Sélection d'un modèle linéaire

Marie Vaugoyeau

21 January 2024

Table of contents

1 import des packages

```
library(tidyverse)
library(palmerpenguins)
```

2 Définition de la `régression linéaire`

Objectif : Trouver une équation de type linéaire qui permet d'expliquer une **variable réponse quantitative** par **une ou plusieurs variable(s) explicative(s)**.

> **Différence entre régression linéaire et modèle linéaire**
>
> Il n'y en a pas !
> Certaines personnes parlent de **modèle de régression linéaire**.

L'équation est de la forme :

$$Y = a_1 X_1 + a_2 X_2 + ... + a_n X_n + b$$

Avec a_i : la pente (ou coefficient directeur) associé à la variable X_i et b : l'ordonnée à l'origine ou **intecept** (en anglais).

3 Les données

Les données utilisées sont celles du jeu de données `penguins` du package `{palmerpenguins}`. Plus d'information sur la page d'aide `help(penguins)`.

```
penguins |>
  glimpse()
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
$ sex               <fct> male, female, female, NA, female, male, female, male~
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

4 Réalisation d'un modèl linéaire

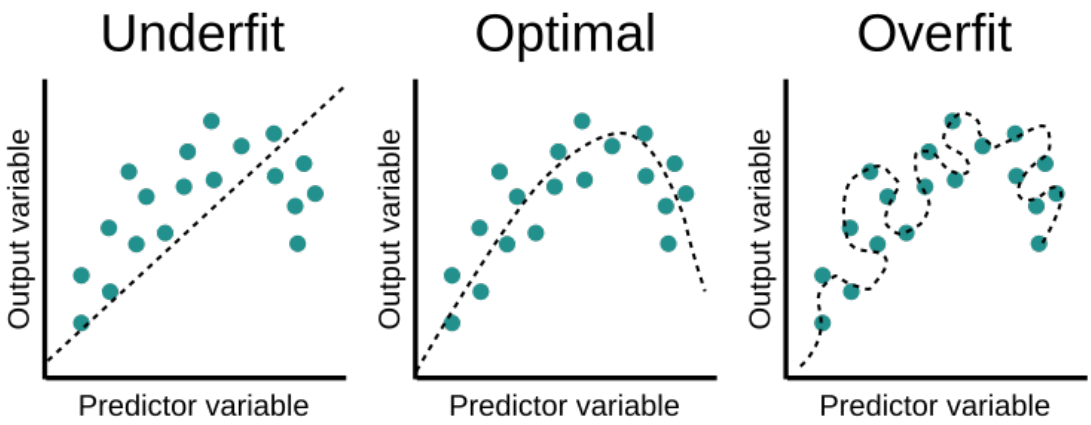
4.1 1ère étape : Choix des variables utilisées

Dans cette exemple, la **variable réponse** est `body_mass_g` et les **variables explicatives** sont les caractéristiques morphologiques mesurées : `bill_length_mm`, `bill_depth_mm` et `flipper_length_mm`.

> **Note**
>
> Les données ajoutées dans un model doit avoir un sens.
> On ne peux pas ajouter toutes les variables **juste pour voir** !

Les risques à mettre toutes les variables possibles dans un modèle :

- Impossibilité d'expliquer le modèle dans la réalité (*expl : l'âge du capitaine*)
- Sur ou sous ajustement (aussi appelé sur ou sous apprentissage et en anglais *over or underfitting*)



@educative

Sur le graphique :

- le schéma de gauche montre un **sous-ajustement**, c'est-à-dire que la droite ne prend pas en compte les variations des données et **simplifie trop**.
- le schéma du milieu montre un **bon ajustement** aux données.
- le schéma de droite montre un **sur-ajustement**. Le courbe ne permet pas de prendre en compte de nouvelles données.

4.2 2ème étape : Vérifier les limites de construction du modèle

Les données doivent être indépendantes et suivre (ou être approximées par) des lois normales.

Test de Shapiro-Wilk

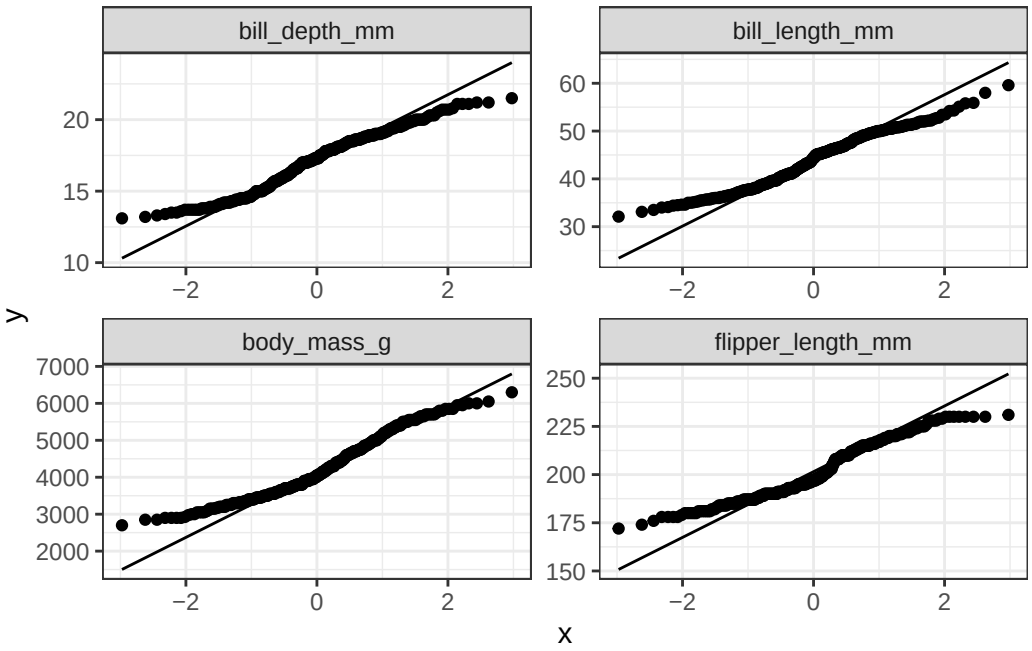
```r
map(
  .x = penguins |>
    select(where(is.numeric), - year),
  .f = shapiro.test
)
```

```
$bill_length_mm

    Shapiro-Wilk normality test

data:  .x[[i]]
W = 0.97485, p-value = 1.12e-05


$bill_depth_mm

    Shapiro-Wilk normality test

data:  .x[[i]]
W = 0.97258, p-value = 4.419e-06


$flipper_length_mm

    Shapiro-Wilk normality test

data:  .x[[i]]
W = 0.95155, p-value = 3.54e-09


$body_mass_g

    Shapiro-Wilk normality test

data:  .x[[i]]
W = 0.95921, p-value = 3.679e-08
```

> **Note**
>
> Selon le test de Shapiro-Wilk, les données ne suivent pas des lois normales

Représentation graphique

```r
penguins |>
  select(
    where(is.numeric),
    - year
  ) |>
  pivot_longer(everything()) |>
  ggplot() +
  aes(sample = value) +
  geom_qq() +
  geom_qq_line() +
  facet_wrap(~ name, scales = "free") +
  theme_bw()
```



Ici la normalité est acceptable, surtout qu'il y a bien plus de 30 données.

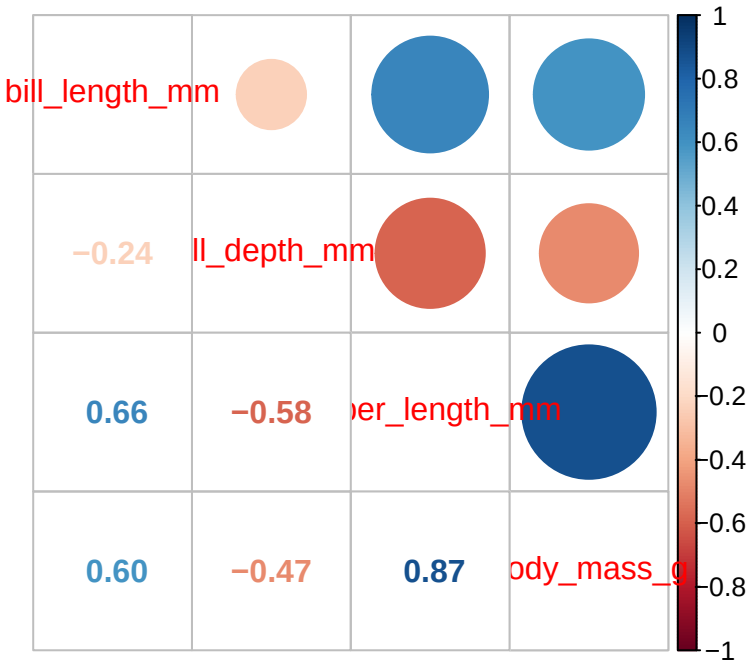
> **Note**
>
> Le modèle linéaire est assez résistant à l'absence de normalité et il est possible de le faire en prenant en compte **la loi des grands nombres**.

Si tu as déjà un modèle linéaire, tu as dû entendre parler de multicolinéarité (comme on m'a posé la question lors du live)

> ⚠ Définitions : multicolinéarité ou corrélation
>
> La colinéarité est une corrélation entre variables indépendantes.
> Quand plusieurs variables sont concernées on parle de multicolinéarité.

Ici il est intéressant de regarder la multicolinéarité même si elle est traité plus loin !

```
penguins |>
  select(where(is.numeric), - year) |>
  drop_na() |>
  cor() |>
  corrplot::corrplot.mixed()
```



4.3 3ème étape : Création du modèle linéaire

Plusieurs packages ont des fonctions qui permettent de réaliser un modèle linéaire.
Ici je vais rester sur la fonction `lm()` du package `{stats}` automatiquement chargé dans l'environnement.
Cette fonction prend comme premier argument la `formula`, c'est-à-dire la formule de type `y ~ x` et en deuxième argument `data`, le jeu de données utilisé.

```r
lm_body_mass <- lm(
  body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm,
  data = penguins
)
```

Pour accéder aux coefficients, il y a plusieurs solutions :

- Rappeler le nom du modèle : Ne donne pas les statistiques de test
- Utiliser la fonction `summary()` du package `{base}` : Le plus complet

```r
lm_body_mass
```

```
Call:
lm(formula = body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm,
    data = penguins)

Coefficients:
      (Intercept)     bill_length_mm      bill_depth_mm  flipper_length_mm
        -6424.765              4.162             20.050             50.269
```

```r
summary(lm_body_mass)
```

```
Call:
lm(formula = body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm,
    data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max
-1054.94  -290.33   -21.91   239.04  1276.64

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6424.765    561.469 -11.443   <2e-16 ***
bill_length_mm        4.162      5.329   0.781    0.435
bill_depth_mm        20.050     13.694   1.464    0.144
flipper_length_mm    50.269      2.477  20.293   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.4 on 338 degrees of freedom
```

```
  (2 observations effacées parce que manquantes)
Multiple R-squared:  0.7615,    Adjusted R-squared:  0.7594
F-statistic: 359.7 on 3 and 338 DF,  p-value: < 2.2e-16
```

Pour aller plus loin :

- Utilisation de la fonction `anova()` du package `{stats}` : Permet d'afficher facilement le tableau des coefficients
- Prendre la fonction `Anova()` du package `{car}` : Même chose que précédent mais type II (et même III s'il y a une interaction)

```r
anova(lm_body_mass)
```

```
Analysis of Variance Table

Response: body_mass_g
                   Df   Sum Sq  Mean Sq F value    Pr(>F)
bill_length_mm      1 77669072 77669072  501.84 < 2.2e-16 ***
bill_depth_mm       1 25591770 25591770  165.36 < 2.2e-16 ***
flipper_length_mm   1 63735497 63735497  411.81 < 2.2e-16 ***
Residuals         338 52311359   154767
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
car::Anova(lm_body_mass)
```

```
Anova Table (Type II tests)

Response: body_mass_g
                    Sum Sq  Df  F value Pr(>F)
bill_length_mm       94393   1   0.6099 0.4354
bill_depth_mm       331766   1   2.1436 0.1441
flipper_length_mm 63735497   1 411.8149 <2e-16 ***
Residuals         52311359 338
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4.4 4ème étape : Validation du modèle

Le modèle est accepté si les **résidus** suivent une **loi normale**.

```
lm_body_mass$residuals |>
  shapiro.test()
```

```

    Shapiro-Wilk normality test

data:  lm_body_mass$residuals
W = 0.99368, p-value = 0.164
```

Les résidus suivent une loi normale (`p-valeur` > 0.05 -> impossible de rejeter l'hypothèse nulle selon laquelle les données suivent une loi normale).

Il est aussi bien de visualiser le modèle grâce à la fonction `plot()`.

```
plot(lm_body_mass)
```

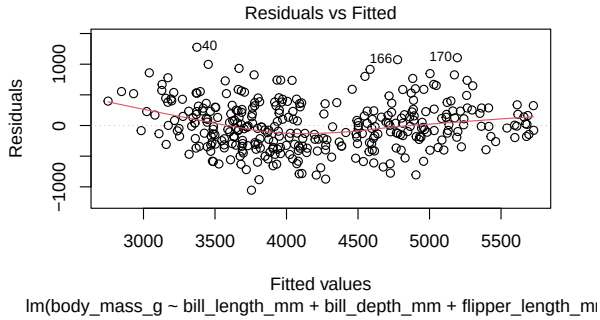
Et la multicolinéarité ?

```
car::vif(lm_body_mass)
```

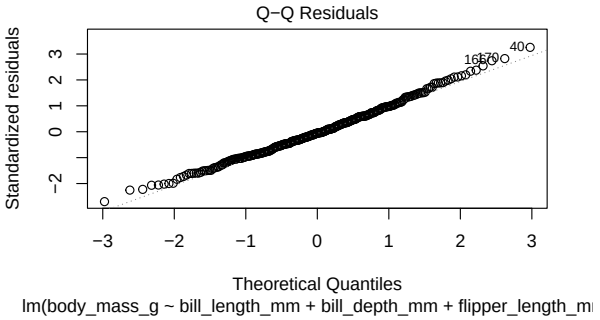
```
   bill_length_mm     bill_depth_mm flipper_length_mm
         1.865090          1.611292          2.673338
```

Il y a pas de multicolinéarité lorsque les facteurs d'inflation de la variance (en anglais *variance inflation factor (VIF)*) sont à `1`.

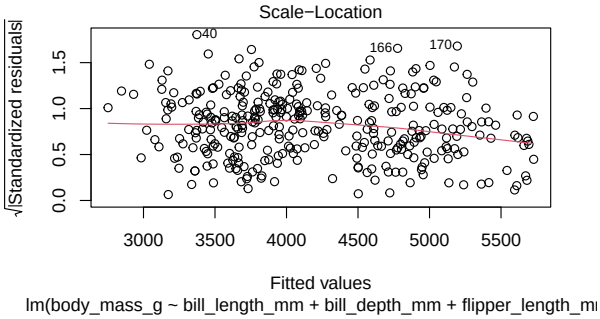
> **Influence de la multicolinéarité**
>
> Si les **FIV** sont supérieurs à 1, la variable est corrélée aux autre et son influence est "augmentée".
> A quel valeur est-ce grave ?
> Pour Paul ALLISON au delà de `2,5` c'est un signe d'inquiétude. Pour d'autres personnes, c'est à partir de `5`.
> **Mon conseil** : Simplifions le modèle et voyons après !



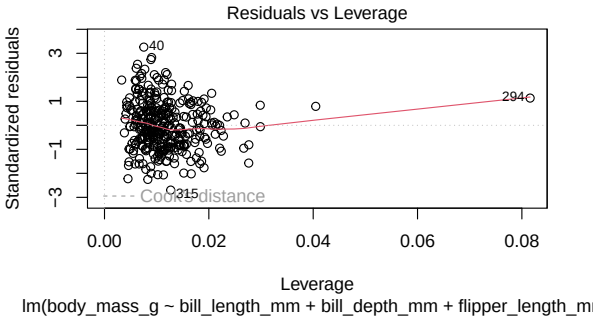
(a) La courbe rouge doit être la plus proche de la droite en pointillée



(a) Les points doivent suivre la première diagonale en pointillée



(a) La courbe rouge doit être la plus plate possible



(a) La courbe rouge doit être proche de la droite horizontale en pointillée

4.5 5ème étape : Sélection de modèle

Ici, réalisation d'une sélection descendante qui revient à supprimer les variables les moins significatives.

```
lm_body_mass_2 <- lm(
  body_mass_g ~ bill_depth_mm + flipper_length_mm,
  data = penguins
)

summary(lm_body_mass_2)
```

```
Call:
lm(formula = body_mass_g ~ bill_depth_mm + flipper_length_mm,
    data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max
-1029.78  -271.45   -23.58   245.15  1275.97

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -6541.907    540.751 -12.098   <2e-16 ***
bill_depth_mm        22.634     13.280   1.704   0.0892 .
flipper_length_mm    51.541      1.865  27.635   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 393.2 on 339 degrees of freedom
  (2 observations effacées parce que manquantes)
Multiple R-squared:  0.761, Adjusted R-squared:  0.7596
F-statistic: 539.8 on 2 and 339 DF,  p-value: < 2.2e-16
```

```
anova(lm_body_mass, lm_body_mass_2)
```

```
Analysis of Variance Table

Model 1: body_mass_g ~ bill_length_mm + bill_depth_mm + flipper_length_mm
Model 2: body_mass_g ~ bill_depth_mm + flipper_length_mm
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1    338 52311359
2    339 52405752 -1    -94393 0.6099 0.4354
```

```r
AIC(lm_body_mass, lm_body_mass_2)
```

```
               df      AIC
lm_body_mass    5 5063.320
lm_body_mass_2  4 5061.937
```

Pour comparer deux modèles, j'utilise ici l'AIC.

> **i** AIC : Critère d'Information d'Akaike (en anglais *Akaike information criterion*)
>
> Permet de comparer deux modèles proches (même données et une ou deux variables en plus ou en moins) pour choisir le plus significatif, c'est-à-dire celui qui a la la valeur d'AIC la plus faible.
> **Attention** : si la différence est inférieure à 2, il faut faire le choix de parcimonie, c'est-à-dire de préférer le modèle le plus simple (avec le moins de variables explicatives).

```r
lm_body_mass_3 <- lm(
  body_mass_g ~ flipper_length_mm,
  data = penguins
)

summary(lm_body_mass_3)
```

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm, data = penguins)

Residuals:
     Min       1Q   Median       3Q      Max
-1058.80  -259.27   -26.88   247.33  1288.69

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5780.831    305.815  -18.90   <2e-16 ***
flipper_length_mm    49.686      1.518   32.72   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 394.3 on 340 degrees of freedom
  (2 observations effacées parce que manquantes)
Multiple R-squared:  0.759, Adjusted R-squared:  0.7583
F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

```r
anova(lm_body_mass_2, lm_body_mass_3)
```

```
Analysis of Variance Table

Model 1: body_mass_g ~ bill_depth_mm + flipper_length_mm
Model 2: body_mass_g ~ flipper_length_mm
  Res.Df      RSS Df Sum of Sq      F  Pr(>F)
1    339 52405752
2    340 52854796 -1   -449044 2.9048 0.08924 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
AIC(lm_body_mass_2, lm_body_mass_3)
```

```
               df      AIC
lm_body_mass_2  4 5061.937
lm_body_mass_3  3 5062.855
```

Le modèle le plus simple avec juste la longueur de la nageoire serait meilleur.

```r
lm_body_mass_3$residuals |>
  shapiro.test()
```

```
    Shapiro-Wilk normality test

data:  lm_body_mass_3$residuals
W = 0.99301, p-value = 0.1123
```

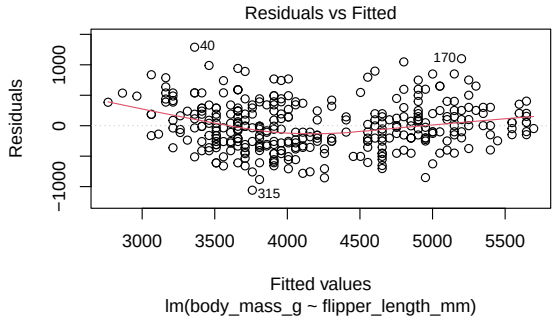
Les résidus suivants une loi normale, le modèle est validé.

```r
plot(lm_body_mass_3)
```

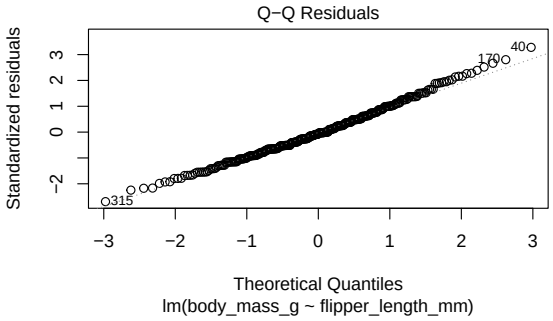
Les sorties graphiques de la fonction `plot()` valide le modèle aussi.

Il ne reste donc plus qu'à valoriser le modèle trouvé via un graphique.

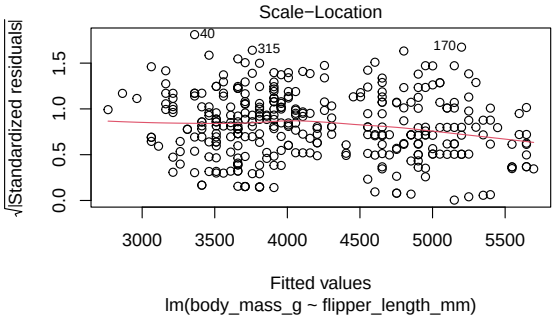
```r
ggplot(penguins) +
  aes(x = flipper_length_mm, y = body_mass_g) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  theme_classic()
```



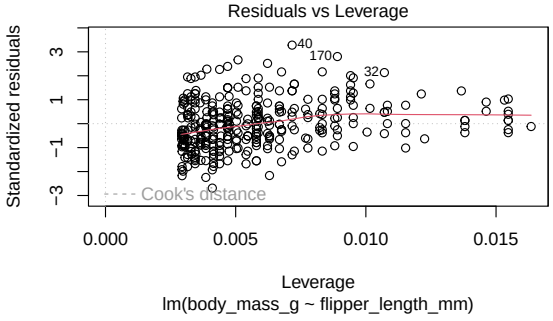
(a) La courbe rouge doit être la plus proche de la droite en pointillée



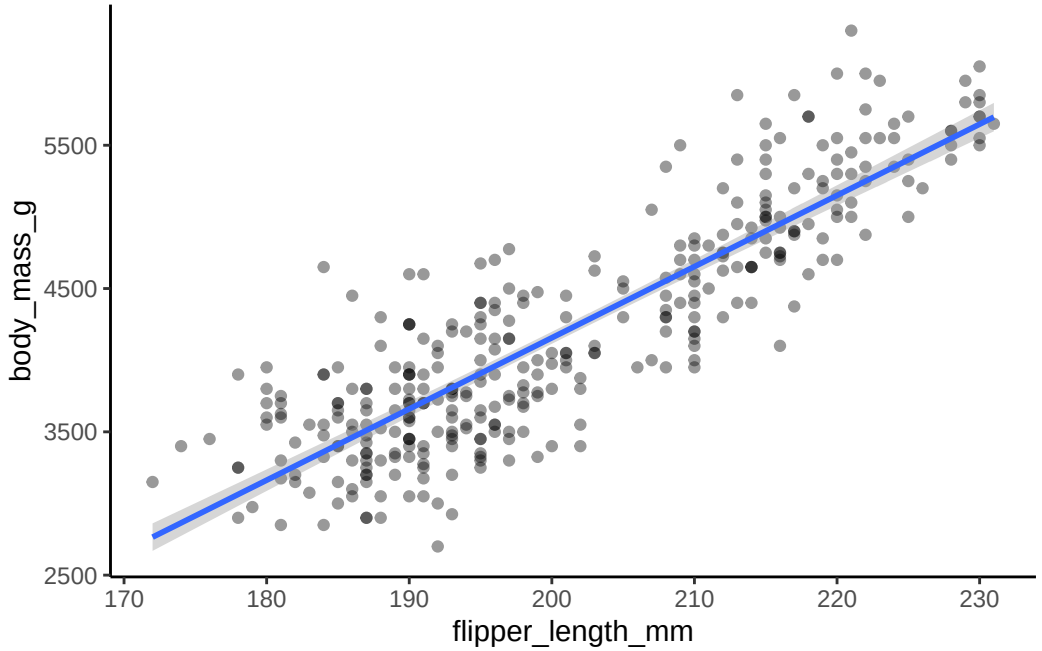
(a) Les points doivent suivre la première diagonale en pointillée



(a) La courbe rouge doit être la plus plate possible

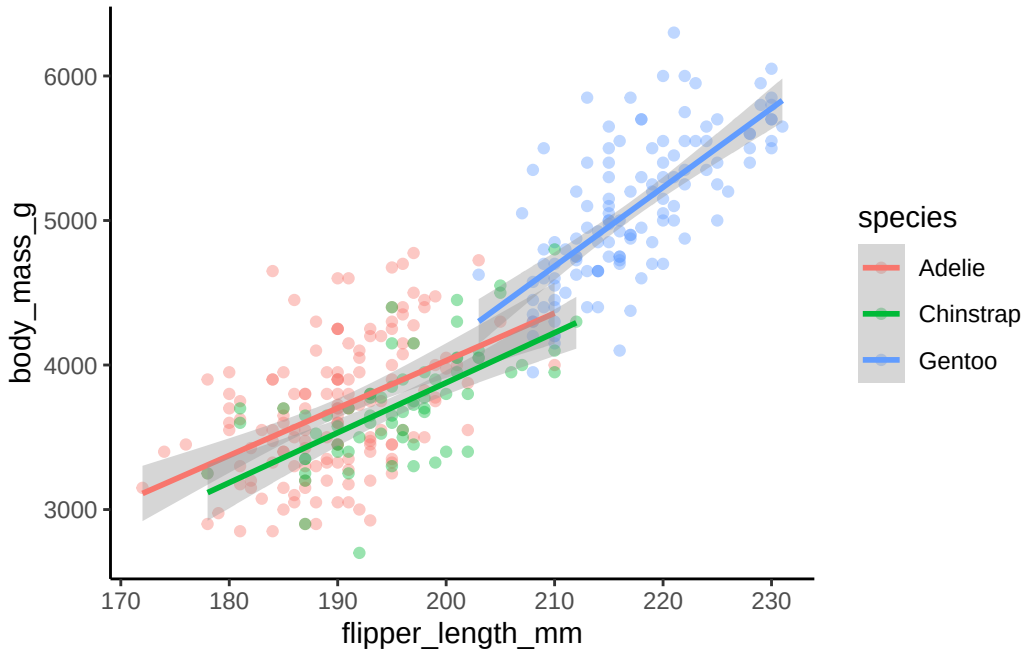


(a) La courbe rouge doit être proche de la droite horizontale en pointillée



Sur le graphique il semble apparaître "2 groupes", **un avec une nageoire de moins de 205 mm** et **un avec plus**.
Il est possible d'explorer graphiquement cette idée en ajoutant l'espèce en couleur.

```
ggplot(penguins) +
  aes(x = flipper_length_mm, y = body_mass_g, colour = species) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  theme_classic()
```



Création d'un modèle avec une interaction espèce et longueur de la nageoire. C'est à dire que l'espèce influence le coefficient directeur associée à la longueur de la nageoire comme vu sur le graphique.
L'interaction est représenté par `:` mais comme les effets simples doivent être présent dans le modèle, il faut utiliser `*` ainsi `A * B = A + B + A:B` avec `A` et `B` sont les effets simples qui ne doivent pas être supprimé du modèle si l'interaction est significative et `A:B` est l'interaction.

```r
lm_body_mass_4 <- lm(
  body_mass_g ~ flipper_length_mm * species,
  data = penguins
)

summary(lm_body_mass_4)
```

```
Call:
lm(formula = body_mass_g ~ flipper_length_mm * species, data = penguins)

Residuals:
    Min      1Q  Median      3Q     Max
-911.18 -251.93  -31.77  197.82 1144.81

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                          -2535.837    879.468  -2.883  0.00419 **
flipper_length_mm                       32.832      4.627   7.095 7.69e-12 ***
speciesChinstrap                      -501.359   1523.459  -0.329  0.74229
speciesGentoo                        -4251.444   1427.332  -2.979  0.00311 **
flipper_length_mm:speciesChinstrap       1.742      7.856   0.222  0.82467
flipper_length_mm:speciesGentoo         21.791      6.941   3.139  0.00184 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 370.6 on 336 degrees of freedom
  (2 observations effacées parce que manquantes)
Multiple R-squared:  0.7896,    Adjusted R-squared:  0.7864
F-statistic: 252.2 on 5 and 336 DF,  p-value: < 2.2e-16
```

```
car::Anova(lm_body_mass_4)
```

```
Anova Table (Type II tests)

Response: body_mass_g
                           Sum Sq  Df F value    Pr(>F)
flipper_length_mm         24776495   1 180.398 < 2.2e-16 ***
species                    5187807   2  18.886 1.686e-08 ***
flipper_length_mm:species  1519564   2   5.532  0.004327 **
Residuals                 46147424 336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La fonction `Anova()` du package `{car}` nous permet de voir que l'interaction est significative.

```
AIC(lm_body_mass_3, lm_body_mass_4)
```

```
               df      AIC
lm_body_mass_3  3 5062.855
lm_body_mass_4  7 5024.443
```

Selon l'AIC, le modèle avec l'interaction est beaucoup plus intéressant que le modèle simple avec que la longueur de la nageoire.

Pour connaître la différence entre les espèces il faut faire un test post-hoc pour effectuer une comparaison multiple, ici un test post-hoc de Tukey.

> ⚠ Attention
>
> Un test post-hoc ne se réalise **que** si la variable concernée est **significative** dans le modèle !

```
library(multcomp)

summary(glht(lm_body_mass_4, linfct = mcp(species="Tukey")))
```

```
     Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Tukey Contrasts


Fit: lm(formula = body_mass_g ~ flipper_length_mm * species, data = penguins)

Linear Hypotheses:
                        Estimate Std. Error t value Pr(>|t|)
Chinstrap - Adelie == 0   -501.4     1523.5  -0.329  0.94178
Gentoo - Adelie == 0     -4251.4     1427.3  -2.979  0.00861 **
Gentoo - Chinstrap == 0  -3750.1     1676.7  -2.237  0.06618 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Adjusted p values reported -- single-step method)
```

> ℹ Note
>
> Pour résumer, `Chinstrap` et `Adelie` sont similaires.
> `Gentoo` est significativement différent de `Adelie` ($p\text{-}valeur < 0.01$) et légèrement différent de `Chinstrap` ($p\text{-}valeur = 0.07$).
> Si tu as l'habitude d'utiliser les lettres, `Gentoo` est `a`, `Adelie` est `b` et `Chinstrap` est `ab`.

5 En savoir un peu plus sur moi

Bonjour,

Je suis Marie Vaugoyeau et je suis disponible pour des **missions en freelance** d'**accompagnement à la formation** à R et à l'analyse de données et/ou en **programmation** (reprise de scripts, bonnes pratiques de codage, développement de package).

Ayant un **bagage recherche en écologie**, j'ai accompagné plusieurs chercheuses en biologie dans leurs analyses de données mais je suis ouverte à d'autres domaines.

Vous pouvez retrouver mes offres ici.

En plus de mes missions de consulting je diffuse mes savoirs en R et analyse de données sur plusieurs plateformes :

- J'ai écrit un **livre** aux éditions ENI
- Tous les mois je fais un **live sur Twitch** pour parler d'un package de R, d'une analyse
- Je rédige une **newsletter** de manière irrégulière pour parler de mes **inspirations** et transmettre **des trucs et astuces sur R**. Pour s'y inscrire, c'est par là. J'ai aussi un **blog** sur lequel vous pourrez retrouver une version de cet article.

Pour en savoir encore un peu plus sur moi, il y a LinkedIn et pour retrouver tous ces liens et plus encore, c'est ici

N'hésitez pas à me contacter sur marie.vaugoyeau@gmail.com !

Bonne journée

Marie