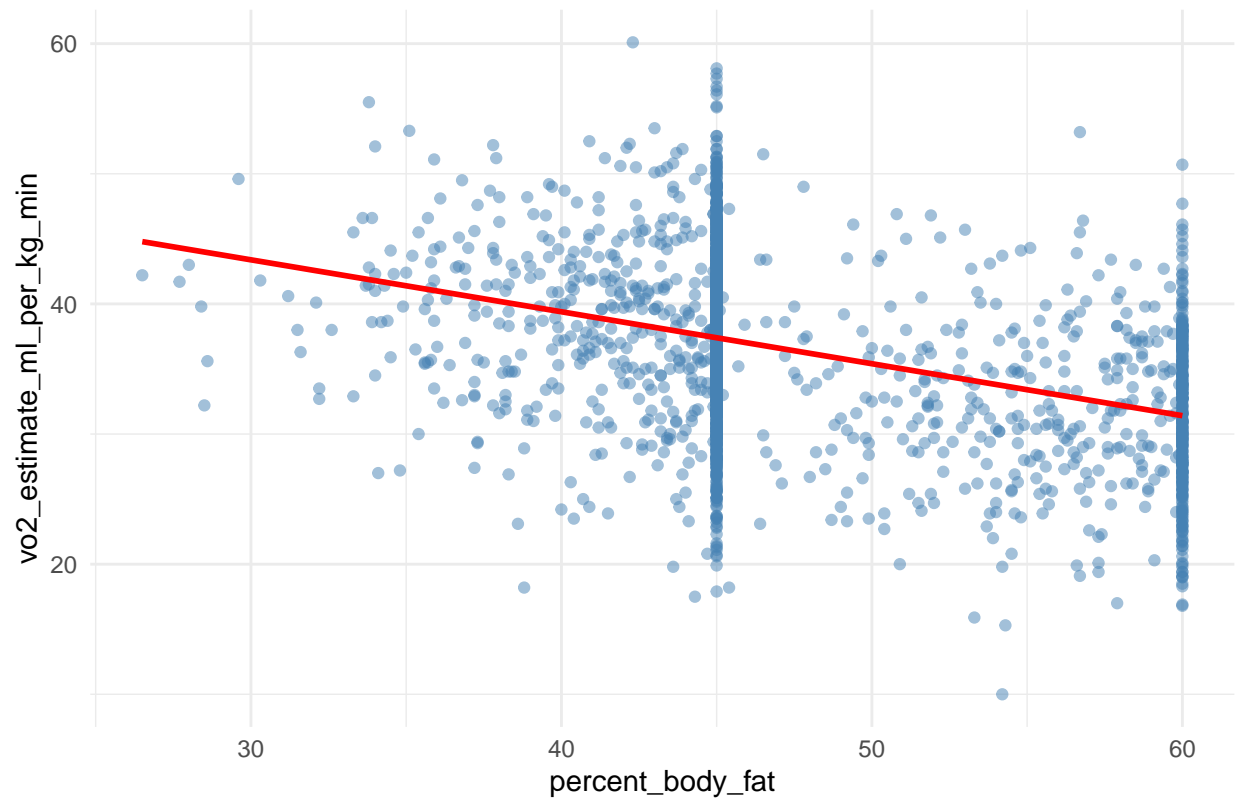
```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre vo2\_estimate\_ml\_per\_kg\_min et bmi



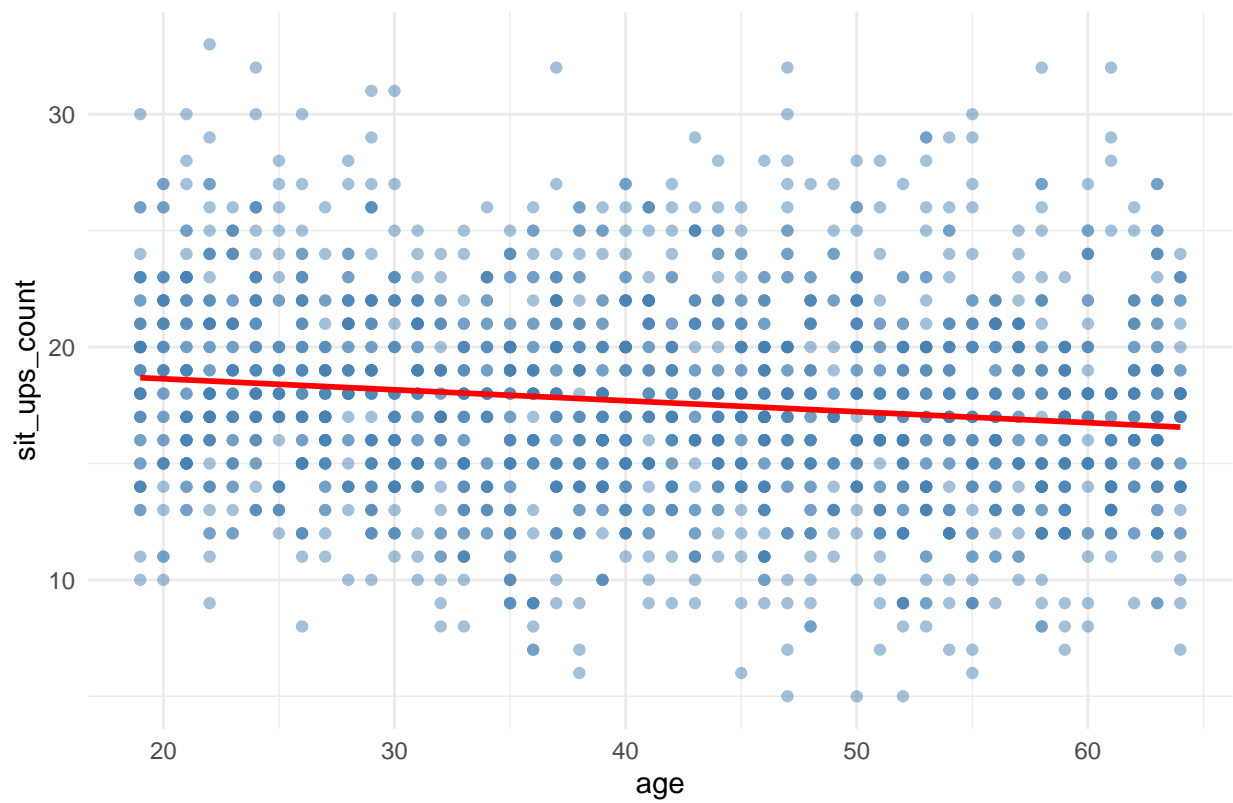
```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre vo2\_estimate\_ml\_per\_kg\_min et percent\_body\_fat



```
## `geom_smooth()` using formula = 'y ~ x'
```

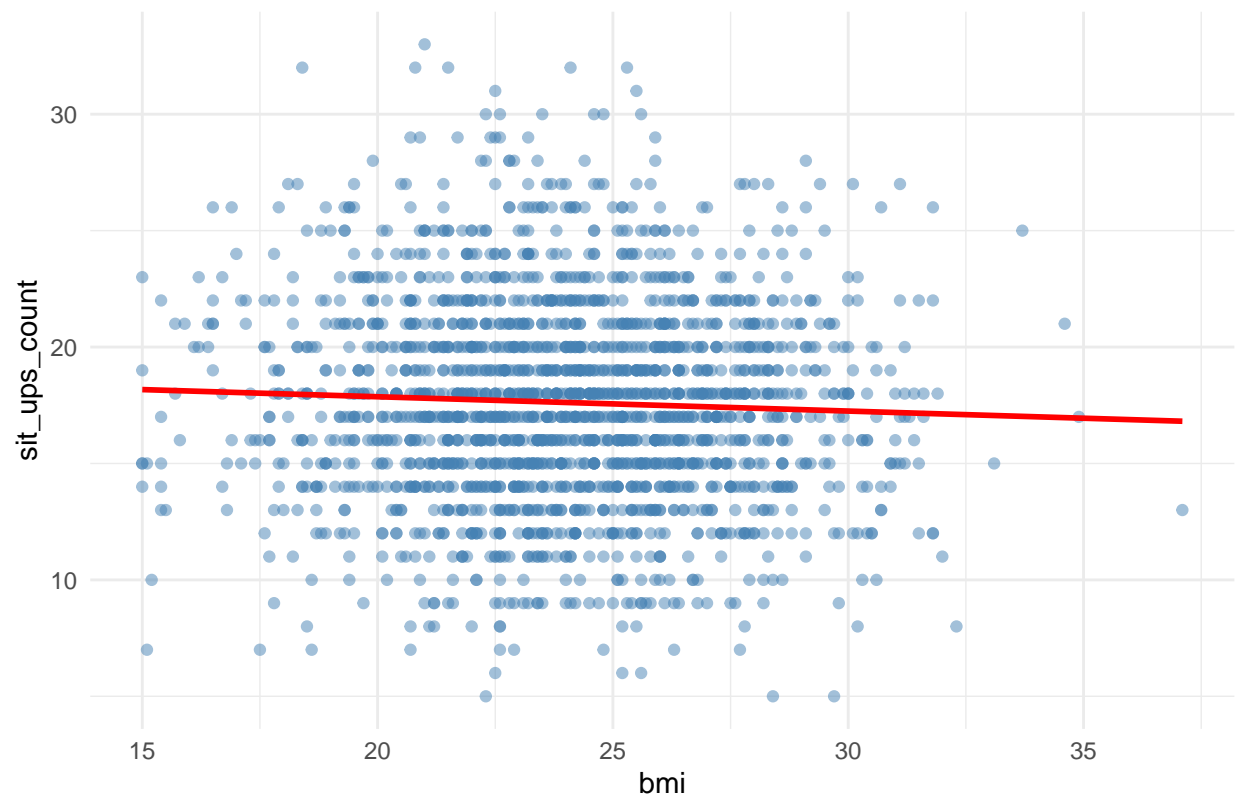
Relation entre sit\_ups\_count et age



```
## `geom_smooth()` using formula = 'y ~ x'
```



Relation entre sit\_ups\_count et bmi



```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre sit\_ups\_count et percent\_body\_fat



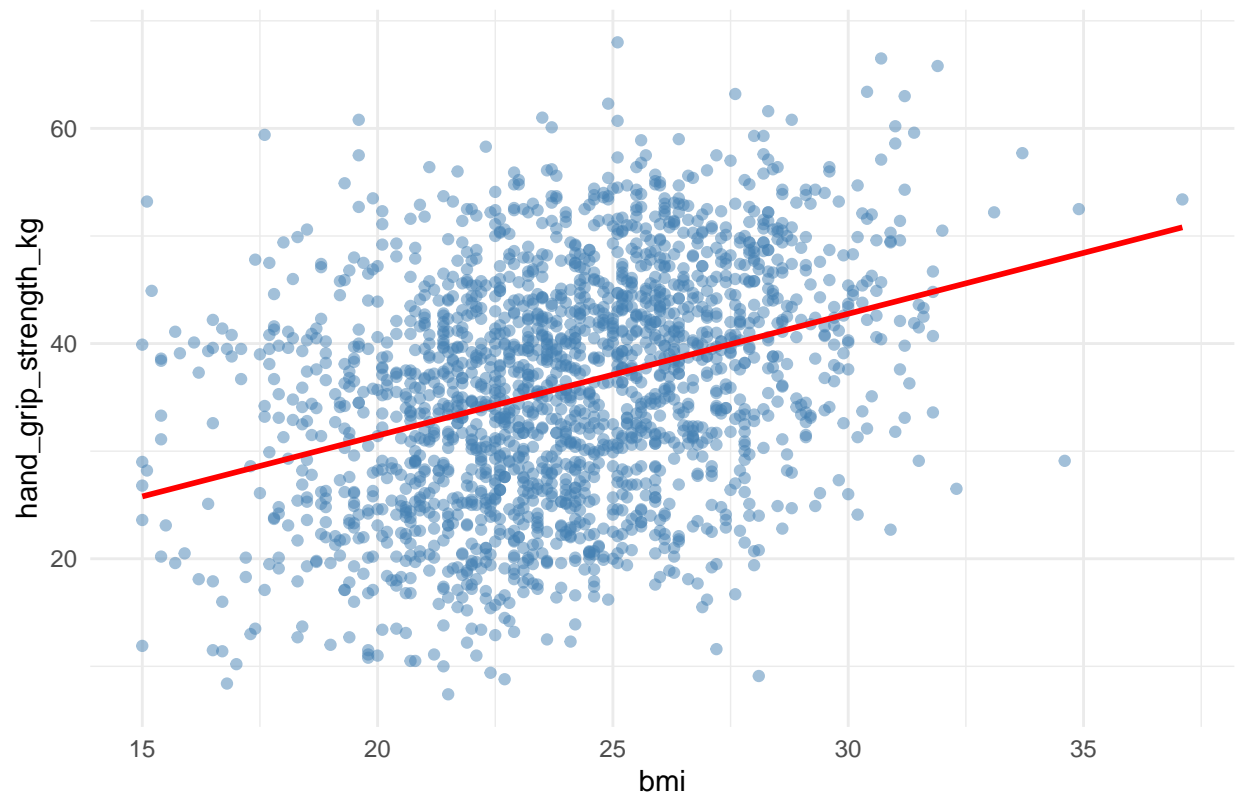
```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre hand\_grip\_strength\_kg et age



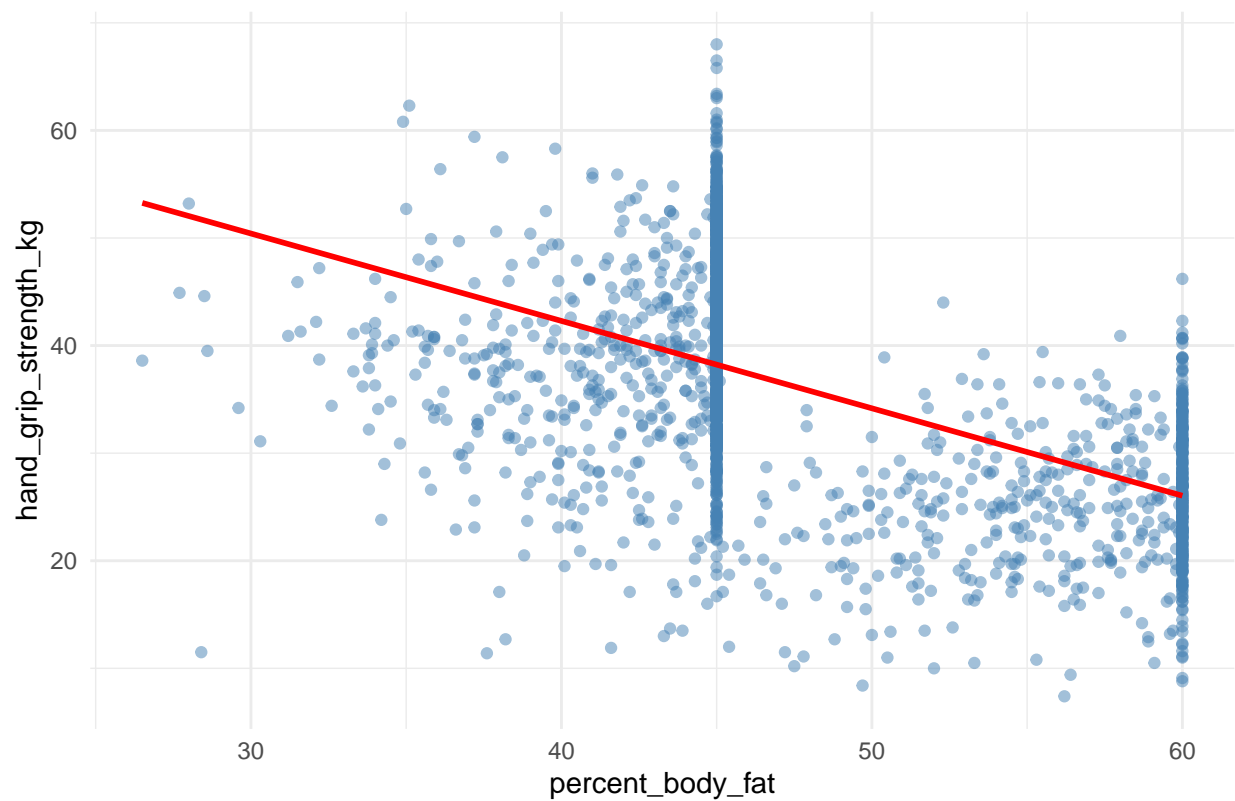
```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre hand\_grip\_strength\_kg et bmi

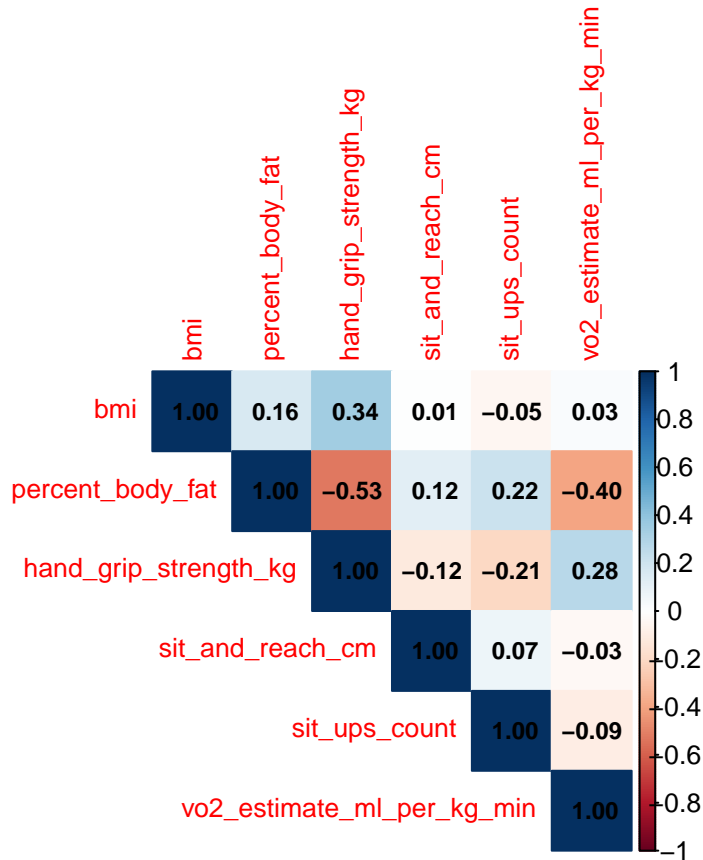


```
## `geom_smooth()` using formula = 'y ~ x'
```

Relation entre hand\_grip\_strength\_kg et percent\_body\_fat



```
cor_matrix <- cor(data[num_vars], use = "complete.obs")
corrplot(cor_matrix,
  method = "color",
  type = "upper",
  addCoef.col = "black",
  number.cex = 0.75,
  tl.cex = 0.8)
```



Après avoir tracé les diagrammes de dispersion entre les variables explicatives (*age*, *bmi*, *percent\_body\_fat*) et les variables à prédire (*hand\_grip\_strength\_kg*, *sit\_ups\_count*, *sit\_and\_reach\_cm*, *vo2\_estimate\_ml\_per\_kg\_min*), on observe les relations suivantes :

- **Hand Grip Strength (kg)** : une tendance linéaire négative est clairement visible avec le pourcentage de masse grasse, ce qui indique que l'augmentation de la masse grasse est associée à une diminution de la force de préhension.
- **VO<sub>2</sub> Estimate (ml/kg/min)** : on note également une tendance linéaire négative avec le pourcentage de masse grasse, suggérant que la capacité cardiovasculaire diminue lorsque le pourcentage de masse grasse augmente.
- **Sit-ups Count et Sit and Reach (cm)** : aucune relation linéaire évidente n'est observée avec les variables explicatives. Les points sont dispersés, indiquant que ces performances ne sont pas fortement prédictibles à partir de l'âge, du BMI ou du pourcentage de masse grasse.

La matrice de corrélation (heatmap) confirme ces observations : les coefficients les plus marqués concernent la relation négative entre le pourcentage de masse grasse et la force musculaire ainsi que la capacité cardiovasculaire. Les autres relations restent faibles ou non significatives.

Ces résultats confirment que le pourcentage de masse grasse est un prédicteur important de la force de préhension et de la VO<sub>2</sub> estimate, tandis que la flexibilité et l'endurance musculaire nécessitent d'autres variables explicatives pour être modélisées de manière fiable.

### 3 Analyse Statistique Avancée

L'Analyse Statistique Avancée vise à approfondir l'étude des données physiologiques et anthropométriques afin de mieux comprendre les relations entre les différentes variables. Cette étape permet d'évaluer la normalité des distributions, de comparer les caractéristiques physiques entre différents groupes d'individus et d'identifier des tendances ou corrélations importantes.

Elle constitue une préparation essentielle à la modélisation statistique, car elle permet de vérifier les hypothèses nécessaires à l'application des tests paramétriques et non paramétriques, ainsi qu'aux modèles de régression. En explorant plus finement les interactions et les variations au sein des données, cette analyse fournit une base solide pour interpréter les résultats et tirer des conclusions fiables sur la condition physique des individus étudiés.

#### 3.1 Vérification de la densité des variables

```
# Select numeric variables
numeric_vars <- data[, sapply(data, is.numeric)]

# Initialize lists to store results
results <- data.frame(
  Variable = character(),
  Shapiro_W = numeric(),
  Shapiro_p = numeric(),
  KS_Statistic = numeric(),
  KS_p = numeric(),
  stringsAsFactors = FALSE
)

# Loop through each numeric variable
for (var in names(numeric_vars)) {

  # Shapiro-Wilk test
  shapiro_res <- shapiro.test(numeric_vars[[var]])

  # Kolmogorov-Smirnov test against normal distribution
  ks_res <- ks.test(
    numeric_vars[[var]],
    "pnorm",
    mean(numeric_vars[[var]], na.rm = TRUE),
    sd(numeric_vars[[var]], na.rm = TRUE)
  )

  # Store results
  results <- rbind(
    results,
    data.frame(
      Variable = var,
      Shapiro_p = round(shapiro_res$p.value, 4),
      KS_p = round(ks_res$p.value, 4)
    )
  )
}
```

```
## Warning in ks.test.default(numeric_vars[[var]], "pnorm",
## mean(numeric_vars[[var]], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
## Warning in ks.test.default(numeric_vars[[var]], "pnorm",
## mean(numeric_vars[[var]], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
## Warning in ks.test.default(numeric_vars[[var]], "pnorm",
## mean(numeric_vars[[var]], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
## Warning in ks.test.default(numeric_vars[[var]], "pnorm",
## mean(numeric_vars[[var]], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
## Warning in ks.test.default(numeric_vars[[var]], "pnorm",
## mean(numeric_vars[[var]], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
## Warning in ks.test.default(numeric_vars[[var]], "pnorm",
## mean(numeric_vars[[var]], : ties should not be present for the one-sample
## Kolmogorov-Smirnov test
```

```
# Print table
results
```

```
##           Variable Shapiro_p   KS_p
## 1             age      0.0000 0.0000
## 2             bmi      0.1413 0.7044
## 3    percent_body_fat      0.0000 0.0000
## 4    hand_grip_strength_kg      0.0000 0.0181
## 5      sit_and_reach_cm      0.4410 0.9692
## 6      sit_ups_count      0.0000 0.0000
## 7 vo2_estimate_ml_per_kg_min      0.9484 0.9331
```

Pour évaluer la normalité des variables quantitatives, deux tests statistiques complémentaires ont été utilisés : le test de Shapiro–Wilk et le test de Kolmogorov–Smirnov. Ces tests permettent de déterminer si la distribution des variables s'écarte significativement d'une distribution normale.

Les résultats obtenus peuvent être résumés comme suit :

- **Age** : Les deux tests indiquent une distribution non normale ( $p < 0.05$ ), suggérant une répartition asymétrique de l'âge dans l'échantillon.
- **BMI** : Le test de Shapiro–Wilk ( $p = 0.1413$ ) et le test de Kolmogorov–Smirnov ( $p = 0.7044$ ) confirment tous deux une distribution approximativement normale.
- **Percent Body Fat** : Les deux tests révèlent une distribution non normale ( $p < 0.05$ ), indiquant une forte variabilité individuelle et une asymétrie.
- **Hand Grip Strength (kg)** : Shapiro–Wilk indique une distribution non normale ( $p < 0.05$ ), tandis que Kolmogorov–Smirnov suggère une légère déviation ( $p = 0.0181$ ), montrant une distribution légèrement asymétrique.
- **Sit and Reach (cm)** : Les deux tests confirment une distribution normale (Shapiro–Wilk :  $p = 0.4410$  ; Kolmogorov–Smirnov :  $p = 0.9692$ ).



- **Sit-ups Count** : Les deux tests montrent une distribution non normale ( $p < 0.05$ ), reflétant une variabilité élevée et la présence d'outliers.
- **VO<sub>2</sub> Estimate (ml/kg/min)** : Les deux tests indiquent une distribution normale (Shapiro–Wilk :  $p = 0.9484$  ; Kolmogorov–Smirnov :  $p = 0.9331$ ).

En comparant les résultats des deux tests, on remarque que :

- La majorité des variables présentent des résultats cohérents entre les deux tests.
- Certaines variables, comme *hand grip strength*, montrent de légères différences, Kolmogorov–Smirnov étant un peu plus tolérant aux écarts par rapport à la normalité.

Globalement, certaines variables peuvent être considérées comme approximativement normales (*BMI*, *Sit and Reach*, *VO<sub>2</sub> Estimate*), tandis que d'autres (*Age*, *Percent Body Fat*, *Hand Grip Strength*, *Sit-ups Count*) nécessitent des méthodes non paramétriques pour les analyses ultérieures ou des transformations pour satisfaire l'hypothèse de normalité.

### 3.2 Convergence vers la normalité des moyennes (Bootstrap)

```
set.seed(123)
n_boot <- 1000
boot_means <- numeric(n_boot)

for(i in 1:n_boot){
  sample_i <- sample(data$age, size = length(data$age), replace = TRUE)
  boot_means[i] <- mean(sample_i)
}

# Check normality of bootstrapped means
shapiro.test(boot_means)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  boot_means
## W = 0.99837, p-value = 0.472
```

```
library(dplyr)

# Identify numeric variables
numeric_vars <- sapply(data, is.numeric)

# Initialize results dataframe
convergence_results <- data.frame(
  Variable = names(data)[numeric_vars],
  Original_Shapiro_p = NA,
  Bootstrap_Shapiro_p = NA
)

set.seed(123) # for reproducibility
n_boot <- 1000 # number of bootstrap iterations

for(i in 1:length(convergence_results$Variable)){
```

```

var_name <- convergence_results$Variable[i]
var_data <- data[[var_name]]

# 1 Original Shapiro-Wilk test
sh <- shapiro.test(var_data)
convergence_results$Original_Shapiro_p[i] <- sh$p.value

# 2 Bootstrap means
boot_means <- replicate(n_boot, mean(sample(var_data, size = length(var_data), replace = TRUE)))

# Shapiro-Wilk test on bootstrap means
sh_boot <- shapiro.test(boot_means)
convergence_results$Bootstrap_Shapiro_p[i] <- sh_boot$p.value
}

# Add interpretation columns
convergence_results <- convergence_results %>%
  mutate(
    Original_Normality = ifelse(Original_Shapiro_p > 0.05, "Normal", "Not Normal"),
    Bootstrap_Normality = ifelse(Bootstrap_Shapiro_p > 0.05, "Converges to Normal", "Does Not Converge")
  )

# View results
print(convergence_results)

```

```

##           Variable Original_Shapiro_p Bootstrap_Shapiro_p
## 1                age      2.661509e-24      0.4719636
## 2                 bmi      1.413056e-01      0.7798773
## 3      percent_body_fat      1.621357e-42      0.6341569
## 4  hand_grip_strength_kg      8.217870e-08      0.5994569
## 5      sit_and_reach_cm      4.410473e-01      0.2965462
## 6      sit_ups_count      7.356676e-09      0.7128611
## 7 vo2_estimate_ml_per_kg_min      9.484053e-01      0.8372852
##   Original_Normality Bootstrap_Normality
## 1      Not Normal Converges to Normal
## 2       Normal Converges to Normal
## 3      Not Normal Converges to Normal
## 4      Not Normal Converges to Normal
## 5       Normal Converges to Normal
## 6      Not Normal Converges to Normal
## 7       Normal Converges to Normal

```

Le **bootstrap** est une méthode statistique qui consiste à générer de nombreux échantillons aléatoires avec remise à partir des données originales, afin d'estimer la distribution d'une statistique (par exemple, la moyenne) sans faire d'hypothèses strictes sur la distribution initiale des données. Cette approche permet d'évaluer la variabilité et la robustesse des estimations et de vérifier des propriétés comme la convergence vers une distribution normale, conformément au théorème central limite.

Pour compléter l'analyse de la distribution des variables numériques, nous avons utilisé le bootstrap pour calculer les moyennes de chaque variable sur 1000 échantillons tirés avec remise, puis testé la normalité de ces moyennes à l'aide du test de Shapiro-Wilk.

Les résultats obtenus peuvent être résumés ainsi :

- Certaines variables, telles que **l'âge** ou le **pourcentage de masse grasse**, étaient non normales selon le test de Shapiro-Wilk sur les données originales. Après bootstrap, les distributions de leurs moyennes convergent vers une distribution normale, conformément au théorème central limite.
- D'autres variables, comme le **BMI**, **sit and reach**, ou le **VO estimé**, étaient proches de la normalité initialement et leurs moyennes bootstrap confirment cette normalité.
- Dans tous les cas, les moyennes bootstrap des variables étudiées montrent une tendance à **converger vers une distribution normale**, même pour celles qui ne sont pas normales initialement.

Cette approche permet de justifier l'utilisation de méthodes statistiques paramétriques basées sur la normalité des moyennes, telles que les tests t ou la régression linéaire, en s'appuyant sur la robustesse apportée par la convergence des échantillons bootstrap.

En résumé, bien que certaines variables originales ne respectent pas strictement la normalité, l'analyse bootstrap démontre que leurs moyennes peuvent être traitées comme approximativement normales pour les analyses statistiques ultérieures.

### 3.3 Test ANOVA

Afin d'examiner les différences entre les groupes pour plusieurs variables quantitatives, nous avons procédé en plusieurs étapes : calcul des moyennes par groupe, vérification de la normalité, puis application des tests statistiques appropriés (T-tests et ANOVA).

#### 3.3.1 Création des groupes

```
tapply(data$hand_grip_strength_kg, data$sex, mean, na.rm = TRUE)
```

```
##           F           M
## 25.14574 41.10384
```

```
tapply(data$vo2_estimate_ml_per_kg_min, cut(data$age, breaks=c(20,30,40,50,60,70)), mean, na.rm = TRUE)
```

```
## (20,30] (30,40] (40,50] (50,60] (60,70]
## 38.58000 37.68804 35.12238 34.61742 32.52249
```

```
tapply(data$percent_body_fat, cut(data$bmi, breaks=c(10,20,25,30,38)), mean, na.rm = TRUE)
```

```
## (10,20] (20,25] (25,30] (30,38]
## 42.27202 48.82615 48.11202 46.69355
```

```
tapply(data$sit_ups_count, cut(data$age, breaks=c(20,30,40,50,60)), mean, na.rm = TRUE)
```

```
## (20,30] (30,40] (40,50] (50,60]
## 18.70805 17.28894 17.69464 16.76244
```

- **Force de préhension** selon le sexe : Les hommes présentent une moyenne de 41.10 kg et les femmes de 25.15 kg, indiquant une différence notable entre les sexes.

- **VO estimé** selon les tranches d'âge : On observe une diminution progressive du VO estimé avec l'âge, de 38.58 ml/kg/min pour les 20-30 ans à 32.52 ml/kg/min pour les 60-70 ans.
- **Pourcentage de masse grasse** selon les catégories de BMI : Les moyennes augmentent légèrement avec l'IMC, reflétant la corrélation attendue entre masse grasse et BMI.
- **Nombre de sit-ups** selon l'âge : On note une diminution progressive des performances physiques avec l'âge.

Pour certaines analyses, nous avons créé des groupes catégoriels :

- **sex\_int** : 0 pour les femmes, 1 pour les hommes.
- **age\_group** : tranches d'âge de 10 ans de 20 à 70 ans.
- **bmi\_group** : catégories de BMI.
- **fat\_group** : catégories de pourcentage de masse grasse.

```
# Création des groupes
data$sex_int <- ifelse(data$sex == "M", 1, 0)
age_group <- cut(data$age, breaks=c(20,30,40,50,60,70), include.lowest = TRUE)
bmi_group <- cut(data$bmi, breaks=c(10,20,30,40), include.lowest = TRUE)
fat_group <- cut(data$percent_body_fat, breaks=c(25,35,40,50,60), include.lowest = TRUE)
```

### 3.3.2 ANOVA et tests de Bartlett

```
# ANOVA + Bartlett
bartlett.test(vo2_estimate_ml_per_kg_min ~ age_group, data = data)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: vo2_estimate_ml_per_kg_min by age_group
## Bartlett's K-squared = 2.2292, df = 4, p-value = 0.6937
```

```
anova_vo2 <- aov(vo2_estimate_ml_per_kg_min ~ age_group, data = data)
summary(anova_vo2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## age_group      4   7795  1948.7    42.58 <2e-16 ***
## Residuals    1951   89278    45.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 44 observations deleted due to missingness
```

```
# ANOVA + Bartlett
bartlett.test(sit_and_reach_cm ~ age_group, data = data)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: sit_and_reach_cm by age_group
## Bartlett's K-squared = 3.324, df = 4, p-value = 0.5051
```

```
anova_sit <- aov(sit_and_reach_cm ~ age_group, data = data)
summary(anova_sit)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## age_group      4     531   132.65    3.599 0.00625 **
## Residuals    1951   71900    36.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 44 observations deleted due to missingness
```

```
# ANOVA + Bartlett
bartlett.test(vo2_estimate_ml_per_kg_min ~ sex_int, data = data)
```

```
##
## Bartlett test of homogeneity of variances
##
## data:  vo2_estimate_ml_per_kg_min by sex_int
## Bartlett's K-squared = 0.021874, df = 1, p-value = 0.8824
```

```
anova_sex <- aov(vo2_estimate_ml_per_kg_min ~ sex_int, data = data)
summary(anova_sex)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## sex_int        1   17372   17372    422.5 <2e-16 ***
## Residuals    1998   82152     41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ANOVA + Bartlett
bartlett.test(sit_and_reach_cm ~ sex_int, data = data)
```

```
##
## Bartlett test of homogeneity of variances
##
## data:  sit_and_reach_cm by sex_int
## Bartlett's K-squared = 2.3237, df = 1, p-value = 0.1274
```

```
anova_sit_sexe <- aov(sit_and_reach_cm ~ sex_int, data = data)
summary(anova_sit_sexe)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## sex_int        1   1782   1782.4    49.64 2.54e-12 ***
## Residuals    1998   71746    35.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# ANOVA + Bartlett
bartlett.test(vo2_estimate_ml_per_kg_min ~ bmi_group, data = data)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: vo2_estimate_ml_per_kg_min by bmi_group
## Bartlett's K-squared = 0.64425, df = 2, p-value = 0.7246
```

```
anova_bmi <- aov(vo2_estimate_ml_per_kg_min ~ bmi_group, data = data)
summary(anova_bmi)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## bmi_group      2     101    50.29    1.01  0.364
## Residuals    1997    99423    49.79
```

```
# ANOVA + Bartlett
bartlett.test(sit_and_reach_cm ~ bmi_group, data = data)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: sit_and_reach_cm by bmi_group
## Bartlett's K-squared = 2.0589, df = 2, p-value = 0.3572
```

```
anova_sit_bmi <- aov(sit_and_reach_cm ~ bmi_group, data = data)
summary(anova_sit_bmi)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## bmi_group      2     139    69.61    1.894  0.151
## Residuals    1997    73390    36.75
```

```
# ANOVA + Bartlett
bartlett.test(sit_and_reach_cm ~ fat_group, data = data)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: sit_and_reach_cm by fat_group
## Bartlett's K-squared = 5.8513, df = 3, p-value = 0.1191
```

```
anova_sit_fat <- aov(sit_and_reach_cm ~ fat_group, data = data)
summary(anova_sit_fat)
```

```
##              Df Sum Sq Mean Sq F value    Pr(>F)
## fat_group      3     1262    420.7    11.62 1.51e-07 ***
## Residuals    1996    72267     36.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
bartlett.test(vo2_estimate_ml_per_kg_min ~ fat_group, data = data)
```

```
##
## Bartlett test of homogeneity of variances
##
## data: vo2_estimate_ml_per_kg_min by fat_group
## Bartlett's K-squared = 3.247, df = 3, p-value = 0.3551

anova_fat <- aov(vo2_estimate_ml_per_kg_min ~ fat_group, data = data)
summary(anova_fat)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## fat_group      3  15875     5292   126.3 <2e-16 ***
## Residuals    1996  83649       42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(anova_fat)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = vo2_estimate_ml_per_kg_min ~ fat_group, data = data)
##
## $fat_group
##              diff          lwr          upr          p adj
## (35,40]-[25,35] -1.115789 -4.233690  2.0021115  0.7940896
## (40,50]-[25,35] -2.208892 -4.948800  0.5310151  0.1623590
## (50,60]-[25,35] -8.260247 -11.049599 -5.4708961  0.0000000
## (40,50]-[35,40] -1.093103 -2.719891  0.5336853  0.3094790
## (50,60]-[35,40] -7.144458 -8.853208 -5.4357077  0.0000000
## (50,60]-[40,50] -6.051355 -6.891363 -5.2113468  0.0000000
```

Pour chaque variable quantitative, nous avons vérifié l'homogénéité des variances avec le test de Bartlett avant d'appliquer l'ANOVA.

- **VO estimé selon l'âge** : Le test de Bartlett montre que les variances sont homogènes ( $p = 0.694$ ). L'ANOVA révèle une différence significative entre les groupes d'âge ( $F = 42.58$ ,  $p < 2e - 16$ ).
- **Sit and reach selon l'âge** : Variances homogènes ( $p = 0.505$ ). L'ANOVA indique une différence significative ( $F = 3.599$ ,  $p = 0.006$ ).
- **VO estimé selon le sexe** : Variances homogènes ( $p = 0.882$ ). L'ANOVA montre une différence très significative entre hommes et femmes ( $F = 422.5$ ,  $p < 2e - 16$ ).
- **Sit and reach selon le sexe** : Variances homogènes ( $p = 0.127$ ). L'ANOVA révèle également une différence significative ( $F = 49.64$ ,  $p < 0.001$ ).
- **VO estimé et sit and reach selon le BMI** : Variances homogènes ( $p > 0.35$ ). L'ANOVA n'indique pas de différences significatives pour VO selon le BMI ( $p = 0.364$ ) ni pour sit and reach ( $p = 0.151$ ).
- **Sit and reach selon le pourcentage de masse grasse** : Variances homogènes ( $p = 0.119$ ). L'ANOVA montre une différence significative entre les groupes ( $F = 11.62$ ,  $p < 0.001$ ).
- **VO estimé selon le pourcentage de masse grasse** : Variances homogènes ( $p = 0.355$ ). L'ANOVA révèle des différences très significatives ( $F = 126.3$ ,  $p < 2e - 16$ ), et le test de comparaisons multiples de Tukey montre que les différences concernent surtout le groupe 50-60% par rapport aux autres.

### 3.3.3 Conclusion

Les tests ANOVA et les tests de Bartlett confirment que certaines variables présentent des différences significatives entre les groupes selon le sexe, l'âge ou la composition corporelle, tandis que d'autres variables ne montrent pas de différences notables selon le BMI. Les comparaisons multiples de Tukey permettent d'identifier plus précisément les groupes qui diffèrent significativement.

## 3.4 Comparaison selon le sexe : analyse des moyennes et des variances

L'objectif principal de cette section est d'étudier l'influence du **sexe** sur les différentes variables anthropométriques et de condition physique. Pour cela, les données ont été divisées en deux groupes (femmes et hommes), puis une analyse statistique en deux étapes a été réalisée : une analyse descriptive afin de comparer les niveaux moyens et la dispersion des variables, suivie de tests inférentiels portant sur l'égalité des variances (test de Fisher) et la comparaison des moyennes (test t de Student). Cette approche permet d'identifier de manière rigoureuse les différences significatives entre les deux sexes.

```
tapply(data$bmi, data$sex, summary)
```

```
## $F
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  15.00  21.20   23.00   23.08  25.00   32.30
##
## $M
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  15.00  22.20   24.40   24.41  26.60   37.10
```

```
tapply(data$sit_and_reach_cm, data$sex, summary)
```

```
## $F
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.900   5.200   9.300   9.255  13.200   28.100
##
## $M
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -14.900   3.000   7.200   7.236  11.250   34.400
```

```
tapply(data$vo2_estimate_ml_per_kg_min, data$sex, summary)
```

```
## $F
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00  27.70   31.90   32.01  36.40   53.20
##
## $M
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.90  34.00   38.40   38.32  42.50   60.10
```

```
tapply(data$age, data$sex, summary)
```

```
## $F
##   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```



```
##    19.00    30.00    42.00    41.82    53.00    64.00
##
## $M
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      19.00    30.00    41.00    41.41    53.00    64.00
```

```
tapply(data$percent_body_fat, data$sex, summary)
```

```
## $F
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      34.30    54.00    59.10    56.43    60.00    60.00
##
## $M
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      26.50    44.20    45.00    43.69    45.00    45.00
```

```
tapply(data$hand_grip_strength_kg, data$sex, summary)
```

```
## $F
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##       7.40    20.40    25.00    25.15    29.70    46.20
##
## $M
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##      11.5     35.6     41.0     41.1     46.7     68.0
```

```
tapply(data$sit_ups_count, data$sex, summary)
```

```
## $F
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##       7.00    16.00    19.00    19.35    22.00    33.00
##
## $M
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##       5.0     14.0     17.0     16.8     20.0     32.0
```

### 3.4.1 Indice de masse corporelle (BMI)

L'analyse descriptive montre que le BMI moyen est légèrement plus élevé chez les hommes (24.41) que chez les femmes (23.08). Le test t de Student indique que cette différence de moyenne est statistiquement significative ( $p < 2.2 \times 10^{-16}$ ). Le test de Fisher met en évidence une différence significative entre les variances ( $p < 0.001$ ), suggérant une dispersion différente du BMI selon le sexe. Ces résultats indiquent que le sexe influence à la fois le niveau moyen et la variabilité du BMI.

```
t.test(bmi ~ sex, data = data, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data:  bmi by sex
```

```
## t = -8.8119, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
## -1.634350 -1.039305
## sample estimates:
## mean in group F mean in group M
##      23.07690      24.41373
```

```
var.test(bmi ~ sex, data = data)
```

```
##
## F test to compare two variances
##
## data:  bmi by sex
## F = 0.77237, num df = 644, denom df = 1354, p-value = 0.0001806
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6774075 0.8834407
## sample estimates:
## ratio of variances
##      0.7723682
```

### 3.4.2 Flexibilité (Sit and Reach)

Les femmes présentent une flexibilité moyenne supérieure (9.26 cm) à celle des hommes (7.24 cm). Le test t confirme que cette différence est statistiquement significative ( $p < 0.001$ ). En revanche, le test de Fisher ne met pas en évidence de différence significative entre les variances ( $p = 0.129$ ), indiquant une dispersion comparable de la flexibilité entre les deux sexes.

```
t.test(sit_and_reach_cm ~ sex, data = data, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data:  sit_and_reach_cm by sex
## t = 7.0453, df = 1998, p-value = 2.536e-12
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
##  1.457437 2.581807
## sample estimates:
## mean in group F mean in group M
##      9.255194      7.235572
```

```
var.test(sit_and_reach_cm ~ sex, data = data)
```

```
##
## F test to compare two variances
##
## data:  sit_and_reach_cm by sex
## F = 0.90133, num df = 644, denom df = 1354, p-value = 0.1294
## alternative hypothesis: true ratio of variances is not equal to 1
```

```
## 95 percent confidence interval:
##  0.7905174 1.0309529
## sample estimates:
## ratio of variances
##           0.9013341
```

### 3.4.3 Capacité cardiorespiratoire (VO<sub>2</sub> estimé)

La moyenne du VO<sub>2</sub> estimé est nettement plus élevée chez les hommes (38.32 ml/kg/min) que chez les femmes (32.01 ml/kg/min). Le test t de Student montre une différence très significative entre les deux groupes ( $p < 2.2 \times 10^{-16}$ ). Le test de Fisher indique que les variances sont homogènes ( $p = 0.889$ ), ce qui confirme que la différence observée porte principalement sur le niveau moyen et non sur la dispersion.

```
t.test(vo2_estimate_ml_per_kg_min ~ sex, data = data, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data: vo2_estimate_ml_per_kg_min by sex
## t = -20.555, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
## -6.906645 -5.703502
## sample estimates:
## mean in group F mean in group M
##      32.01116      38.31624
```

```
var.test(vo2_estimate_ml_per_kg_min ~ sex, data = data)
```

```
##
## F test to compare two variances
##
## data: vo2_estimate_ml_per_kg_min by sex
## F = 0.99003, num df = 644, denom df = 1354, p-value = 0.8888
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8683075 1.1324028
## sample estimates:
## ratio of variances
##           0.990029
```

### 3.4.4 Âge

Les moyennes d'âge sont très proches entre les femmes (41.82 ans) et les hommes (41.41 ans). Le test t ne révèle aucune différence significative entre les deux groupes ( $p = 0.515$ ). De plus, le test de Fisher confirme l'égalité des variances ( $p = 0.576$ ). Ainsi, l'âge ne constitue pas un facteur différenciant entre les sexes dans cet échantillon.

```
# Age
var.test(age ~ sex, data = data)
```

```
##
## F test to compare two variances
##
## data: age by sex
## F = 1.038, num df = 644, denom df = 1354, p-value = 0.5755
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.910381 1.187273
## sample estimates:
## ratio of variances
## 1.038001
```

```
t.test(age ~ sex, data = data, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data: age by sex
## t = 0.65075, df = 1998, p-value = 0.5153
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
## -0.8282222 1.6508201
## sample estimates:
## mean in group F mean in group M
## 41.82016 41.40886
```

### 3.4.5 Pourcentage de masse grasse

Le pourcentage de masse grasse est significativement plus élevé chez les femmes (56.43%) que chez les hommes (43.69%). Le test t de Student confirme une différence extrêmement significative des moyennes ( $p < 2.2 \times 10^{-16}$ ). Le test de Fisher indique également une différence marquée des variances ( $p < 2.2 \times 10^{-16}$ ), suggérant une plus grande hétérogénéité de la masse grasse selon le sexe.

```
# Percent body fat
var.test(percent_body_fat ~ sex, data = data)
```

```
##
## F test to compare two variances
##
## data: percent_body_fat by sex
## F = 2.9038, num df = 644, denom df = 1354, p-value < 2.2e-16
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 2.546780 3.321382
## sample estimates:
## ratio of variances
## 2.903794
```

```
t.test(percent_body_fat ~ sex, data = data, var.equal = TRUE)
```

```
##
```

```
## Two Sample t-test
##
## data: percent_body_fat by sex
## t = 72.862, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
## 12.39715 13.08297
## sample estimates:
## mean in group F mean in group M
## 56.42744 43.68738
```

### 3.4.6 Force de préhension (Hand Grip Strength)

La force de préhension moyenne est nettement plus élevée chez les hommes (41.10 kg) que chez les femmes (25.15 kg). Le test t révèle une différence très significative ( $p < 2.2 \times 10^{-16}$ ). Le test de Fisher indique une différence significative entre les variances ( $p < 0.001$ ), montrant que la dispersion de la force de préhension varie également selon le sexe.

```
# Hand grip strength
var.test(hand_grip_strength_kg ~ sex, data = data)
```

```
##
## F test to compare two variances
##
## data: hand_grip_strength_kg by sex
## F = 0.6384, num df = 644, denom df = 1354, p-value = 1.307e-10
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 0.5599122 0.7302093
## sample estimates:
## ratio of variances
## 0.6384021
```

```
t.test(hand_grip_strength_kg ~ sex, data = data, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data: hand_grip_strength_kg by sex
## t = -43.195, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
## -16.68264 -15.23356
## sample estimates:
## mean in group F mean in group M
## 25.14574 41.10384
```

### 3.4.7 Endurance musculaire (Nombre de sit-ups)

Les femmes réalisent en moyenne davantage de sit-ups (19.35) que les hommes (16.80). Le test t de Student montre que cette différence est statistiquement significative ( $p < 2.2 \times 10^{-16}$ ). Le test de Fisher met

en évidence une différence significative des variances ( $p = 0.0014$ ), indiquant une variabilité différente de l'endurance musculaire entre les sexes.

```
# Sit-ups count
var.test(sit_ups_count ~ sex, data = data)
```

```
##
## F test to compare two variances
##
## data: sit_ups_count by sex
## F = 1.2381, num df = 644, denom df = 1354, p-value = 0.001376
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
## 1.085859 1.416122
## sample estimates:
## ratio of variances
## 1.238077
```

```
t.test(sit_ups_count ~ sex, data = data, var.equal = TRUE)
```

```
##
## Two Sample t-test
##
## data: sit_ups_count by sex
## t = 12.59, df = 1998, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group F and group M is not equal to 0
## 95 percent confidence interval:
## 2.150427 2.943998
## sample estimates:
## mean in group F mean in group M
## 19.34574 16.79852
```

### 3.4.8 Conclusion

L'analyse des moyennes et des variances selon le sexe met en évidence des différences significatives pour la majorité des variables de condition physique et de composition corporelle. À l'exception de l'âge, toutes les variables étudiées présentent des différences significatives de moyenne entre les hommes et les femmes, confirmant le rôle important du sexe dans l'évaluation de la condition physique et des indicateurs de santé.

## 3.5 Analyse comparative selon les groupes d'âge

Cette sous-section a pour objectif d'analyser l'influence de l'âge sur plusieurs indicateurs anthropométriques et de performance physique. Pour cela, la population a été divisée en deux groupes indépendants : les individus âgés de 35 ans ou moins (*Under\_35*) et ceux âgés de plus de 35 ans (*Over\_35*). Dans un premier temps, des statistiques descriptives ont été calculées afin de comparer les distributions des variables dans chaque groupe. Ensuite, des tests de comparaison de variances (test de Fisher) ont été appliqués afin de vérifier l'homogénéité des dispersions entre les deux groupes, condition préalable à l'utilisation de tests paramétriques de comparaison des moyennes.

```
# Create age group variable
data$age_group_35 <- ifelse(data$age <= 35, "Under_35", "Over_35")
tapply(data$bmi, data$age_group_35, summary)
```

```
## $Over_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      15.0   21.8   24.0   24.0   26.2   37.1
##
## $Under_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      15.00  21.90  23.90  23.95  26.10  34.90
```

```
tapply(data$percent_body_fat, data$age_group_35, summary)
```

```
## $Over_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      26.50  45.00  45.00  47.89  54.02  60.00
##
## $Under_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      27.70  45.00  45.00  47.62  51.85  60.00
```

```
tapply(data$hand_grip_strength_kg, data$age_group_35, summary)
```

```
## $Over_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       7.4   28.0   37.2   36.2   44.2   68.0
##
## $Under_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      8.40  27.70  35.50  35.53  43.00  66.50
```

```
tapply(data$sit_ups_count, data$age_group_35, summary)
```

```
## $Over_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       5.00  14.00  17.00  17.25  20.00  32.00
##
## $Under_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       8.00  15.00  18.00  18.27  21.00  33.00
```

```
tapply(data$sit_and_reach_cm, data$age_group_35, summary)
```

```
## $Over_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -14.900   3.600   7.500   7.611  11.600  25.700
##
## $Under_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -12.700   4.100   8.400   8.373  12.800  34.400
```

```
tapply(data$vo2_estimate_ml_per_kg_min, data$age_group_35, summary)
```

```
## $Over_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    10.00  30.38   35.10   35.02  39.80   57.30
##
## $Under_35
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.30  34.20   38.40   38.51  43.33   60.10
```

```
# BMI
```

```
var.test(bmi ~ age_group_35, data = data)
```

```
##
## F test to compare two variances
##
## data:  bmi by age_group_35
## F = 0.98894, num df = 1275, denom df = 723, p-value = 0.8611
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8681316 1.1239436
## sample estimates:
## ratio of variances
##           0.9889446
```

```
# Percent body fat
```

```
var.test(percent_body_fat ~ age_group_35, data = data)
```

```
##
## F test to compare two variances
##
## data:  percent_body_fat by age_group_35
## F = 1.0702, num df = 1275, denom df = 723, p-value = 0.3071
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9394955 1.2163363
## sample estimates:
## ratio of variances
##           1.07024
```

```
# Hand grip strength
```

```
var.test(hand_grip_strength_kg ~ age_group_35, data = data)
```

```
##
## F test to compare two variances
##
## data:  hand_grip_strength_kg by age_group_35
## F = 1.0445, num df = 1275, denom df = 723, p-value = 0.5133
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9169036 1.1870873
```



```
## sample estimates:
## ratio of variances
##          1.044504
```

```
# Sit-ups
var.test(sit_ups_count ~ age_group_35, data = data)
```

```
##
## F test to compare two variances
##
## data:  sit_ups_count by age_group_35
## F = 1.1143, num df = 1275, denom df = 723, p-value = 0.1033
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9781855 1.2664271
## sample estimates:
## ratio of variances
##          1.114314
```

```
# Sit-and-reach
var.test(sit_and_reach_cm ~ age_group_35, data = data)
```

```
##
## F test to compare two variances
##
## data:  sit_and_reach_cm by age_group_35
## F = 0.95231, num df = 1275, denom df = 723, p-value = 0.4535
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.8359735 1.0823096
## sample estimates:
## ratio of variances
##          0.9523113
```

```
# VO2
var.test(vo2_estimate_ml_per_kg_min ~ age_group_35, data = data)
```

```
##
## F test to compare two variances
##
## data:  vo2_estimate_ml_per_kg_min by age_group_35
## F = 1.0474, num df = 1275, denom df = 723, p-value = 0.4865
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.9194649 1.1904034
## sample estimates:
## ratio of variances
##          1.047422
```

**3.5.0.1 Indice de masse corporelle (BMI)** Les moyennes du BMI sont très proches entre les deux groupes (23.95 pour les individus de 35 ans ou moins contre 24.00 pour les individus de plus de 35 ans). Le test de Fisher ne met pas en évidence de différence significative entre les variances ( $p = 0.8611$ ), indiquant une dispersion similaire du BMI dans les deux groupes d'âge.

**3.5.0.2 Pourcentage de masse grasse** Le pourcentage moyen de masse grasse est comparable entre les deux groupes (47.62 pour les moins de 35 ans et 47.89 pour les plus de 35 ans). Le test de comparaison des variances n'est pas significatif ( $p = 0.3071$ ), ce qui suggère une variabilité homogène de la masse grasse indépendamment de l'âge.

**3.5.0.3 Force de préhension** La force de préhension moyenne est légèrement plus élevée chez les individus de plus de 35 ans (36.20 kg) que chez les plus jeunes (35.53 kg). Toutefois, le test de Fisher indique une égalité des variances entre les deux groupes ( $p = 0.5133$ ), montrant une dispersion comparable des performances musculaires.

**3.5.0.4 Nombre de sit-ups** Les individus de 35 ans ou moins présentent en moyenne un nombre de sit-ups légèrement supérieur (18.27) par rapport aux individus de plus de 35 ans (17.25). Le test de Fisher ne révèle pas de différence significative entre les variances ( $p = 0.1033$ ), indiquant une variabilité similaire de l'endurance musculaire entre les deux groupes.

**3.5.0.5 Souplesse (sit-and-reach)** La souplesse moyenne est légèrement plus élevée chez les individus de moins de 35 ans (8.37 cm) que chez les individus plus âgés (7.61 cm). Le test de Fisher confirme l'homogénéité des variances entre les deux groupes ( $p = 0.4535$ ).

**3.5.0.6 VO<sub>2</sub> estimé** Le VO<sub>2</sub> estimé est plus élevé chez les individus de 35 ans ou moins (38.51 ml/kg/min) comparativement aux individus de plus de 35 ans (35.02 ml/kg/min), traduisant une diminution de la capacité cardiorespiratoire avec l'âge. Le test de comparaison des variances ne montre toutefois aucune différence significative ( $p = 0.4865$ ).

**3.5.0.7 Conclusion partielle** Les tests de Fisher indiquent une homogénéité des variances pour l'ensemble des variables étudiées entre les deux groupes d'âge. Les différences observées concernent principalement les moyennes, notamment pour les variables de performance physique, ce qui permet l'utilisation de tests paramétriques pour approfondir la comparaison des groupes.

## 3.6 Test de Proportions

Afin d'évaluer si les proportions observées pour certaines caractéristiques de l'échantillon diffèrent significativement de 50%, nous avons réalisé des **tests de proportion à un échantillon** avec correction de continuité. Les variables analysées sont : le sexe, l'âge et l'IMC.

```
data$age_group <- ifelse(data$age <= 35, "Under_35", "Over_35")
data$bmi_group <- ifelse(data$bmi < 25, "Under_25", "Over_25")
sex_table <- table(data$sex)
prop.test(sex_table)
```

```
##
## 1-sample proportions test with continuity correction
##
## data:  sex_table, null probability 0.5
## X-squared = 251.34, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3021259 0.3435629
## sample estimates:
```

```
##      p
## 0.3225
```

```
age_table <- table(data$age_group)
prop.test(age_table)
```

```
##
## 1-sample proportions test with continuity correction
##
## data:  age_table, null probability 0.5
## X-squared = 151.8, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.6164394 0.6590252
## sample estimates:
##      p
## 0.638
```

```
bmi_table <- table(data$bmi_group)
prop.test(bmi_table)
```

```
##
## 1-sample proportions test with continuity correction
##
## data:  bmi_table, null probability 0.5
## X-squared = 96.361, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.3686071 0.4118196
## sample estimates:
##      p
## 0.39
```

### 3.6.1 Sexe (Homme vs Femme)

Le test de proportion pour le sexe donne les résultats suivants :

- Statistique du test :  $X^2 = 251.34$ ,  $df = 1$
- Valeur p :  $p < 2.2 \times 10^{-16}$
- Intervalle de confiance à 95% pour la proportion de femmes :  $[0.302, 0.344]$
- Proportion observée :  $p = 0.323$

**Interprétation :** La proportion de femmes dans l'échantillon (32,3%) est significativement différente de 50% ( $p < 0.001$ ). On peut conclure que l'échantillon contient une majorité d'hommes.

### 3.6.2 Âge (Moins de 35 ans vs 35 ans et plus)

Le test de proportion pour l'âge montre :

- Statistique du test :  $X^2 = 151.8$ ,  $df = 1$
- Valeur p :  $p < 2.2 \times 10^{-16}$
- Intervalle de confiance à 95% pour la proportion de participants de 35 ans et plus : [0.616, 0.659]
- Proportion observée :  $p = 0.638$

**Interprétation :** La proportion de participants âgés de 35 ans et plus (63,8%) est significativement différente de 50%, indiquant que la majorité de l'échantillon est constituée d'adultes de 35 ans et plus.

### 3.6.3 IMC (Inférieur à 25 vs Supérieur ou égal à 25)

Le test de proportion pour l'IMC fournit :

- Statistique du test :  $X^2 = 96.361$ ,  $df = 1$
- Valeur p :  $p < 2.2 \times 10^{-16}$
- Intervalle de confiance à 95% pour la proportion d'individus ayant un IMC supérieur ou égal à 25 : [0.369, 0.412]
- Proportion observée :  $p = 0.390$

**Interprétation :** La proportion d'individus avec un IMC supérieur ou égal à 25 (39,0%) est significativement différente de 50%, ce qui indique que la majorité de l'échantillon a un IMC inférieur à 25.

## 3.7 Modèle pour la force de préhension (Hand Grip Strength)

**Modèle :** Régression linéaire multiple avec les variables explicatives : `sex`, `age`, `bmi`, `percent_body_fat`.

**Étapes :**

1. Régression initiale avec toutes les variables.
2. Sélection des variables par méthode stepwise (critère AIC).
3. Suppression des outliers (résidus étudiants > 2).
4. Estimation finale du modèle.

```
# Step 5: Stepwise regression
model_grip <- lm(hand_grip_strength_kg ~ sex + age + bmi + percent_body_fat, data = data)
step_grip <- step(model_grip, direction = "both")
```

```
## Start: AIC=8001.34
## hand_grip_strength_kg ~ sex + age + bmi + percent_body_fat
##
##               Df Sum of Sq    RSS    AIC
## - percent_body_fat  1      63.0 108788 8000.5
```

```
## <none> 108725 8001.3
## - age 1 202.1 108927 8003.1
## - bmi 1 6834.1 115559 8121.3
## - sex 1 14080.9 122805 8242.9
##
## Step: AIC=8000.5
## hand_grip_strength_kg ~ sex + age + bmi
##
## Df Sum of Sq RSS AIC
## <none> 108788 8000.5
## + percent_body_fat 1 63 108725 8001.3
## - age 1 197 108985 8002.1
## - bmi 1 10119 118906 8176.4
## - sex 1 94852 203639 9252.4
```

```
# Step 6: Remove outliers
```

```
library(MASS)
sre_grip <- studres(step_grip)
keep_grip <- which(abs(sre_grip) < 2)
clean_grip <- data[keep_grip, ]
```

```
# Step 7: Evaluate
```

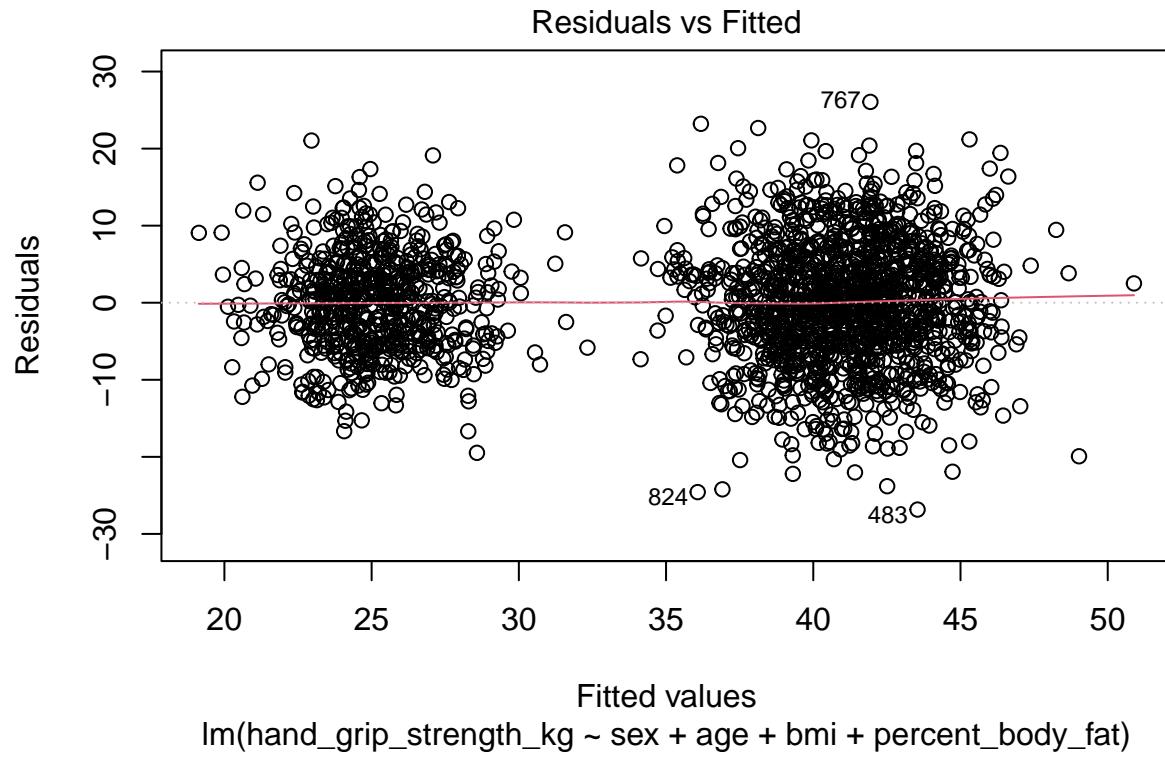
```
final_grip <- lm(hand_grip_strength_kg ~ sex + age + bmi + percent_body_fat, data = clean_grip)
summary(final_grip)
```

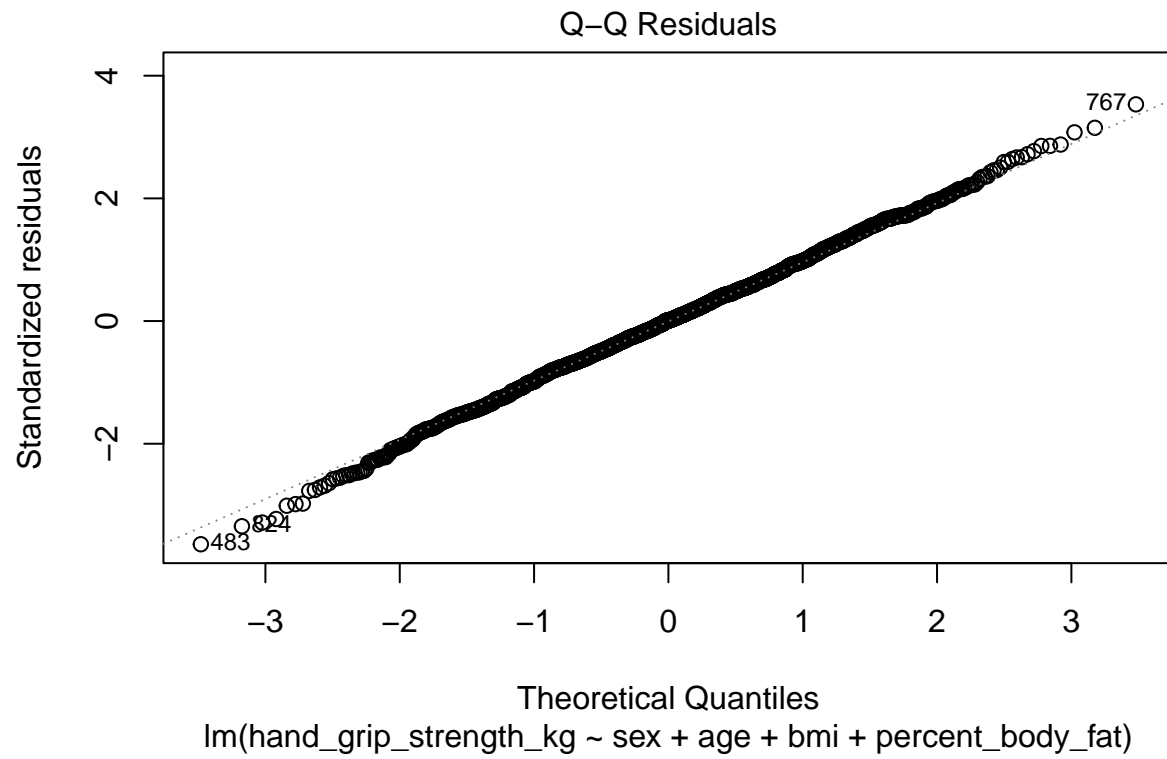
```
##
## Call:
## lm(formula = hand_grip_strength_kg ~ sex + age + bmi + percent_body_fat,
## data = clean_grip)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.9085  -4.6648  -0.0039   4.4648  14.5613
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   6.9100757  2.3654001   2.921  0.00353 **
## sexM          15.2203713  0.7800149  19.513 < 2e-16 ***
## age           0.0346255  0.0110865   3.123  0.00182 **
## bmi           0.7270529  0.0601564  12.086 < 2e-16 ***
## percent_body_fat -0.0008463  0.0520276  -0.016  0.98702
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.408 on 1901 degrees of freedom
## Multiple R-squared:  0.6085, Adjusted R-squared:  0.6076
## F-statistic: 738.6 on 4 and 1901 DF, p-value: < 2.2e-16
```

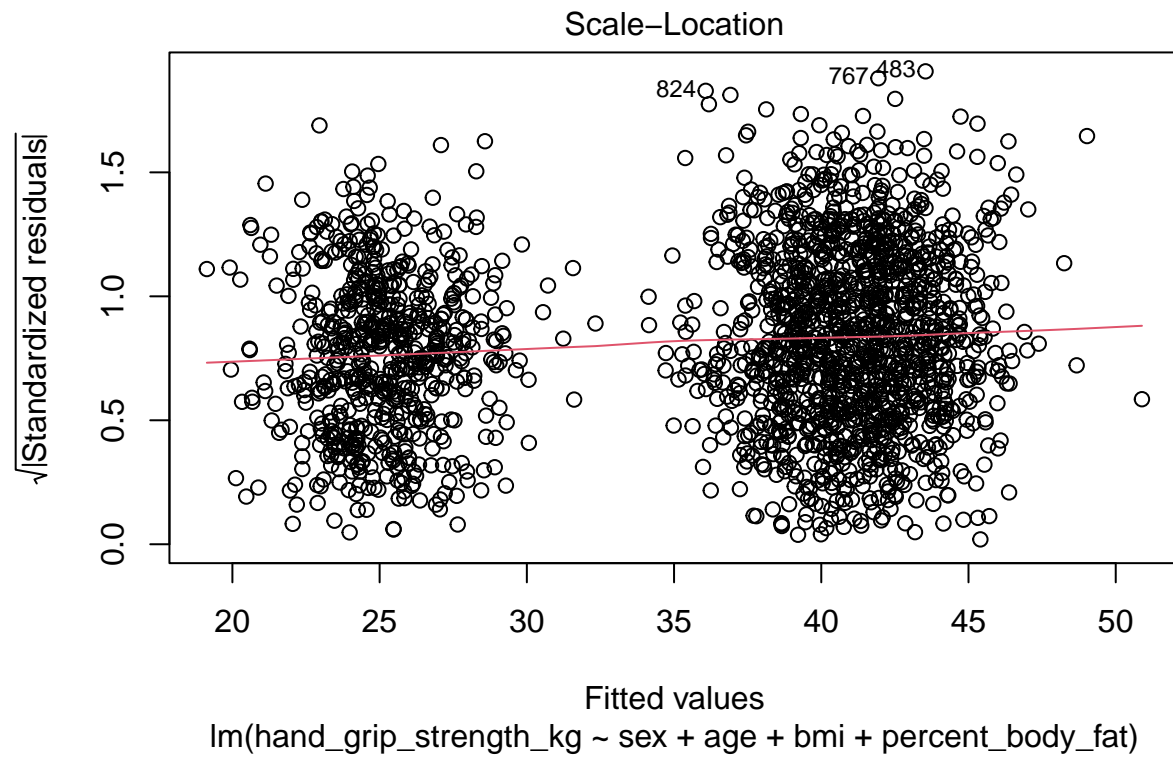
```
SEE_grip <- sqrt(mean(final_grip$residuals^2))
SEE_grip
```

```
## [1] 6.399141
```

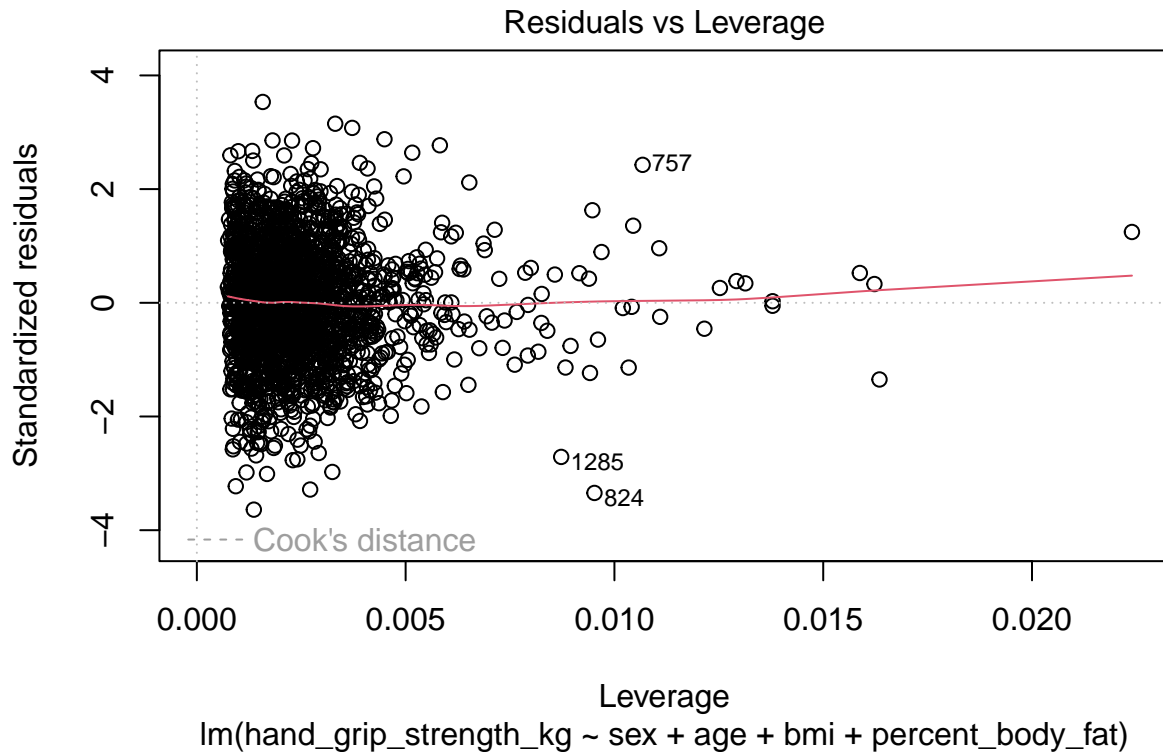
```
plot(model_grip)
```











Résultats principaux :

Variable	Estimate	p-value	Interprétation
Intercept	6.91	0.0035 **	Valeur de référence si toutes les variables = 0
sex (M)	15.22	$< 2e - 16$ ***	Les hommes ont une force plus élevée de 15 kg en moyenne
age	0.035	0.0018 **	L'âge est faiblement mais positivement associé à la force
bmi	0.727	$< 2e - 16$ ***	IMC positif corrélé à une force plus élevée
percent_body_fat	-0.0008	0.987	Pas significatif

Table 2: Coefficients du modèle de la force de préhension

Indicateurs du modèle :

- $R^2$  ajusté : 0.608 (le modèle explique environ 61% de la variance)
- SEE (Erreur standard de l'estimation) : 6.4 kg

**Analyse :** Sexe, âge et IMC sont significatifs, tandis que la masse grasse n'est pas. Le modèle est relativement performant.

### 3.8 Modèle pour la $VO_2$ estimée ( $VO_2$ max)

**Modèle :** Régression linéaire multiple avec sex, age, bmi, percent\_body\_fat.

**Étapes :** Même approche que pour la force de préhension (stepwise + suppression des outliers).

```

model_vo2 <- lm(vo2_estimate_ml_per_kg_min ~ sex + age + bmi + percent_body_fat, data = data)
step_vo2 <- step(model_vo2, direction = "both")

## Start: AIC=7240.17
## vo2_estimate_ml_per_kg_min ~ sex + age + bmi + percent_body_fat
##
##           Df Sum of Sq  RSS    AIC
## - bmi           1      11.4 74320 7238.5
## <none>                74309 7240.2
## - percent_body_fat 1      196.1 74505 7243.4
## - sex             1     1667.4 75976 7282.6
## - age             1     7269.5 81578 7424.8
##
## Step: AIC=7238.48
## vo2_estimate_ml_per_kg_min ~ sex + age + percent_body_fat
##
##           Df Sum of Sq  RSS    AIC
## <none>                74320 7238.5
## + bmi           1      11.4 74309 7240.2
## - percent_body_fat 1      438.1 74758 7248.2
## - sex             1     2541.4 76862 7303.7
## - age             1     7270.6 81591 7423.1

sre_vo2 <- studres(step_vo2)
keep_vo2 <- which(abs(sre_vo2) < 2)
clean_vo2 <- data[keep_vo2, ]

final_vo2 <- lm(vo2_estimate_ml_per_kg_min ~ sex + age + bmi + percent_body_fat, data = clean_vo2)
summary(final_vo2)

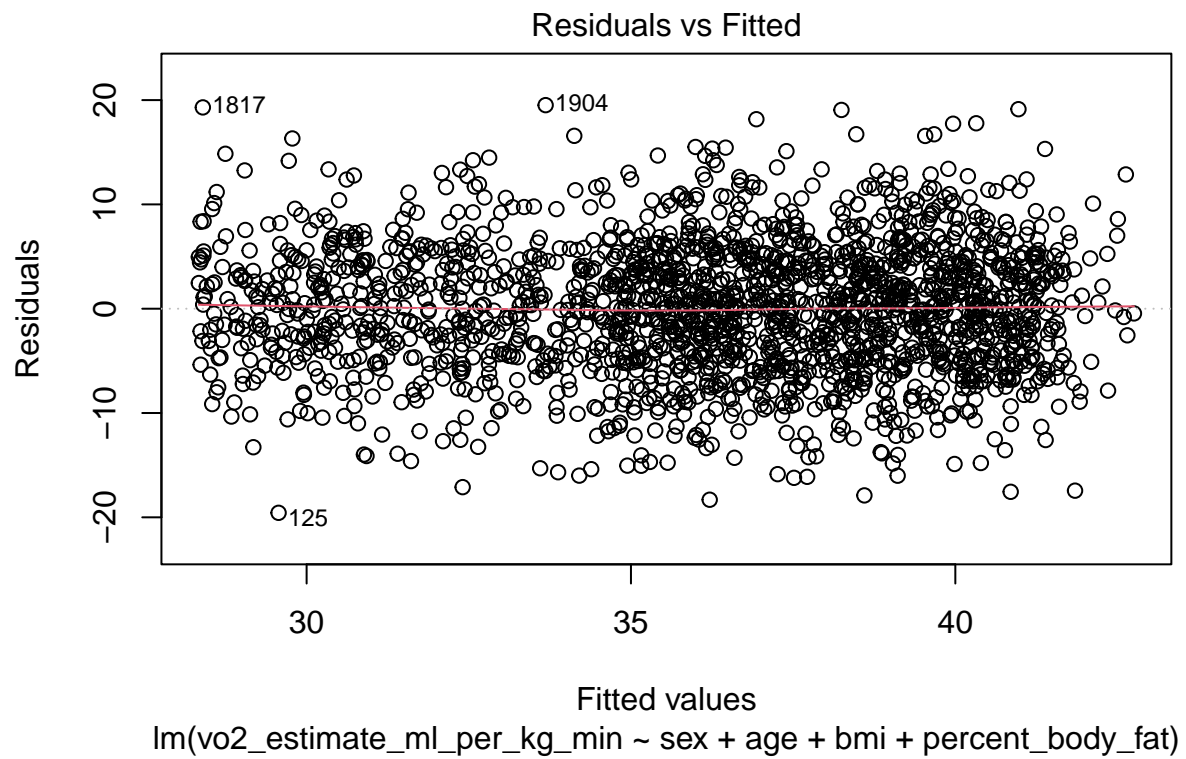
##
## Call:
## lm(formula = vo2_estimate_ml_per_kg_min ~ sex + age + bmi + percent_body_fat,
##     data = clean_vo2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.1901  -4.0854   0.0551   3.9784  12.1608
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.724239   1.997917  22.886 < 2e-16 ***
## sexM           4.738568   0.659663   7.183 9.73e-13 ***
## age          -0.146650   0.009219 -15.907 < 2e-16 ***
## bmi          -0.029765   0.049796  -0.598  0.55009
## percent_body_fat -0.122472  0.043859  -2.792  0.00528 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.315 on 1900 degrees of freedom
## Multiple R-squared:  0.3125, Adjusted R-squared:  0.3111
## F-statistic: 215.9 on 4 and 1900 DF, p-value: < 2.2e-16

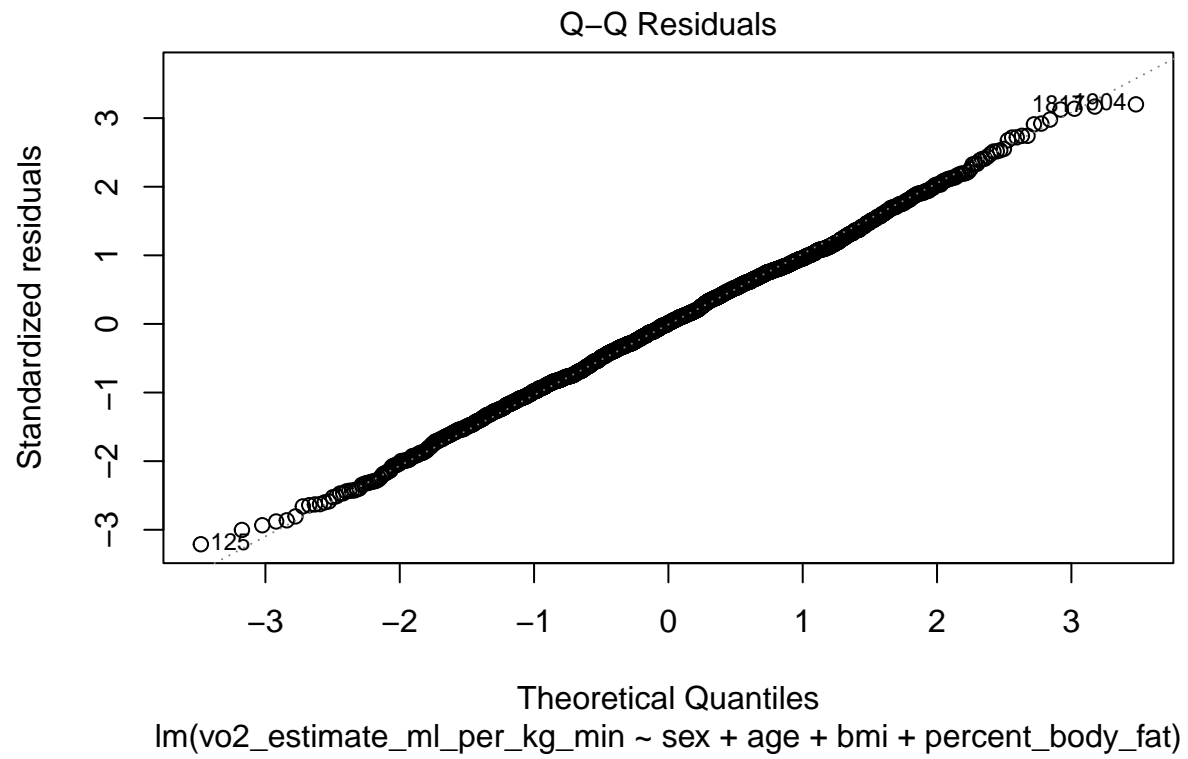
```

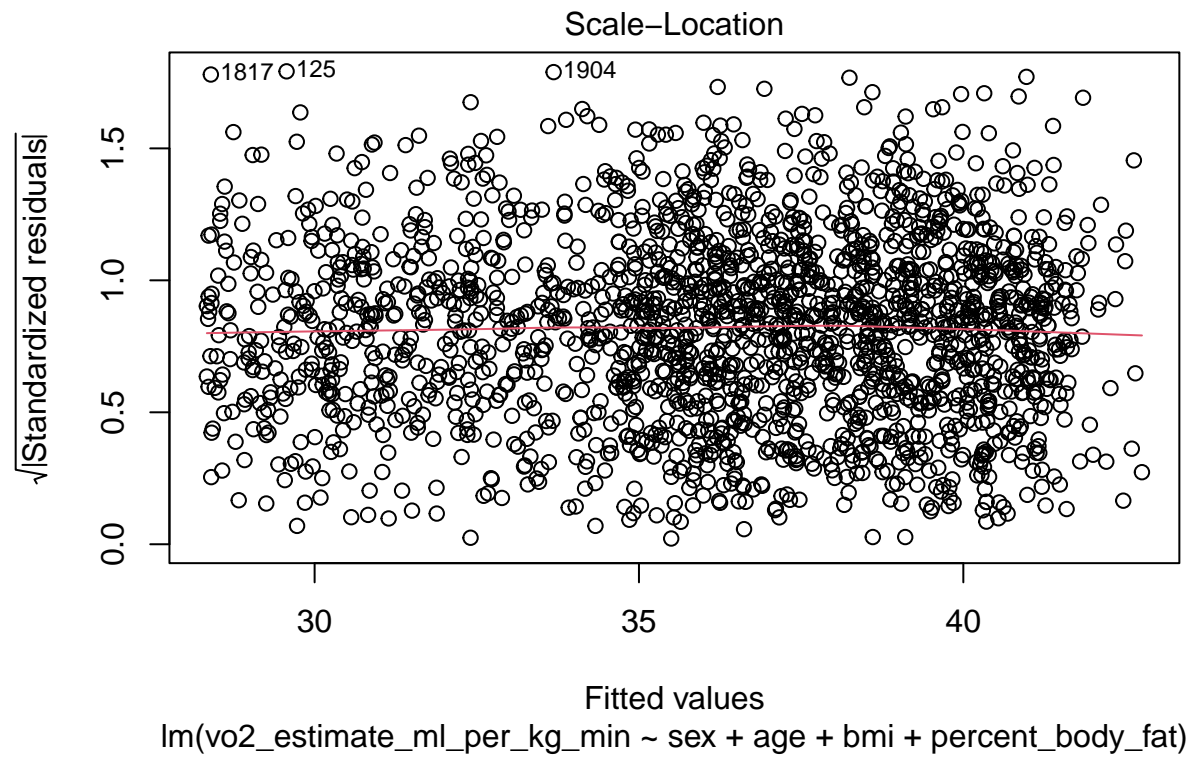
```
SEE_vo2 <- sqrt(mean(final_vo2$residuals^2))  
SEE_vo2
```

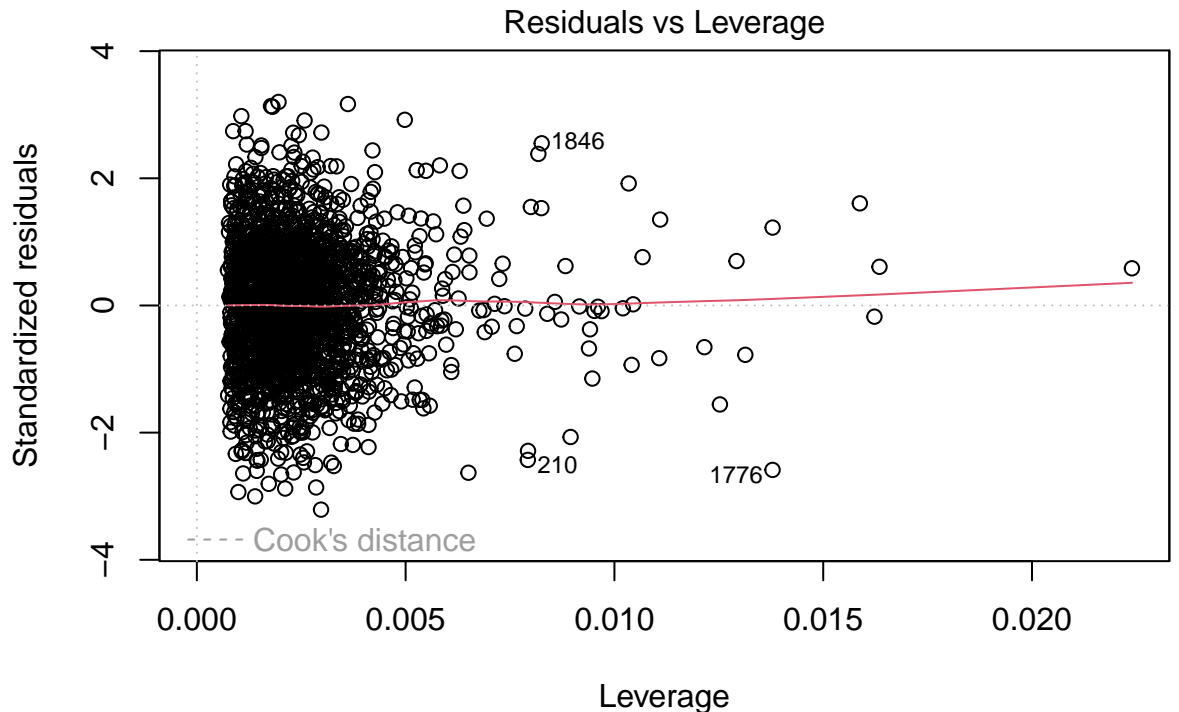
```
## [1] 5.308369
```

```
plot(model_vo2)
```









Résultats principaux :

Variable	Estimate	p-value	Interprétation
Intercept	45.72	$< 2e - 16$ ***	VO <sub>2</sub> estimée de référence
sex (M)	4.74	$9.73e-13$ **	Les hommes ont une VO <sub>2</sub> plus élevée
age	-0.147	$< 2e - 16$ ***	La VO <sub>2</sub> diminue avec l'âge
bmi	-0.03	0.55	Non significatif
percent_body_fat	-0.122	0.005 **	La masse grasse diminue légèrement la VO <sub>2</sub>

Table 3: Coefficients du modèle VO<sub>2</sub> estimée

Indicateurs du modèle :

- $R^2$  ajusté : 0.311 (le modèle explique 31% de la variance)
- SEE : 5.31 ml/kg/min

**Analyse :** Sexe, âge et masse grasse sont significatifs, mais l'IMC ne l'est pas. Ce modèle est moins précis que celui de la force de préhension.

### 3.9 Synthèse et Comparaison des Modèles

- Force de préhension : modèle plus précis ( $R^2 \approx 0.61$ )
- VO<sub>2</sub> : modèle moins précis ( $R^2 \approx 0.31$ )
- Variables clés : sexe et âge influencent fortement les deux mesures

- Implication : estimation possible des indicateurs physiques à partir de données simples, avec des précisions variables.

## 4 Conclusion Générale

Ce projet a permis de réaliser une analyse statistique complète des données physiologiques et anthropométriques afin de mieux comprendre les relations entre les caractéristiques individuelles et les performances physiques.

L'analyse exploratoire a mis en évidence les principales tendances des variables étudiées, ainsi que la présence de valeurs aberrantes susceptibles d'influencer certaines mesures. Les tests de normalité (Shapiro–Wilk et Kolmogorov–Smirnov) ont montré que certaines variables ne respectaient pas strictement la normalité, ce qui a orienté le choix des méthodes statistiques appropriées. L'utilisation du bootstrap a également permis de confirmer que les moyennes des variables pouvaient être considérées comme approximativement normales.

Les analyses comparatives ont révélé l'influence significative de plusieurs facteurs démographiques et corporels. Le sexe apparaît comme un déterminant majeur de la condition physique, avec des différences significatives observées pour la force de préhension, la flexibilité, la capacité cardiorespiratoire, le pourcentage de masse grasse et l'endurance musculaire. L'âge influence également certaines performances physiques, notamment la capacité cardiovasculaire et l'endurance musculaire, qui tendent à diminuer avec l'avancée en âge.

L'étude des corrélations a mis en évidence des relations importantes entre certaines variables, comme la corrélation négative entre le pourcentage de masse grasse et la force musculaire ou la capacité cardiovasculaire. Ces résultats confirment l'importance de la composition corporelle dans l'évaluation de la condition physique.

En résumé, ce projet démontre la pertinence des outils statistiques pour analyser et interpréter des données réelles issues du domaine de la santé et du sport. Il souligne également l'importance d'une approche méthodologique rigoureuse dans la collecte, l'exploration et l'analyse des données afin de tirer des conclusions fiables et interprétables sur la condition physique des individus étudiés.