Analyse de variance sur le jeu de données Penguins

NDAO Pape Semou

28 février 2025

Dans ce TP, nous travaillons sur le jeu de données «penguins.csv» qui rassemble différentes mesures sur 344 manchots en Antarctique. L'objectif est d'étudier une variable quantitative (ici, bill_length_mm) en fonction de variables qualitatives et quantitatives. Nous aborderons :

- Le prétraitement des données pour retirer les valeurs manquantes.
- L'analyse de variance (ANOVA) avec une première variable qualitative (Qual1), puis une seconde (Qual2) et enfin l'ajustement d'un modèle intégrant plusieurs variables qualitatives.
- L'analyse de covariance en ajoutant une seconde variable quantitative (Quant2) et en explorant l'interaction avec une variable qualitative.

1.Traitement et visualisation des données de

```
# Chargement des données
df <- read.csv("penguins.csv", header = TRUE, stringsAsFactors = TRUE)
summary(df)
```

```
##       species          island    bill_length_mm  bill_depth_mm
##  Adelie   :152   Biscoe   :168   Min.   :32.10   Min.   :13.10
##  Chinstrap: 68   Dream    :124   1st Qu.:39.23   1st Qu.:15.60
##  Gentoo   :124   Torgersen: 52   Median :44.45   Median :17.30
##                                  Mean   :43.92   Mean   :17.15
##                                  3rd Qu.:48.50   3rd Qu.:18.70
##                                  Max.   :59.60   Max.   :21.50
##                                  NA's   :2       NA's   :2
##  flipper_length_mm  body_mass_g       sex           year
##  Min.   :172.0     Min.   :2700   female:165   Min.   :2007
##  1st Qu.:190.0     1st Qu.:3550   male  :168   1st Qu.:2007
##  Median :197.0     Median :4050   NA's  : 11   Median :2008
##  Mean   :200.9     Mean   :4202                Mean   :2008
##  3rd Qu.:213.0     3rd Qu.:4750                3rd Qu.:2009
##  Max.   :231.0     Max.   :6300                Max.   :2009
##  NA's   :2         NA's   :2
```

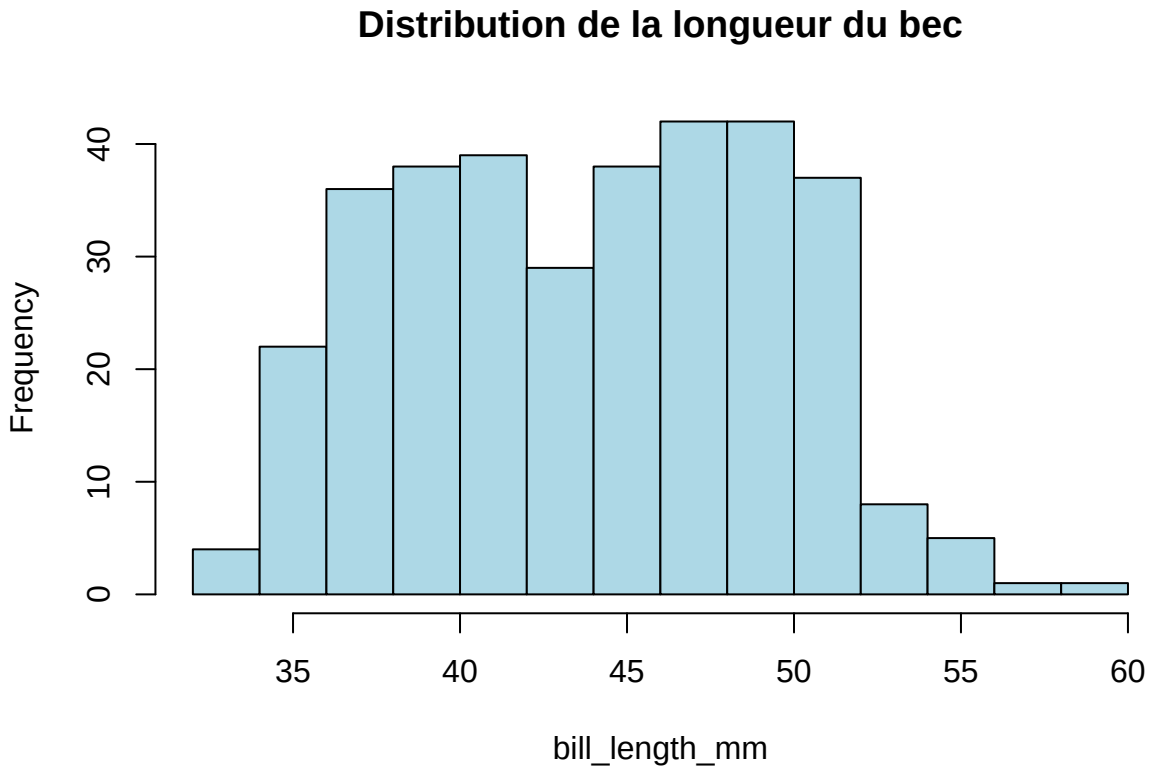
```
# On retire les lignes avec des valeurs manquantes
df_clean <- df[!is.na(df$bill_length_mm), ]
# Vérification
cat("Nbre de lignes avec les valeurs manquantes", nrow(df), "\n")
```

```
## Nbre de lignes avec les valeurs manquantes 344
```

```r
cat("Nbre de lignes apres retrait des valeurs manquantes", nrow(df_clean), "\n")
```

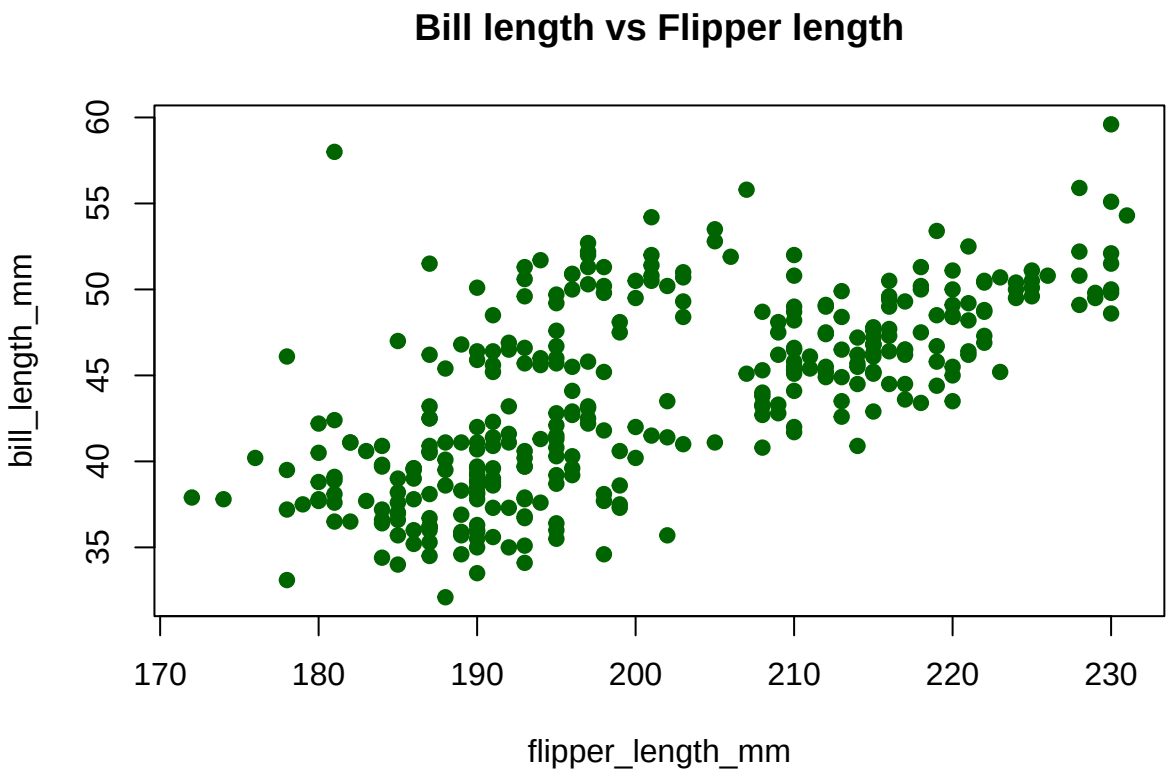
```
## Nbre de lignes apres retrait des valeurs manquantes 342
```

```r
# Histogramme de bill_length_mm
hist(df_clean$bill_length_mm, main="Distribution de la longueur du bec", xlab="bill_length_mm", col="lig
```



La plupart des individus se situent dans une plage de 39-47 mm. La distribution n'est pas strictement normale (à vérifier avec un test ou un QQ plot), mais ne présente pas de fortes anomalies non plus. Quelques valeurs à l'extrémité supérieure (au-delà de 55 mm) pourraient être des individus plus grands ou de certaines espèces de manchots dont le bec est plus long.

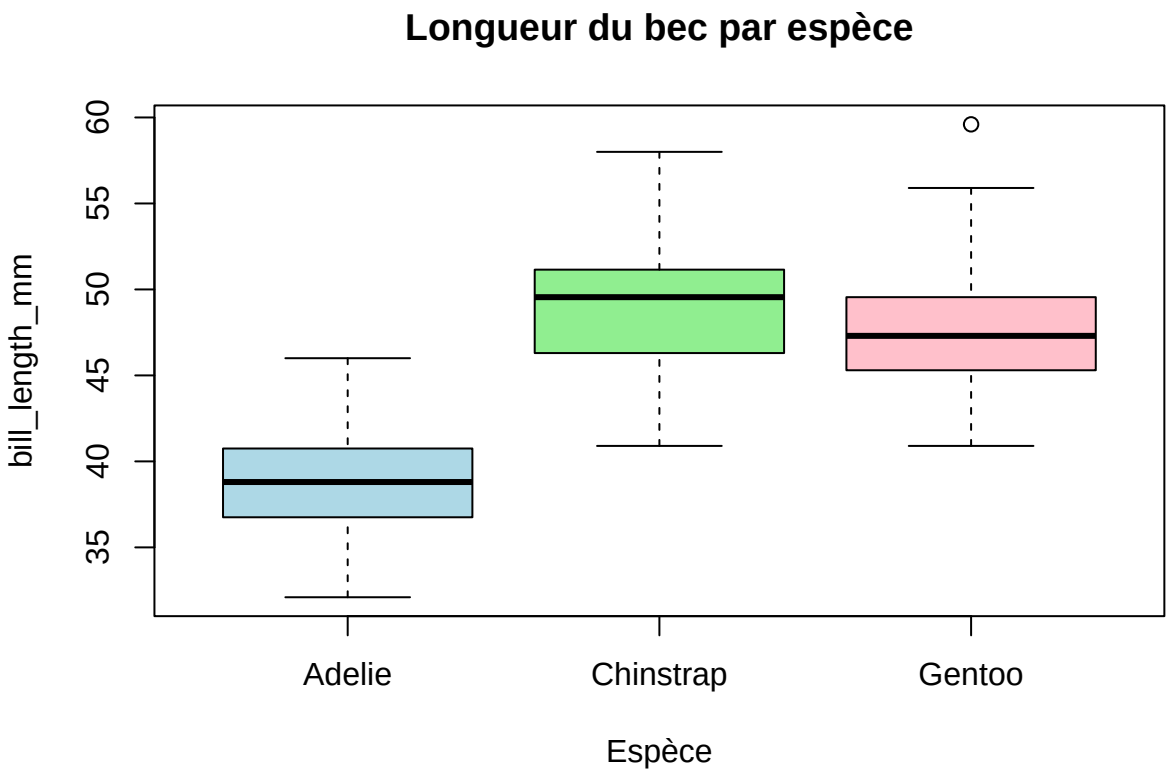
```r
# Diagramme de dispersion de bill_length_mm par rapport  flipper_length_mm)
plot(df_clean$flipper_length_mm, df_clean$bill_length_mm,
     xlab="flipper_length_mm", ylab="bill_length_mm",
     main="Bill length vs Flipper length", col="darkgreen", pch=19)
```



3. Analyse de Variance (ANOVA) L'ANOVA nous permet d'étudier l'influence d'une variable qualitative sur notre variable quantitative choisie.

3.1. Analyse avec une première variable qualitative (Qual1) Nous choisissons species (espèce du manchot) comme Qual1.

```
# Boxplot de bill_length_mm par espèce
boxplot(bill_length_mm ~ species, data = df_clean,
        main="Longueur du bec par espèce",
        xlab="Espèce", ylab="bill_length_mm", col=c("lightblue","lightgreen","pink"))
```



-Adelie présente la médiane la plus basse, autour de 38-39 mm. -Chinstrap a la médiane la plus élevée, autour de 48-49 mm. -Gentoo se situe entre les deux, avec une médiane proche de 45 mm. Cela suggère que, de manière générale, la longueur du bec diffère selon l'espèce, Chinstrap étant la plus grande, Adelie la plus petite et Gentoo intermédiaire.

```r
modele_species <- lm(bill_length_mm ~ species, data = df_clean)
summary(modele_species)
```

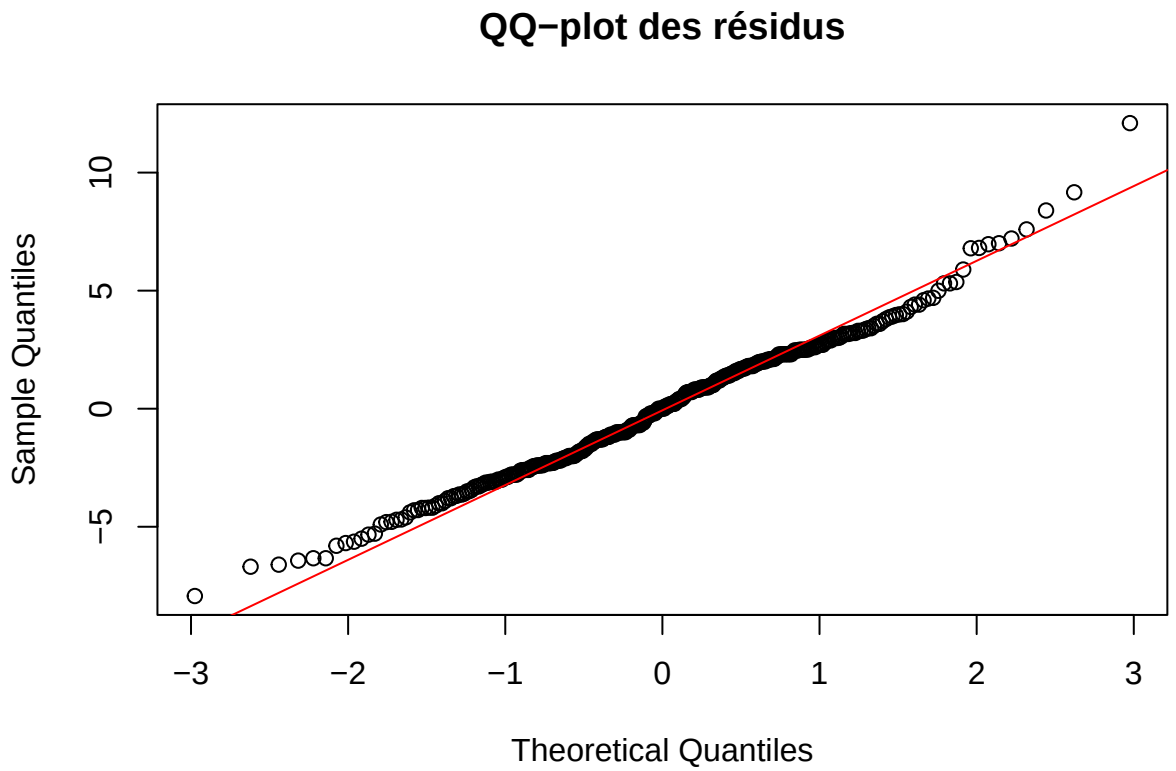
```
##
## Call:
## lm(formula = bill_length_mm ~ species, data = df_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9338 -2.2049  0.0086  2.0662 12.0951
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       38.7914     0.2409  161.05   <2e-16 ***
## speciesChinstrap  10.0424     0.4323   23.23   <2e-16 ***
## speciesGentoo      8.7135     0.3595   24.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.96 on 339 degrees of freedom
## Multiple R-squared:  0.7078, Adjusted R-squared:  0.7061
## F-statistic: 410.6 on 2 and 339 DF,  p-value: < 2.2e-16
```

```r
anova_species <- anova(modele_species)
print(anova_species)
```

```
## Analysis of Variance Table
##
## Response: bill_length_mm
##            Df Sum Sq Mean Sq F value    Pr(>F)
## species     2 7194.3  3597.2   410.6 < 2.2e-16 ***
## Residuals 339 2969.9     8.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

la p-value est inférieure à 2.2e-16, on rejette très fortement l'hypothèse nulle. Cela signifie qu'il existe une différence statistiquement significative de la longueur du bec selon l'espèce. En d'autres termes, species est un facteur qui explique une part importante de la variabilité de bill_length_mm.

```r
# QQ plot pour vérifier la normalité
qqnorm(modele_species$residuals, main="QQ-plot des résidus")
qqline(modele_species$residuals, col="red")
```

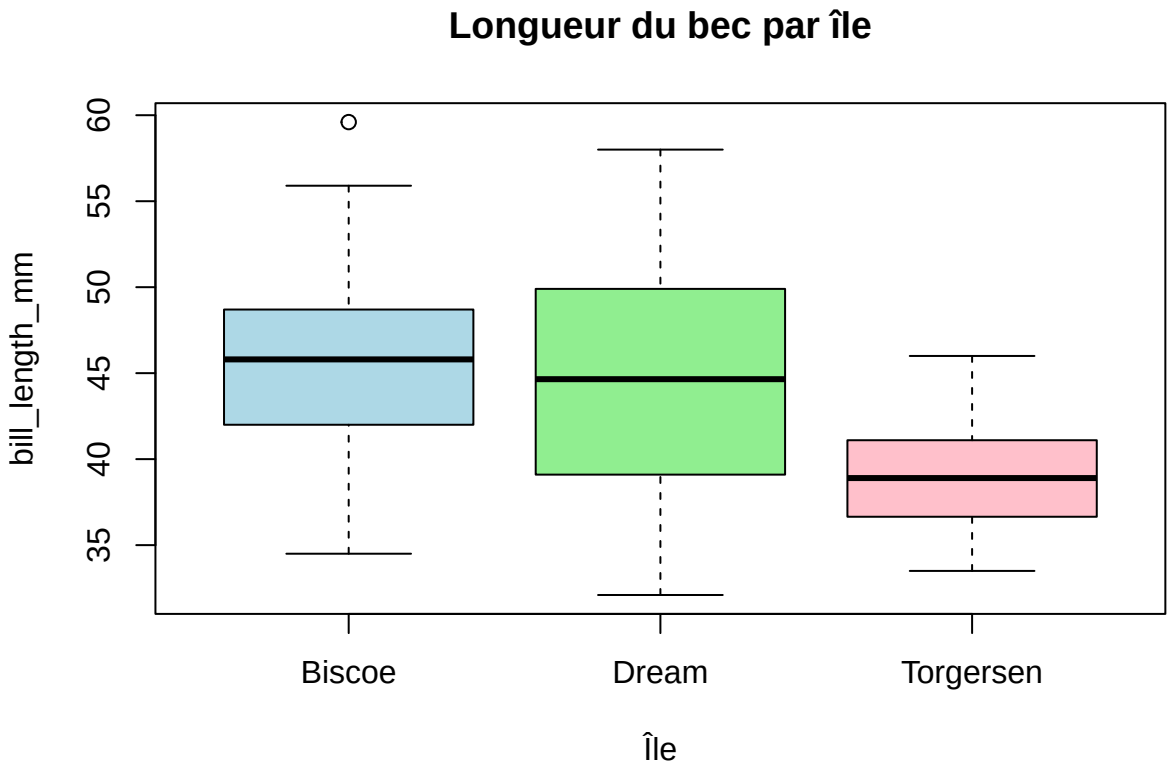


La majorité des points se trouvent proches de la ligne rouge, indiquant que la distribution des résidus est globalement compatible avec une distribution normale. l'hypothèse de normalité des résidus est validée.

3.2. Analyse avec une deuxième variable qualitative (Qual2)

Prenons island comme variable qualitative.

```r
# Boxplot de bill_length_mm par island
boxplot(bill_length_mm ~ island, data = df_clean,
        main="Longueur du bec par île",
        xlab="Île", ylab="bill_length_mm", col=c("lightblue", "lightgreen", "pink"))
```




```r
modele_island <- lm(bill_length_mm ~ island, data = df_clean)
summary(modele_island)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ island, data = df_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0677  -3.8559   0.2958   3.8175  14.3425
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      45.2575     0.3897 116.127  < 2e-16 ***
## islandDream      -1.0897     0.5970  -1.825   0.0688 .
## islandTorgersen  -6.3065     0.8057  -7.827 6.44e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.036 on 339 degrees of freedom
```

```
## Multiple R-squared:  0.154,  Adjusted R-squared:  0.149
## F-statistic: 30.86 on 2 and 339 DF,  p-value: 4.86e-13
```

```
anova_island <- anova(modele_island)
print(anova_island)
```

```
## Analysis of Variance Table
##
## Response: bill_length_mm
##            Df Sum Sq Mean Sq F value   Pr(>F)
## island      2 1565.6  782.80  30.862 4.86e-13 ***
## Residuals 339 8598.6   25.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

-l'île a un impact significatif sur la longueur du bec (p-value < 0.001). -Torgersen a un bec significativement plus court que Biscoe (de l'ordre de 6 mm). -Dream ne diffère pas significativement de Biscoe au seuil de 5 %, bien qu'on observe une légère tendance.

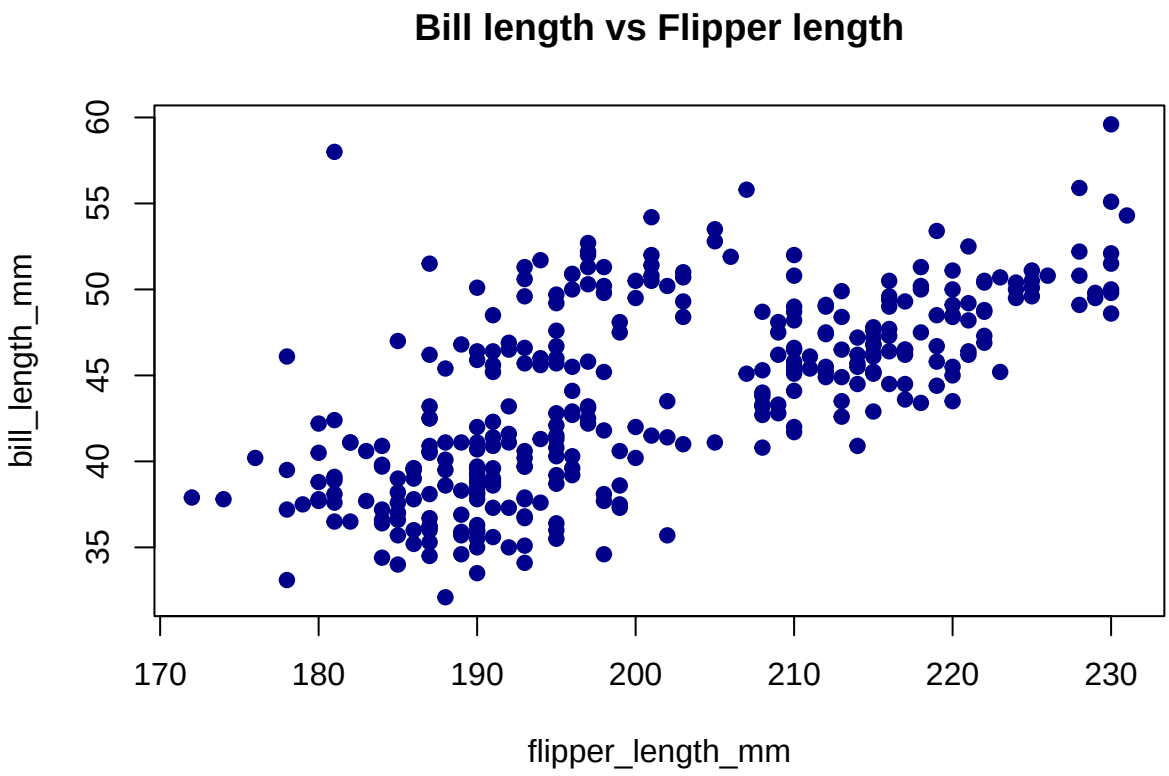
4. Analyse de Covariance (ANCOVA)

L'analyse de covariance permet d'étudier l'effet d'une variable quantitative (Quant2) tout en tenant compte d'une variable qualitative.

4.1. Étude de la relation entre Quant1 et Quant2

-species (Qual1) -flipper_length_mm (Quant2)

```
plot(df_clean$flipper_length_mm, df_clean$bill_length_mm,
     xlab="flipper_length_mm", ylab="bill_length_mm",
     main="Bill length vs Flipper length", pch=19, col="darkblue")
```



Le graphique suggère une corrélation positive notable entre la longueur du bec et la longueur de la nageoire chez ces manchots.

```
modele_cov <- lm(bill_length_mm ~ flipper_length_mm, data = df_clean)
summary(modele_cov)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ flipper_length_mm, data = df_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5792 -2.6715 -0.5721  2.0148 19.1518
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -7.26487    3.20016   -2.27   0.0238 *
## flipper_length_mm  0.25477    0.01589   16.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.126 on 340 degrees of freedom
## Multiple R-squared:  0.4306, Adjusted R-squared:  0.4289
## F-statistic: 257.1 on 1 and 340 DF,  p-value: < 2.2e-16
```

pour chaque millimètre supplémentaire de flipper_length_mm, la longueur du bec (bill_length_mm) augmente de 0.25477 mm.La variable flipper_length_mm est un prédicteur statistiquement très significatif de

bill_length_mm, expliquant environ 43 % de la variabilité, avec une marge d'erreur résiduelle d'environ 4 mm.

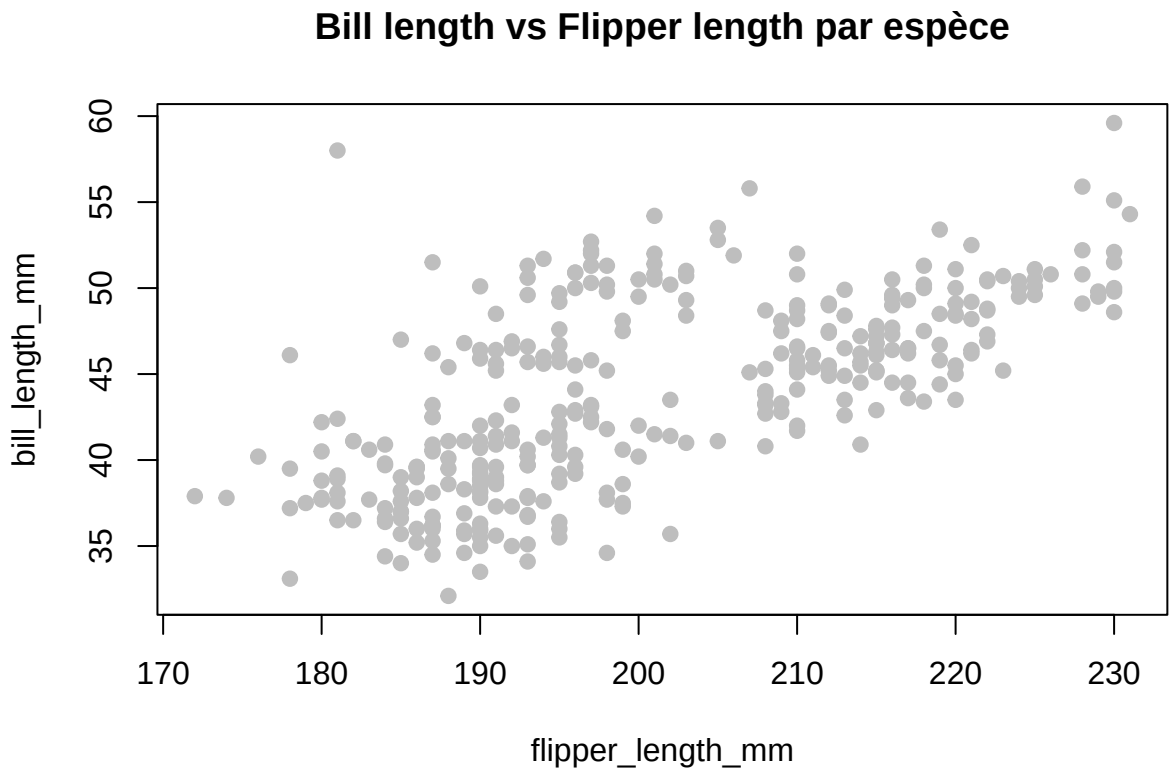
4.2. Régression par modalité de la variable qualitative (species)

4.2.a. Ajustement par modalité

Pour chaque espèce, nous ajustons un modèle linéaire de bill_length_mm en fonction de flipper_length_mm.

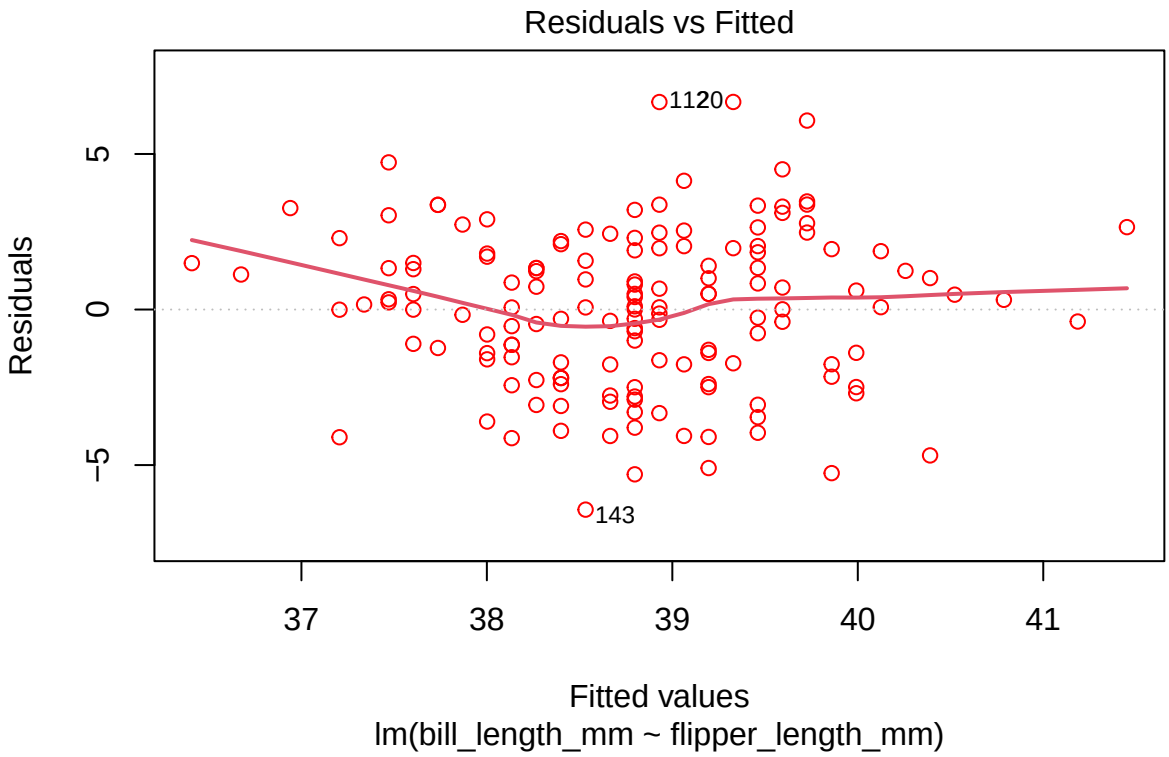
```
# Liste des espèces
especes <- levels(df_clean$species)
couleurs <- c("red", "blue", "green")
```

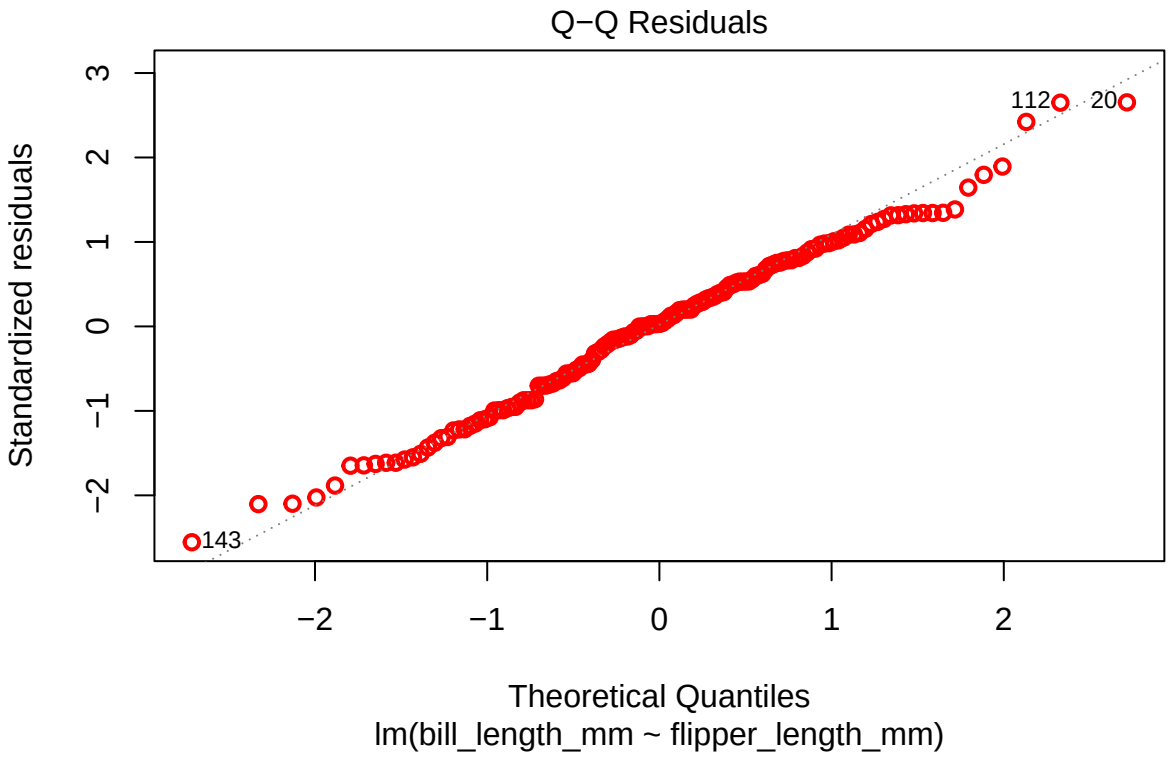
```
# Création du graphique global
plot(df_clean$flipper_length_mm, df_clean$bill_length_mm,
     xlab="flipper_length_mm", ylab="bill_length_mm",
     main="Bill length vs Flipper length par espèce", pch=19, col="gray")
```

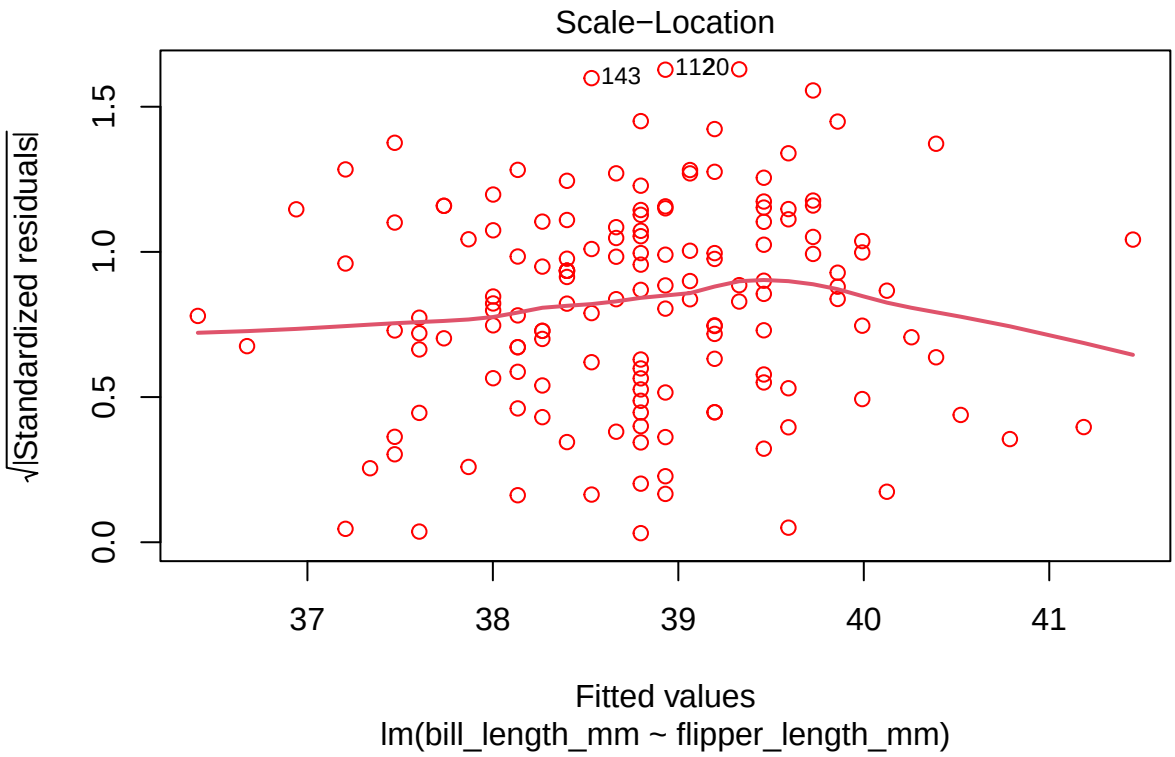


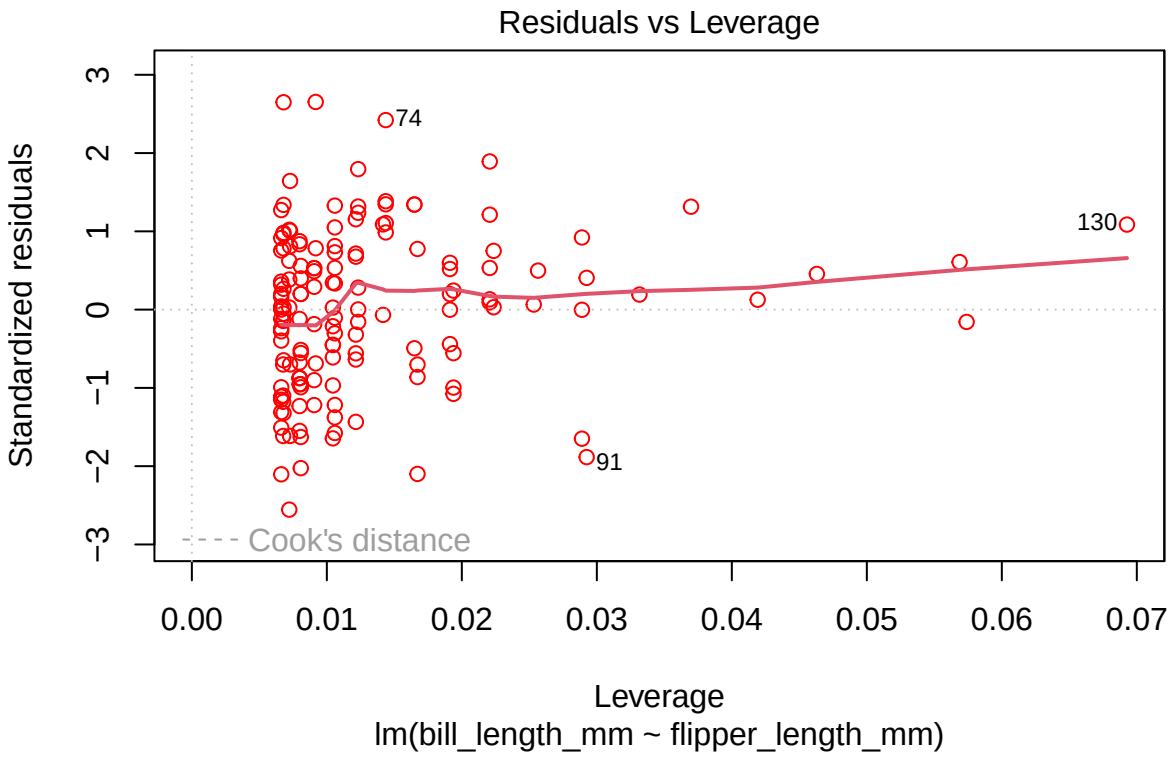
On observe une tendance linéaire positive : plus la nageoire est longue, plus le bec est long

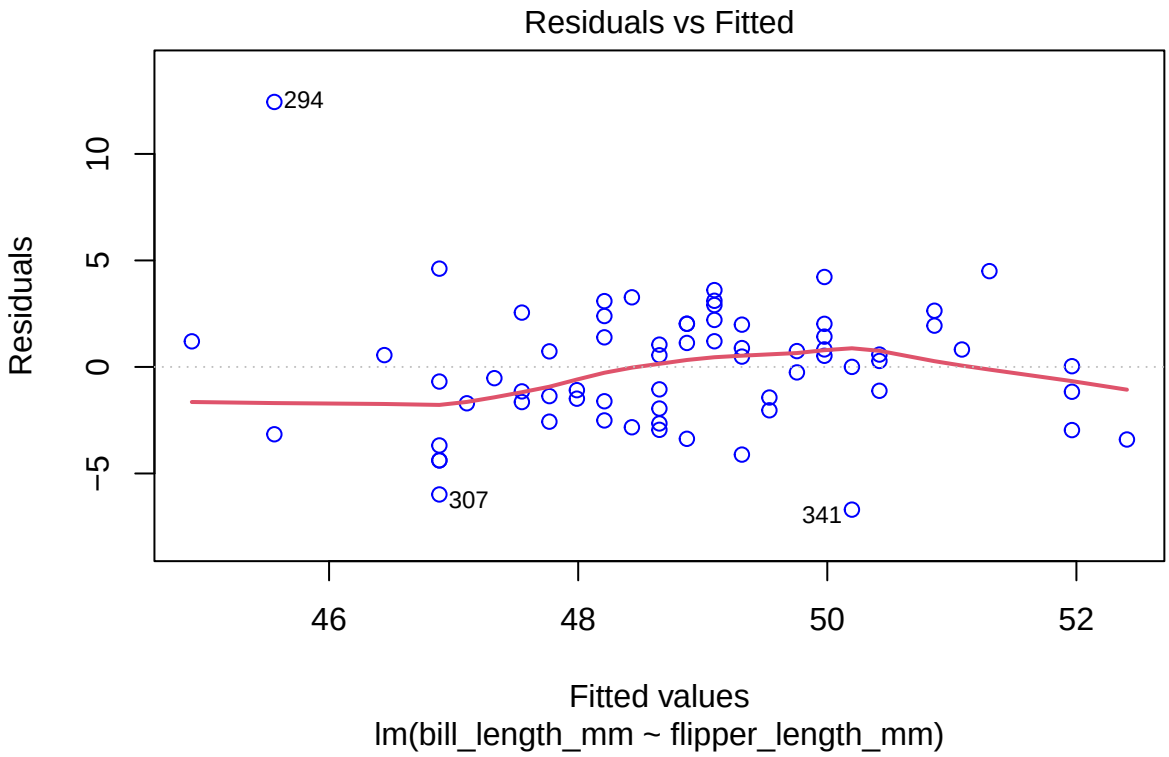
```
# Boucle sur chaque espèce pour ajuster et tracer les droites de régression
for(i in seq_along(especes)) {
  sous_donnees <- subset(df_clean, species == especes[i])
  modele_temp <- lm(bill_length_mm ~ flipper_length_mm, data = sous_donnees)
  plot(modele_temp, col=couleurs[i], lwd=2)
}
```

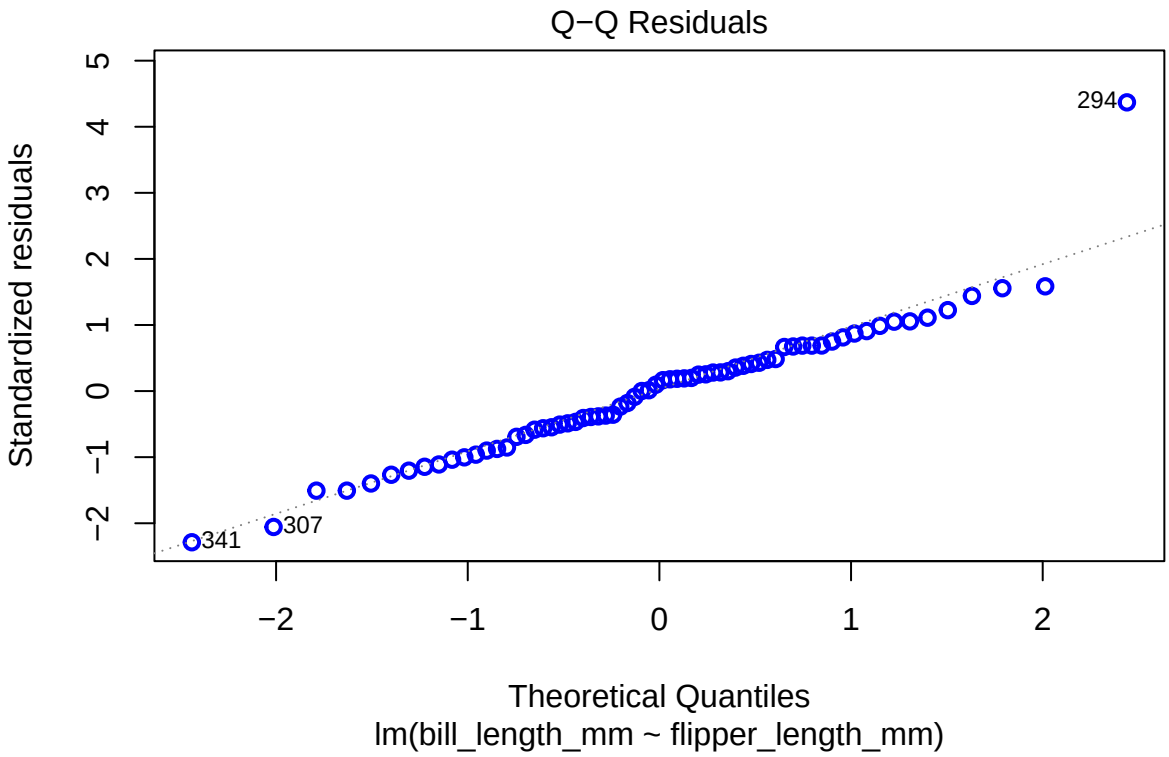

Residuals vs Fitted
1120
143
Residuals
5
0
−5
37
38
39
40
41
Fitted values
lm(bill_length_mm ~ flipper_length_mm)

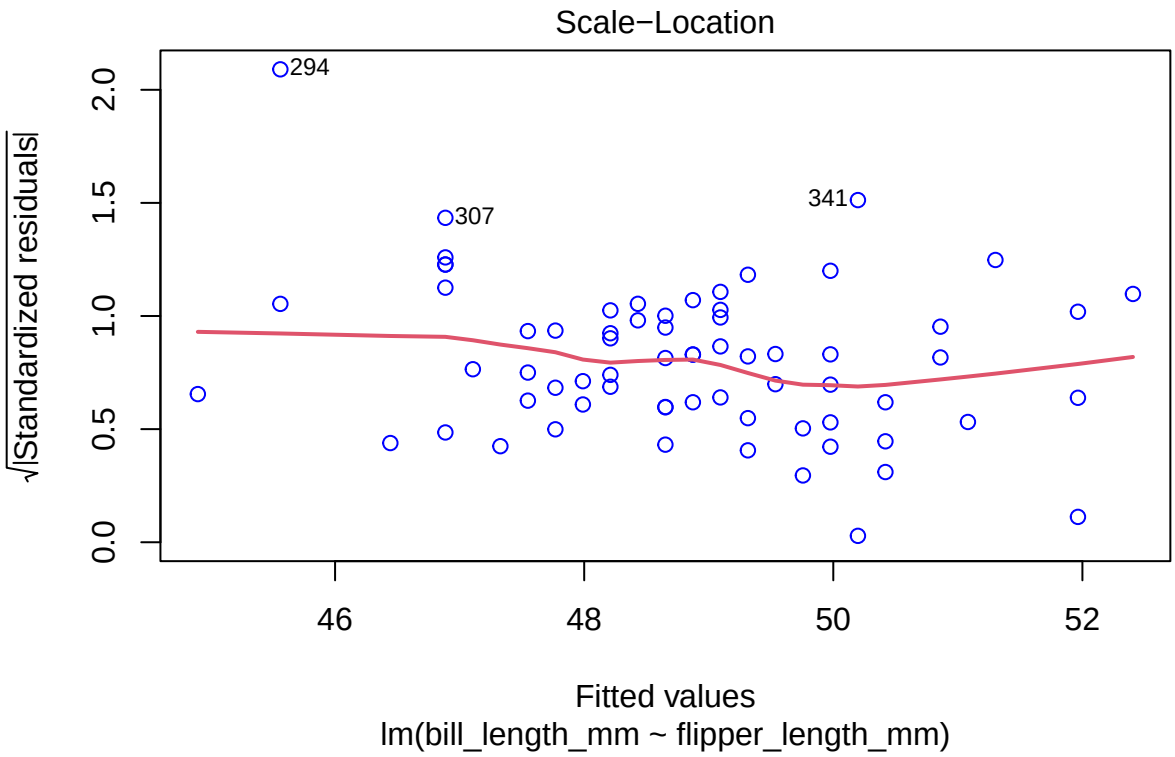

Q−Q Residuals
Standardized residuals
3
2
1
0
−1
−2
112
20
143
−2
−1
0
1
2
Theoretical Quantiles
lm(bill_length_mm ~ flipper_length_mm)

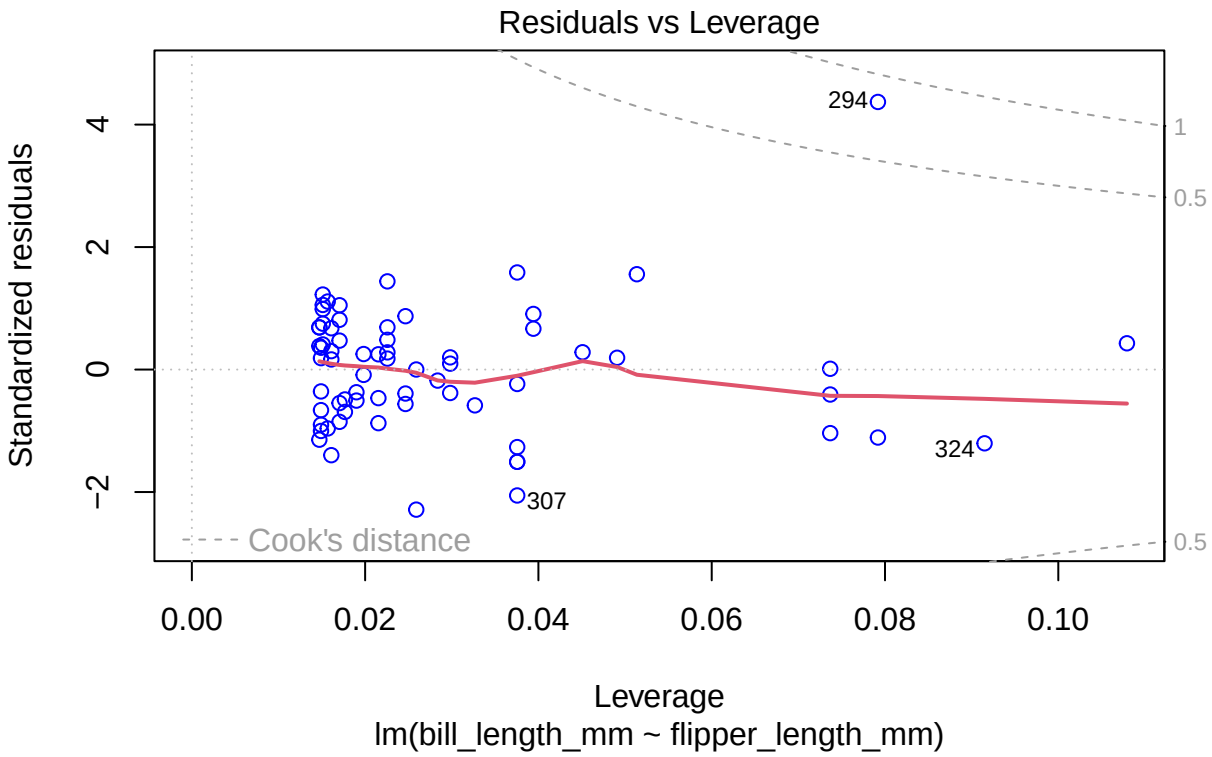

Scale−Location
143
1120
√|Standardized residuals|
0.0
0.5
1.0
1.5
37
38
39
40
41
Fitted values
lm(bill_length_mm ~ flipper_length_mm)

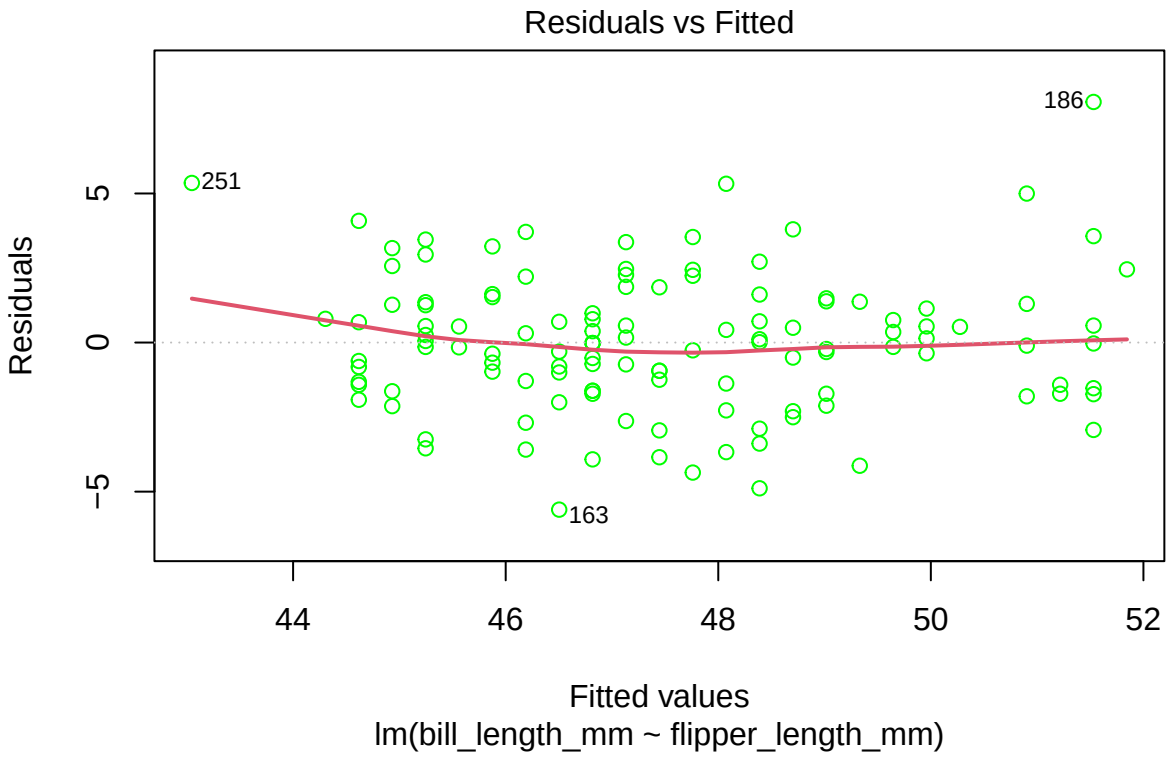

Residuals vs Leverage
Standardized residuals
3
2
1
0
−1
−2
−3
74
130
91
Cook's distance
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
Leverage
lm(bill_length_mm ~ flipper_length_mm)

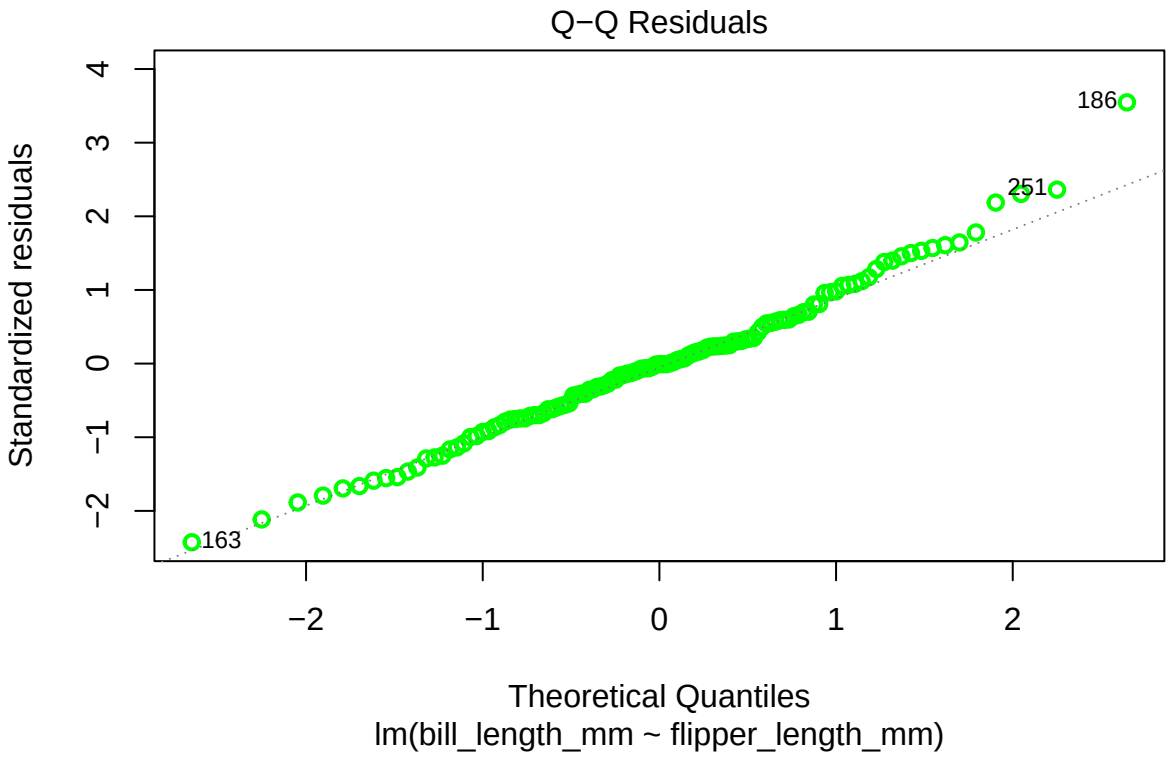

Residuals vs Fitted
294
307
341
Residuals
10
5
0
−5
46
48
50
52
Fitted values
lm(bill_length_mm ~ flipper_length_mm)

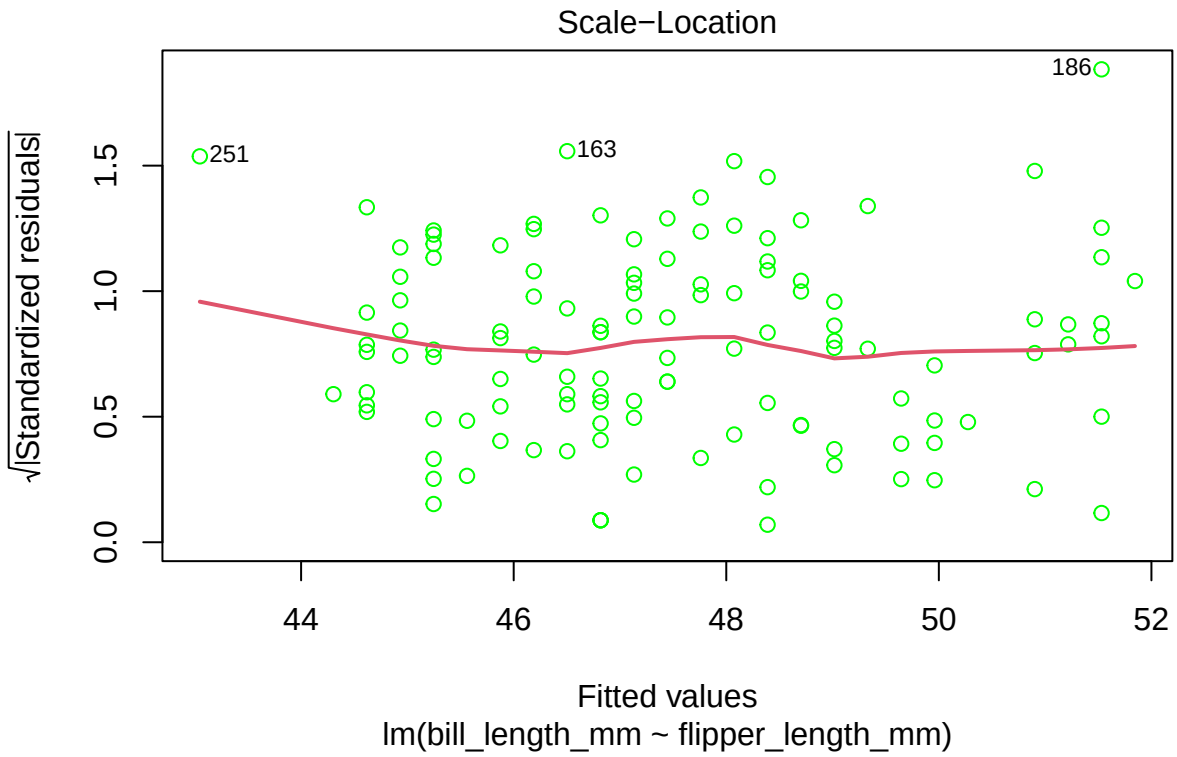

Q−Q Residuals
Standardized residuals
5
4
3
2
1
0
−1
−2
294
307
341
−2
−1
0
1
2
Theoretical Quantiles
lm(bill_length_mm ~ flipper_length_mm)


Scale−Location
294
307
341
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
46
48
50
52
Fitted values
lm(bill_length_mm ~ flipper_length_mm)

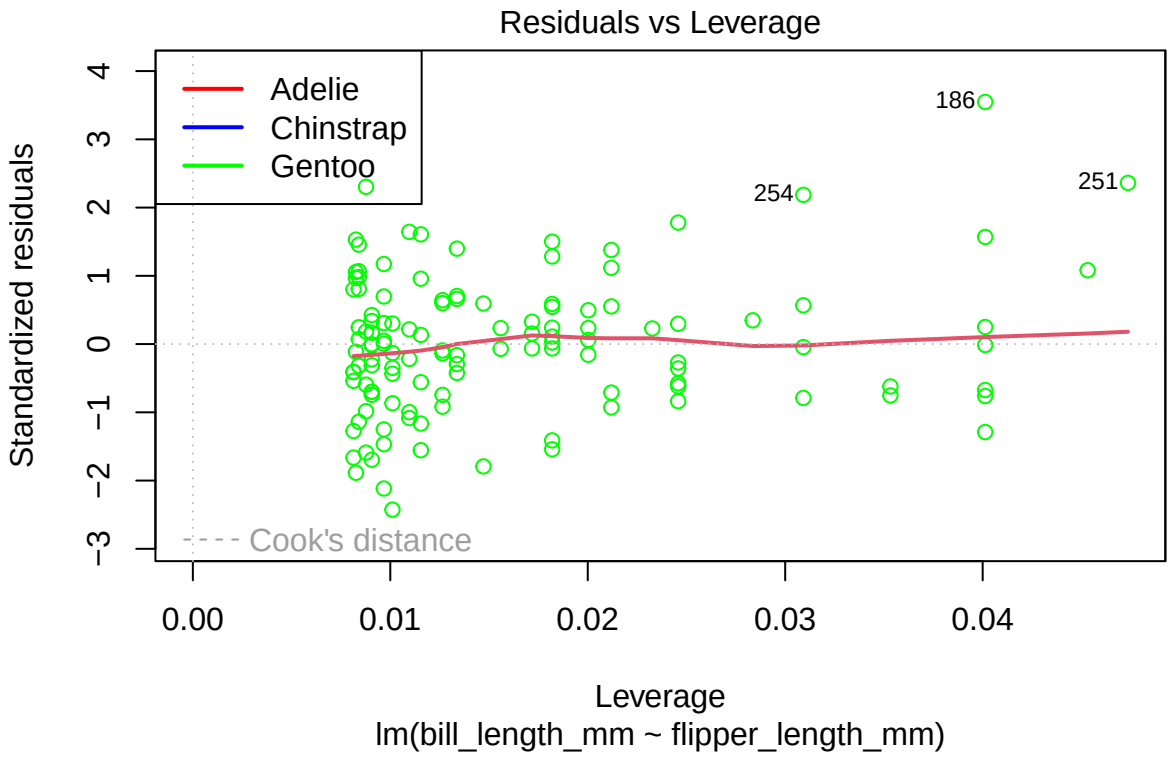

Residuals vs Leverage
Standardized residuals
4
2
0
−2
294
307
324
1
0.5
0.5
Cook's distance
0.00
0.02
0.04
0.06
0.08
0.10
Leverage
lm(bill_length_mm ~ flipper_length_mm)


Residuals vs Fitted
Residuals
Fitted values
lm(bill_length_mm ~ flipper_length_mm)
251
186
163
5
0
−5
44
46
48
50
52


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
3
4
186
251
163
−2
−1
0
1
2
Theoretical Quantiles
lm(bill_length_mm ~ flipper_length_mm)




```
legend("topleft", legend = especes, col = couleurs, lwd = 2)
```



4.2.b.Ajustement du modèle ANCOVA

```
modele_interaction <- lm(bill_length_mm ~ flipper_length_mm * species, data = df_clean)
summary(modele_interaction)
```

```
##
## Call:
## lm(formula = bill_length_mm ~ flipper_length_mm * species, data = df_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.6977 -1.7046  0.0596  1.5571 12.4394
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        13.58714    6.05061   2.246 0.025380 *
## flipper_length_mm                   0.13269    0.03183   4.168 3.91e-05 ***
## speciesChinstrap                   -7.99376   10.48117  -0.763 0.446190
## speciesGentoo                     -34.32335    9.81983  -3.495 0.000537 ***
## flipper_length_mm:speciesChinstrap  0.08813    0.05405   1.631 0.103915
## flipper_length_mm:speciesGentoo     0.18152    0.04775   3.801 0.000171 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.55 on 336 degrees of freedom
## Multiple R-squared:  0.7851, Adjusted R-squared:  0.7819
## F-statistic: 245.5 on 5 and 336 DF,  p-value: < 2.2e-16
```

```r
anova_interaction <- anova(modele_interaction)
print(anova_interaction)
```

```
## Analysis of Variance Table
##
## Response: bill_length_mm
##                             Df Sum Sq Mean Sq  F value    Pr(>F)
## flipper_length_mm            1 4376.4  4376.4 673.2170 < 2.2e-16 ***
## species                      2 3509.4  1754.7 269.9231 < 2.2e-16 ***
## flipper_length_mm:species    2   94.1    47.0   7.2354 0.0008385 ***
## Residuals                  336 2184.3     6.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

-La pente pour Gentoo (0.31421) est nettement plus élevée que celle d'Adelie (0.13269). -La pente pour Chinstrap (0.22082) est entre les deux, mais la différence n'est pas statistiquement claire au seuil 5 %. -L'intercept pour Gentoo est beaucoup plus bas que pour Adelie, ce qui peut paraître contre-intuitif, mais n'est pas forcément un problème dans la plage de valeurs de flipper_length_mm où se situe Gentoo.

La relation entre bill_length_mm et flipper_length_mm dépend de l'espèce. Gentoo présente à la fois un intercept plus faible et une pente plus forte, ce qui suggère qu'à mesure que la nageoire s'allonge, la longueur du bec augmente particulièrement vite chez Gentoo. Pour Chinstrap, la différence de pente n'est pas statistiquement claire à 5 % de signification.