Projet sur la régression linéaire

L’objectif de l’étude est de modéliser linéairement le tour de taille en fonction des autres variables

KOSINSKI Justine

★ Introduction

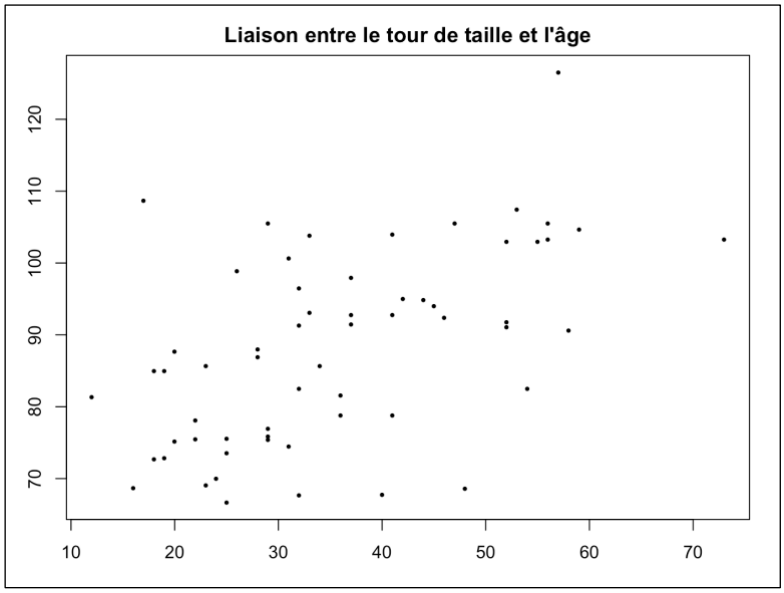
Cette étude se penche sur les données de 60 individus aux États-Unis, explorant le surpoids et l'obésité en utilisant l'Indice de Masse Corporelle (IMC) défini par l'Organisation Mondiale de la Santé (OMS). L'objectif principal est de comprendre comment le tour de taille, un indicateur de santé important, est influencé par des facteurs tels que le sexe, l'âge, la taille et le taux de cholestérol. En simplifiant ces relations, nous espérons identifier des tendances qui pourraient contribuer à la prévention des maladies cardiaques et du diabète de type 2. En somme, cette étude vise à fournir des informations utiles pour améliorer notre compréhension des risques liés au surpoids dans la population américaine.

★ Partie 1 : régression linéaire simple

Le tour de taille peut-il être modélisé linéairement en fonction de l'âge ?

Pour répondre à cette problématique, nous allons suivre le cheminement suivant : réalisation du nuage de points du tour de taille en fonction de l'âge et calcul du coefficient de corrélation linéaire entre ces deux variables, écriture du modèle de régression linéaire simple théorique correspondant, estimation du modèle et représentation de la droite de régression du tour de taille en fonction de l'âge sur le nuage de points, vérification de la normalité des résidus.

Réalisons le nuage de points du tour de taille en fonction de l'âge et calculons le coefficient de corrélation linéaire entre ces deux variables.




```
> round(correlation, 2)
[1] 0.55
```

```
        Pearson's product-moment correlation

data:  AGE and TTAILLE
t = 4.9767, df = 58, p-value = 6.112e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.3403893 0.7032693
sample estimates:
      cor
0.5470323
```

Le nuage de points semble dispersé mais avec une très légère tendance linéaire. En ce qui concerne le coefficient de corrélation que nous avons calculé avec `cor(AGE, TTAILLE)`, la valeur obtenue est d'environ 0.55. Cela suggère une corrélation positive modérée entre le tour de taille (TTAILLE) et l'âge (AGE). La valeur p du test de corrélation est très faible (p-value = 6.112e-06), ce qui indique que la corrélation observée est statistiquement significative.

Le modèle de régression linéaire simple théorique correspondant à notre analyse est exprimé mathématiquement comme suit :

$$\text{TTAILLE}_i = \beta_0 + \beta_1 \cdot \text{AGE}_i + \varepsilon_i$$

où :
- TTAILLE_i est la valeur observée du tour de taille pour l'individu,
- AGE_i est la valeur observée de l'âge pour l'individu,
- β_0 est l'intercept (ordonnée à l'origine) du modèle, représentant la valeur attendue de TTAILLE lorsque AGE est égal à zéro,
- β_1 est la pente du modèle, représentant le changement moyen attendu dans TTAILLE pour une unité d'augmentation de l'âge,
- ε_i est le terme d'erreur, représentant la variation résiduelle ou non expliquée dans TTAILLE.

```
Call:
lm(formula = TTAILLE ~ AGE)

Coefficients:
(Intercept)          AGE
    68.9336       0.5388

> rounded_coefficients <- round(coef(regression), 2)
> cat("Coefficients arrondis:", rounded_coefficients, "\n")
Coefficients arrondis: 68.93 0.54
```

En utilisant les coefficients estimés de notre modèle, l'équation de la droite de régression linéaire est :

$$\text{TTAILLE}_i = 68.93 + 0.54 \cdot \text{AGE}_i + \varepsilon_i$$

Cette équation indique comment le tour de taille TTAILLE varie en fonction de l'âge AGE selon notre modèle de régression linéaire. Nous pouvons interpréter β_1 = 0,54 comme la pente de la relation, indiquant le changement moyen dans le tour de taille pour chaque année d'augmentation de l'âge, et β_0 = 68,93 comme l'intercept, la valeur attendue du tour de taille lorsque l'âge est égal à zéro.

Afin d'estimer ce modèle, nous allons étudier le coefficient de détermination et réaliser deux tests (validité globale du modèle et significativité du paramètre associé à la variable AGE).

```
Call:
lm(formula = TTAILLE ~ AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-26.198  -7.835   1.296   6.127  30.606

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  68.9336     4.1122  16.763  < 2e-16 ***
AGE           0.5388     0.1083   4.977 6.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.28 on 58 degrees of freedom
Multiple R-squared:  0.2992,    Adjusted R-squared:  0.2872
F-statistic: 24.77 on 1 and 58 DF,  p-value: 6.112e-06
```

Le R2 mesure la proportion de la variance totale de la variable dépendante qui est expliquée par le modèle. Plus le R2 est proche de 1, meilleure est l'ajustement. Dans notre cas, un R2 de 0.30 indique que le modèle explique 30% de la variance dans TTAILLE en fonction de l'âge. Cela signifie qu'il y a encore une part importante de la variance qui n'est pas expliquée par le modèle.

Le test de validité globale du modèle évalue l'hypothèse nulle que tous les paramètres du modèle, à l'exception de l'intercept (constante), sont nuls. Mathématiquement, l'hypothèse nulle H0 peut être formulée comme suit :

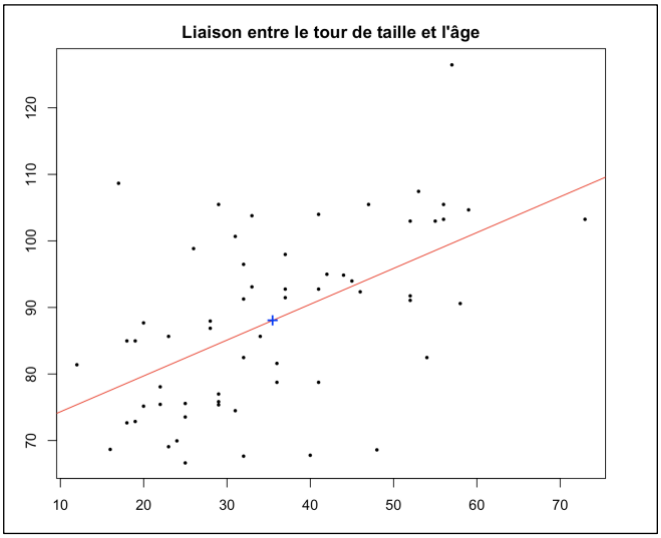
$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

où β1,β2,…,βk sont les coefficients de régression pour toutes les variables explicatives autres que l'intercept dans le modèle.

Dans le tableau de coefficients, on recherche la ligne associée à l'hypothèse nulle que tous les coefficients sont nuls, sauf l'intercept. La statistique de test F et la valeur p associée sont fournies. Si la valeur p est suffisamment petite (typiquement < 0.05), cela fournirait des preuves en faveur du rejet de l'hypothèse nulle, ce qui signifie qu'au moins un des coefficients est significativement différent de zéro, indiquant que le modèle dans son ensemble est statistiquement significatif. Ici la statistique de test F est de 24,77 avec une p-value très petite (6.112e−06). Cela fournit des preuves en faveur du rejet de l'hypothèse nulle que tous les coefficients, sauf l'intercept, sont nuls. En d'autres termes, le modèle dans son ensemble est statistiquement significatif.

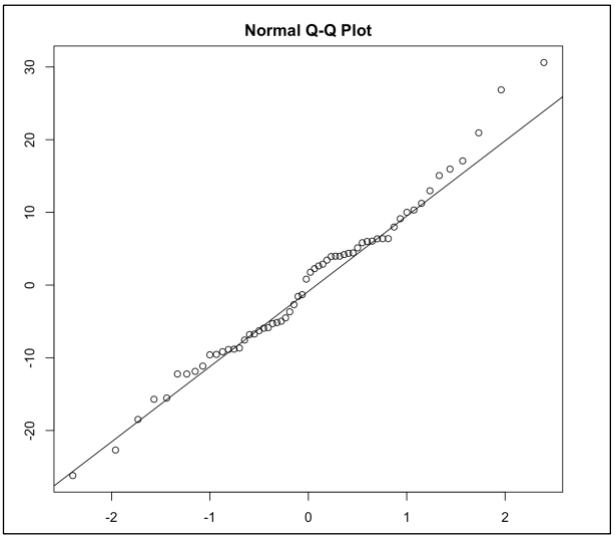
Dans le contexte de notre modèle actuel, le test de significativité du paramètre associé à la variable âge indique que l'âge a un effet significatif sur le tour de taille. La p-value, égale à 6.112e-06 (lue dans la dernière colonne du tableau coefficients) et inférieure au seuil de 5%, permet de rejeter l'hypothèse nulle (stipulant l'absence d'un effet significatif de l'âge) en faveur de l'hypothèse alternative (H1 : $\beta_{\text{AGE}} \neq 0$). Ainsi, la variable âge est considérée comme significative dans le modèle de régression linéaire. Le paramètre estimé associé à l'âge, noté a_chapeau, est égal à 0.54. Cette valeur indique que, en moyenne, le tour de taille augmente d'environ 0.54 unité pour chaque augmentation d'une unité de l'âge. En résumé, l'âge est un facteur significatif et positivement associé aux variations du tour de taille dans le cadre de notre analyse.

L'ajout de la droite de régression et du barycentre du nuage de points permet de visualiser la relation entre les variables "TTAILLE" et "AGE" dans notre modèle. En particulier, la droite de régression (en rouge) représente la tendance linéaire estimée par le modèle, tandis que le point marqué d'un signe plus (en bleu) représente le barycentre du nuage de points. En résumé, l'ajout de ces éléments visuels renforce la compréhension de la modélisation de la relation entre "TTAILLE" et "AGE" et permet une interprétation plus intuitive des résultats obtenus.

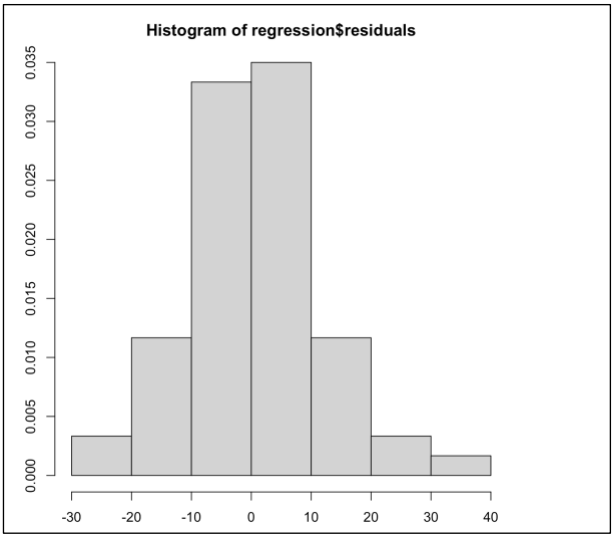


La vérification que la droite de régression passe par le point moyen du nuage est une propriété importante qui confirme l'ajustement du modèle aux données observées. Dans notre cas, cette propriété garantit que la tendance linéaire estimée par le modèle est représentative du comportement moyen de la relation entre le tour de taille et l'âge dans votre échantillon.

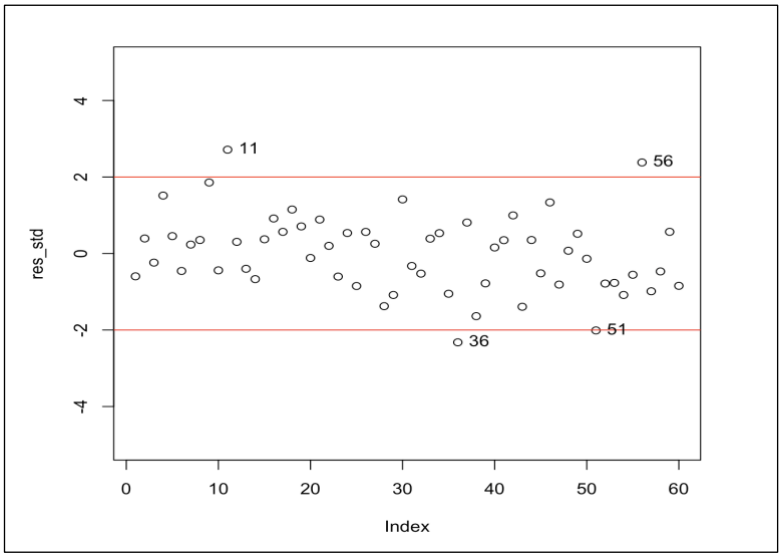
Le graphique QQ-Plot des résidus est utilisé pour évaluer la normalité des erreurs dans le modèle de régression linéaire. Il compare les quantiles théoriques de la distribution normale avec les quantiles observés des résidus. Idéalement, les points sur le graphique devraient être alignés de manière linéaire avec la ligne de référence pour indiquer une distribution normale des résidus. Le QQ-Plot peut aider à repérer d'éventuels points aberrants. Si certains résidus s'éloignent considérablement de la ligne, cela peut indiquer la présence d'observations atypiques ou influentes.



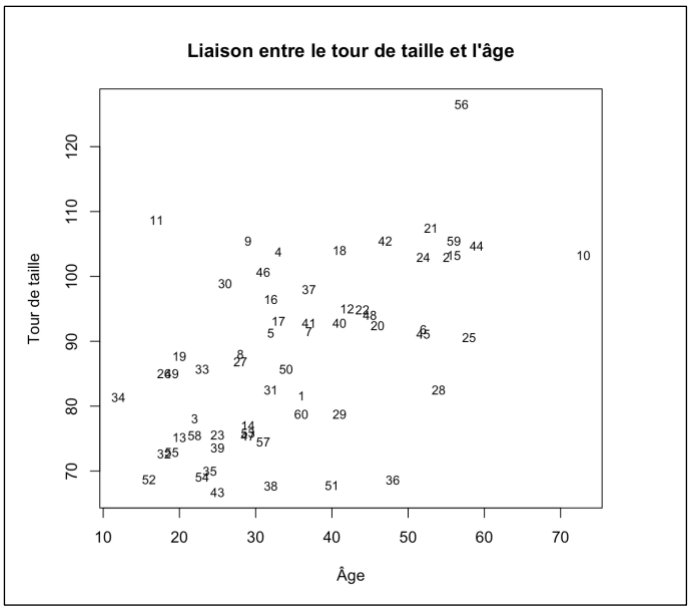
Les points ne suivent pas parfaitement la ligne de référence. Cela suggère que la distribution des résidus n'est pas tout à fait normale, remettant en question l'hypothèse de normalité des erreurs dans notre modèle de régression linéaire. Des déviations importantes peuvent indiquer des violations de cette hypothèse.

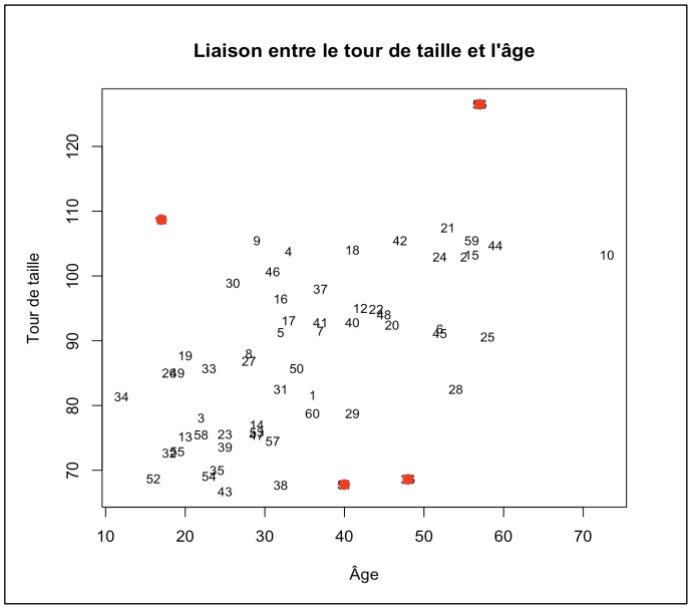


L'histogramme des résidus révèle une légère asymétrie avec un étirement vers la droite. L'asymétrie suggère que les résidus peuvent présenter une certaine tendance ou structure qui ne correspond pas à une distribution normale. Une distribution normale aurait une symétrie parfaite, ce qui n'est pas le cas ici. Cette observation suggère que les résidus peuvent présenter une structure ou une tendance qui ne correspond pas à une distribution normale, renforçant ainsi l'idée que l'hypothèse de normalité des erreurs dans le modèle de régression linéaire n'est pas entièrement vérifiée sur ces données.



Nous avons identifié les indices des observations aberrantes avec les résidus standardisés supérieurs à 2 ou inférieurs à -2. Ensuite, nous affichons ces points aberrants sur le graphique en les mettant en évidence avec une couleur différente (ici, rouge). Cela nous permet de visualiser quelles observations sont considérées comme aberrantes selon le critère des résidus standardisés. Un résidu standardisé élevé indique que cette observation a une influence plus importante sur la modélisation que les autres. Tandis qu'un résidu standardisé faible indique que l'observation a une influence moindre sur la modélisation par rapport aux autres observations. En d'autres termes, les résidus standardisés mesurent à quel point chaque observation s'écarte de la tendance générale du modèle, en tenant compte de l'écart-type des résidus.





Ces points peuvent être des valeurs atypiques ou avoir des caractéristiques particulières qui les rendent différents du reste de la population. L'individu 56 a un tour de taille très élevé par rapport à son âge, ce qui explique que son tour de taille soit mal prédit par la droite. Il en est de même pour l'individu 11, le modèle sous-estime probablement le tour de taille de cet individu par rapport à ce qui est prévu en fonction de l'âge. En ce qui concerne les individus 51 et 36, il s'agit de résidu standardisé : en dessous de -2. Le modèle surestime probablement le tour de taille de ces individus par rapport à ce qui est prévu en fonction de l'âge.

★ Partie 2 : régression linéaire multiple

Déterminer un modèle permettant d'expliquer le tour de taille en fonction des autres variables.

Dans cette étape, nous réaliserons une régression linéaire multiple du tour de taille en utilisant toutes les variables mentionnées précédemment. Nous évaluerons ensuite les résultats en mettant l'accent sur le coefficient de détermination (R2) et le test de validité globale du modèle. Si certains coefficients ne sont pas significatifs, nous procéderons à un ajustement par la méthode de pas à pas descendant, créant éventuellement des variables indicatrices. Enfin, nous vérifierons la normalité des résidus à l'aide d'un histogramme et d'un QQ-plot, incluant un graphique pour repérer d'éventuels points aberrants. Cette approche méthodique nous permettra d'obtenir un modèle significatif tout en assurant la validité des résultats.

Posons le modèle théorique correspondant :

$$\mathrm{TTAILLE}_i = \beta_0 + \beta_1 \times \mathrm{SEXE}_i + \beta_2 \times \mathrm{TAILLE}_i + \beta_3 \times CHOL + \beta_4 \times IMCC_{Surpoids} + \beta_5 \times IMCC_{Obésité} + \varepsilon$$

```
Call:
lm(formula = TTAILLE ~ SEXE + TAILLE + CHOL + IMCC)

Residuals:
     Min       1Q   Median       3Q      Max
-11.7999  -5.2737  -0.9017   3.5520  18.6286

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               9.2788    22.4133   0.414  0.68052
SEXE                      2.3992     2.8443   0.844  0.40267
TAILLE                    0.3985     0.1259   3.164  0.00255 **
CHOL                      0.4360     0.3748   1.163  0.24982
IMCCSurpoids 25<=IMC<30  14.4068     1.9853   7.257 1.58e-09 ***
IMCCObésité IMC >=30     29.4241     2.6934  10.925 2.72e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.844 on 54 degrees of freedom
Multiple R-squared:   0.76,     Adjusted R-squared:  0.7378
F-statistic: 34.21 on 5 and 54 DF,  p-value: 1.396e-15
```

Le coefficient de détermination est 0.76, ce qui signifie que 76% de la variation dans le tour de taille (TTAILLE) est expliqué par les variables explicatives du modèle, à savoir le sexe, la taille, le taux de cholestérol et l'IMCC. Un R2 proche de 1 indique un bon ajustement du modèle aux données.

Le test global (F-test) rejette l'hypothèse nulle (H0) selon laquelle tous les coefficients sont nuls, sauf l'intercept (pvalue< 1.396e-15 <5%). Ainsi, le modèle dans son ensemble est globalement significatif.

Les variables qui ont un effet significatif sur le tour de taille sont celles dont la p-valeur du test de significativité est inférieure à 5 %. On lit la p-valeur dans la dernière colonne du tableau coefficients. Par exemple, test de H0 :a_2=0 contre H1 a_2 différent de 0. On observe la p-value< 0.00255 <5% donc on rejette H0. La variable TAILLE a un effet significatif sur le tour de taille. En comparant toutes les p-valeurs à 5 %, on peut donc conclure que la taille et l'IMC des personnes en surpoids et en obésité sont

significatives, ce qui suggère qu'ils ont un effet significatif sur le tour de taille. En revanche, le sexe (p=0.40 > 0.05) et le taux de cholestérol (p=0.25 > 0.05) ne sont pas significatifs (p>0.05) et ne contribuent pas de manière significative à la prédiction du tour de taille.

Ici, on veut un pas à pas avec un critère qui permette d'avoir à la fin toutes les variables significatives. On va le faire "à la main". On enlève d'abord la variable la moins significative du modèle (celle avec la plus grande p-valeur de la sortie de reg à savoir SEXE (p=0.40).

```
Call:
lm(formula = TTAILLE ~ TAILLE + CHOL + IMCC)

Residuals:
    Min      1Q  Median      3Q     Max
-10.629  -4.650  -1.520   3.843  20.017

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              23.58172   14.61850   1.613   0.1124
TAILLE                    0.32335    0.08879   3.642   0.0006 ***
CHOL                      0.29588    0.33507   0.883   0.3811
IMCCSurpoids 25<=IMC<30 14.39206    1.98005   7.269 1.37e-09 ***
IMCCObésité IMC >=30    29.06611    2.65273  10.957 1.91e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.826 on 55 degrees of freedom
Multiple R-squared:  0.7569,    Adjusted R-squared:  0.7392
F-statistic: 42.81 on 4 and 55 DF,  p-value: 2.809e-16
```

La variable CHOL est non significative du modèle initial (p=0.38), il faut donc la retirer à son tour.

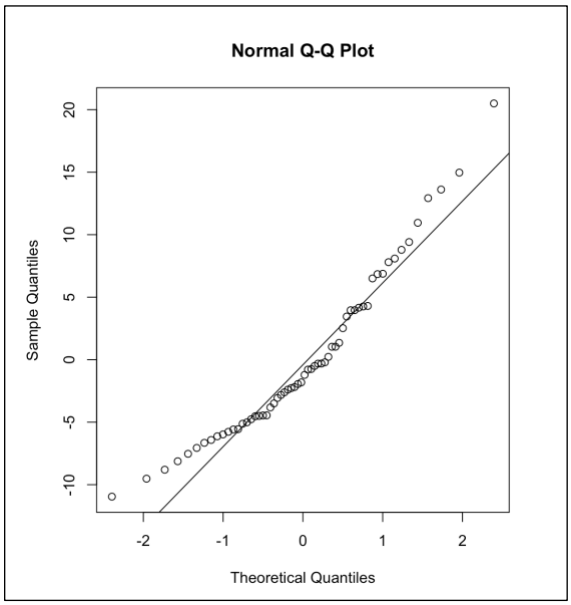
```
Call:
lm(formula = TTAILLE ~ TAILLE + IMCC)

Residuals:
    Min      1Q  Median      3Q     Max
-10.957  -4.826  -1.518   4.010  20.498

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              22.0620    14.4883   1.523 0.133448
TAILLE                    0.3378     0.0871   3.878 0.000279 ***
IMCCSurpoids 25<=IMC<30  14.5391     1.9691   7.383 8.04e-10 ***
IMCCObésité IMC >=30     29.5531     2.5897  11.412 3.08e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.812 on 56 degrees of freedom
Multiple R-squared:  0.7534,    Adjusted R-squared:  0.7402
F-statistic: 57.04 on 3 and 56 DF,  p-value: < 2.2e-16
```

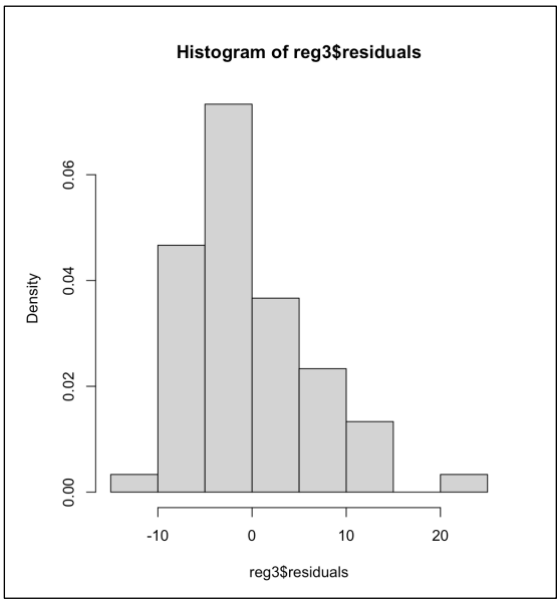
Toutes les variables sont significatives, ce qui signifie qu’on a finit le pas à pas. On retient donc comme modèle finale le modèle 3.

Pour le modèle final obtenu par le pas à pas, on pose le modèle de régression linéaire théorique suivant :

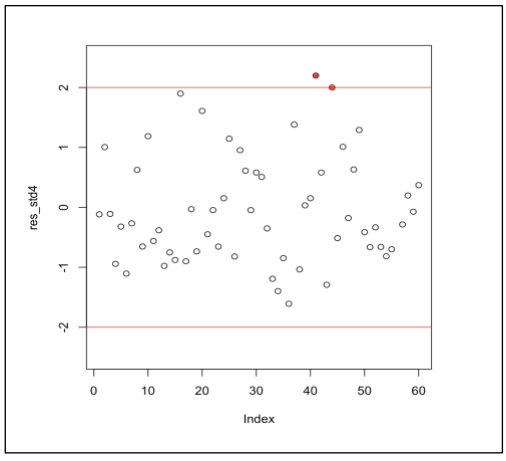
$$TTAILLE = \beta_0 + \beta_1 \times TAILLE + \beta_2 \times IMCC_{Surpoids} + \beta_3 \times IMCC_{Obésité} + \varepsilon$$

On conclut ici aussi que le modèle est globalement valide (avec le test : pvalue< 2.2e-16 <5% donc on rejette H0 car au moins l'un des paramètres n'est pas nul) quant au R2 ajusté, il est quasiment identique, légèrement inférieur, mais pas significativement et comme on préfère garder un modèle avec moins de paramètres et où toutes les variables sont significatives, on retient bien le modèle 3. En ce qui concerne les paramètres estimés, le coefficient de 0.34 suggère que, toutes choses égales par ailleurs, une augmentation d'un cm de la taille est associée à une augmentation de 0.34 cm dans le tour de taille. Le coefficient de 14.54 suggère que, toutes choses égales par ailleurs, une augmentation d'une unité dans l'indice de masse corporelle (IMC) spécifiquement pour les personnes en surpoids, est associée à une augmentation de 14.54 cm dans le tour de taille. Le coefficient de 29.55 suggère que, toutes choses égales par ailleurs, une augmentation d'une unité dans l'indice de masse corporelle (IMC) spécifiquement pour les personnes en obésité, est associée à une augmentation de 29.55 cm dans le tour de taille. Ainsi, ces résultats suggèrent que la taille et l'indice de masse corporelle des personnes en surpoids et en obésité sont des prédicteurs significatifs du tour de taille dans ce modèle. Une augmentation de la taille et de l'IMC des personnes en surpoids et en obésité est associée à une augmentation du tour de taille.





Si les points du QQ-plot étaient alignés le long de la ligne droite, cela suggérerait une distribution normale des résidus. Cependant, ici, les points ne sont pas alignés, indiquant une possible non-normalité des résidus. Ce n'est donc pas gaussien. Un histogramme symétrique suggère une distribution normale. Ici, comme l'histogramme n'est pas symétrique, cela suggère également une possible non-normalité des résidus, pas gaussien.




```
> which(abs(res_std4)> 2)
41 56
```

En utilisant une échelle standardisée (résidus standardisés), on a identifié les points aberrants en fonction d'un seuil de ± 2. Les points aberrants sont marqués en rouge sur le graphique. Certains points aberrants sont identifiés, mais ils ne remettent pas fondamentalement en question la validité du modèle.

En résumé, le modèle final, bien que simplifié, explique de manière significative la variabilité du tour de taille en fonction de la taille et de l'IMC des individus en surpoids et en obésité.

★ Conclusion

L'étude entreprise visait à modéliser la relation entre le tour de taille et plusieurs variables explicatives, en se concentrant principalement sur l'âge dans une première étape, puis en intégrant d'autres facteurs tels que le sexe, la taille, le taux de cholestérol, l'Indice de Masse Corporelle (IMC) avec des catégories distinctes pour le surpoids et l'obésité dans une seconde étape. Malgré la robustesse du modèle, des limitations doivent être notées. La présence de points aberrants dans les données souligne la nécessité d'une exploration plus approfondie, peut-être en identifiant des variables supplémentaires qui pourraient expliquer ces observations atypiques. Des projets futurs pourraient explorer d'autres variables pertinentes, telles que l'activité physique, les habitudes alimentaires, ou des facteurs génétiques, pour enrichir la compréhension de la variation du tour de taille. Les résultats de cette étude peuvent avoir des implications significatives, notamment dans le domaine de la santé publique. Comprendre les déterminants du tour de taille peut contribuer à identifier des facteurs de risque potentiels de maladies métaboliques. Ainsi, cette étude fournit une base solide pour comprendre et prédire le tour de taille en fonction de variables clés. Cependant, une approche multidimensionnelle et des analyses plus approfondies sont nécessaires pour élargir notre compréhension et tirer des conclusions plus robustes.

★ Annexe

```
#Projet 2 sur la régression linéaire
setwd("/Users/justinekosinski/Desktop/MIAGE/L3/S1/Statistiques/Proje
t2")
data_imc_TP2_groupe1=read.table("data_imc_TP2_groupe1.txt",header=TR
UE)
is.data.frame(data_imc_TP2_groupe1)
View(data_imc_TP2_groupe1)
attach(data_imc_TP2_groupe1)


IMCC <- cut(data_imc_TP2_groupe1$IMC, breaks = c(-Inf, 25, 30, Inf),
labels = c("Corpulence normale IMC<25", "Surpoids 25<=IMC<30",
"Obésité IMC >=30"))
IMCC

CHOL <- data_imc_TP2_groupe1$CHOL / 100
CHOL

#Partie 1 : régression linéaire simple
# Question 1
# nuage de points
par(mfrow=c(1,1))
plot(AGE,TTAILLE, pch="•",main="Liaison entre le tour de taille et
l'âge", xlab="Âge",ylab="Tour de taille",
     xlim=c(min(AGE), max(AGE)), ylim=c(min(TTAILLE), max(TTAILLE)))

# coefficient de corrélation linéaire et test associé
correlation <- cor(AGE,TTAILLE)
cor.test(AGE,TTAILLE)
round(correlation, 2)
# Il existe une liaison linéaire significative (p<5%) positive
modérée entre ttaille et age (r=0,55)

# Question 2
# Modèle théorique : ttaille_i = beta_0 + beta_1 age_i + epsilon_i
# ttaille_i = a age_i + b + e_i pour i=1,...,n
regression =lm(TTAILLE~AGE) # droite de regression
regression
rounded_coefficients <- round(coef(regression), 2)
cat("Coefficients arrondis:", rounded_coefficients, "\n")
# a_chapeau=0.54  et b_chapeau=68.93
# L'équation de la droite de régression est : y = 0,54 x + 68.93

summary(regression)
# R2=0,30
# 30% de la variation totale des tours de taille est expliquée par
le modèle linéaire,
# c'est-à-dire par la variation de l'âge.
```

```
# Ce coefficient de détermination est proche de 0 donc le modèle
ajuste pas bien les données.

# Test de validité globale du modèle
# H0 : tous les paramètres sont nuls sauf la constante
# p-value = 6.112e-06<5% donc on rejette H0 donc le modèle est
globalement valide

# Test de significativité du paramètre de Surface
# H0 : a=0 contre H1 : a différent de 0
# p-value = 6.112e-06<5% (lue dans la dernière colonne du tableau
coefficients) donc on rejette H0
# et la variable age est significative
# On peut donc commenter son paramètre estimé  a_chapeau=0,54
# Le coefficient de "AGE" (0.54) indique que, en moyenne, le tour de
taille augmente d'environ 0.54 unité pour
# chaque augmentation d'une unité de l'âge.

# Question 3
# Ajout de la droite de régression et du barycentre du nuage de
points
abline(regression,col="red")
points(mean(AGE),mean(TTAILLE),pch="+",col="blue",cex=1.5)
# argument cex=1.5 pour augmenter la taille du marqueur
# on vérifie la propriété vue en cours :
# la droite de régression passe par le point moyen du nuage

# Question 4
# sert à vérifier la normalité des erreurs et à repérer d'éventuels
points aberrants
qqnorm(regression$residuals)
qqline(regression$residuals)
# on voit que les points ne sont pas vraiment
# alignés ce qui remet en cause l'hypothèse de normalité des erreurs
hist(regression$residuals,freq=FALSE)
#l'histogramme des résidus est légèrement asymétrique avec un
étalement à gauche
#ce qui confirme que l'hypothèse de normalité des erreurs n'est pas
vérifiée tout à fait sur ces données.

res_std=regression$residuals/11.28
#calcul des résidus standardisés (11.28 désigne l'écart-type des
résidus)
# cette valeur est la "residual standard error " et se lit dans la
sortie de summary (regression)

plot(res_std,ylim=c(-5,5)) #Représentation graphique des résidus
standardisés
abline(h=-2,col="red")
abline(h=2,col="red")

# Les points aberrants sont ceux qui ont un résidu standardisé > 2
en valeur absolue
identify(1:length(regression$residuals),regression$residuals/11.28)
```

```
help(identify)
which(res_std< -2) # points aberrants (obs 36 et 51)
which(res_std> 2) # points aberrants (obs 11 et 56)
# Obtenir les indices des observations aberrantes dans les données
d'origine
aberrant_indices <- c(which(res_std < -2), which(res_std > 2))
# Afficher les indices des observations aberrantes
print(aberrant_indices)
# Afficher le graphique avec les points aberrants mis en évidence
plot(AGE, TTAILLE, type="n", main="Liaison entre le tour de taille
et l'âge", xlab="Âge",
     ylab="Tour de taille")
text(AGE, TTAILLE, labels=1:length(regression$residuals), cex=0.8)
points(AGE[aberrant_indices], TTAILLE[aberrant_indices], pch=16,
col="red", cex=1.5)
# l'individu 56 a un tour de taille très eleve par rapport à son
âge, ce qui
# explique que son tour de taille soit mal prédit par la droite

#Partie 2 : régression linéaire multiple
# Question 1
reg = lm(TTAILLE~SEXE+TAILLE+CHOL+IMCC)
summary(reg)
# *** ==> très significative
# coeff de determination : R2 = 0.76 76% de la variation du tour de
taille est expliqué par la variation
# des variables explicatives du modèle, à savoir le sexe (recodé
avec les labels "homme" et "femme"),
# la taille (en cm), le  taux de cholestérol (en g/L), l'imc
surpoids et obésité
# ce coefficient de détermination est proche de 1, donc on a affaire
à un
# bon modèle qui ajuste bien les données

# Test de validité globale du modèle
#H0 : Tous les paramètres sont nuls sauf la constante
#H0 : a_1=a_2=a_3=a_4=0
# pvalue< 1.396e-15 <5% donc on rejette H0 : le modèle est
gobalement valide (au moins l'un des
# paramètres n'est pas nul)

#Tests de significativité des paramètres
# les variables qui ont un effet significatif sur le tour de taille
# sont celles dont la p-valeur du test de significativité est
inférieure
#  à 5 %. On lit la p-valeur dans la dernière colonne du tableau
coefficients
# par exemple, test de H0 :a_2=0 contre H1 a_2 différent de 0
# p-value< 0.00255 <5% donc on rejette H0
# et la variable TAILLE a un effet significatif sur le tour de
taille
# en comparant toutes les p-valeurs à 5 %, on peut donc conclure que
la taille et l'imc surpoids et obésité
```

```
# sont significatives mais pas le sexe (p=0.40 > 0.05) et le taux de
cholestérol (p=0.25 > 0.05)

# Question 2
#Ici, on veut un pas à pas avec
# un critère qui permette d'avoir à la fin toutes les variables
significatives
# On va le faire "à la main"
# on enlève d'abord la variable la moins significative du modèle
(celle avec
# la plus grande p-valeur de la sortie de reg
#  à savoir SEXE
reg2 = lm(TTAILLE~TAILLE+CHOL+IMCC)
summary(reg2)
# CHOL est NS donc on l'enlève
reg3 = lm(TTAILLE~TAILLE+IMCC)
summary(reg3)
# toutes les variables sont significatives donc on a fini le pas à
pas
# et on retient comme modèle final le modèle 3

#Modèle 3 :  TTAILLE_i = c_0+ c_1*TAILLE_i+c_2+*TAILLE_i+c_3+e_i
pour i=1,...,n
# On conclut ici aussi que le modèle est globalement valide (avec le
test)
# quant au R2 ajusté, il est quasiment identique et comme on préfère
# garder un modèle avec moins de paramètres et où toutes les
variables
# sont significatives, on retient bien le modèle 3

# Test de validité globale du modèle
#H0 : Tous les paramètres sont nuls sauf la constante
#H0 : a_1=a_2=0
# pvalue< 2.2e-16 <5% donc on rejette H0 : le modèle est gobalement
valide (au moins l'un des
# paramètres n'est pas nul)

#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept)              22.0620    14.4883   1.523 0.133448
#TAILLE                    0.3378     0.0871   3.878 0.000279 ***
#IMCCSurpoids 25<=IMC<30  14.5391     1.9691   7.383 8.04e-10 ***
#IMCCObésité IMC >=30     29.5531     2.5897  11.412 3.08e-16 ***

# c1_chapeau = 0.3378
# Toutes choses égales par ailleurs, 1 cm supplémentaire  fait
augmenter le tour de taille de 0.34 cm

# c2_chapeau = 14.5391
# Toutes choses égales par ailleurs, une augmentation d'une unité
dans l'indice de masse corporelle (IMC)
# spécifiquement pour les personnes en surpoids, est associée à une
augmentation de 14.54 cm dans le tour de taille
```

```
# c3_chapeau = 29.5531
# Toutes choses égales par ailleurs, une augmentation d'une unité
dans l'indice de masse corporelle (IMC)
# spécifiquement pour les personnes en obésité, est associée à une
augmentation de 29.55 cm dans le tour de taille

# Question 3
qqnorm(reg3$residuals)
qqline(reg3$residuals) # pas alignés donc pas gaussien
hist(reg3$residuals,freq=FALSE) # pas très symétrique donc pas
gaussien
res_std4=reg3$residuals/6.81
plot(res_std4,ylim=c(-2.5,2.5))
abline(h=-2,col="red")
abline(h=2,col="red")
# Identification des points aberrants
aberrant_points <- which(abs(res_std4) > 2) # 41 et 56
# Mettre en surbrillance les points aberrants sur le graphique
points(aberrant_points, res_std4[aberrant_points], pch = 20, col =
"red")

detach(data_imc_TP2_groupe1)
```