Penguin Assignment - Question 4

2022-11-26

Loading packages and visualising the data

Install and load packages, load the data (palmerpenguins):

```
library(ggplot2)
library(janitor)
```

```
##
## Attaching package: 'janitor'

## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)
```

```
##
## 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(palmerpenguins)
```

Set the working directory in files (can also set working directory using the command 'setwd()').

Visualising the data:

```
summary(penguins_raw)
```

```
##   studyName         Sample Number      Species             Region
##  Length:344         Min.   :  1.00   Length:344         Length:344
##  Class :character   1st Qu.: 29.00   Class :character   Class :character
##  Mode  :character   Median : 58.00   Mode  :character   Mode  :character
##                     Mean   : 63.15
```

```
##                     3rd Qu.: 95.25
##                     Max.   :152.00
##
##     Island             Stage           Individual ID      Clutch Completion
##  Length:344         Length:344         Length:344         Length:344
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Date Egg          Culmen Length (mm) Culmen Depth (mm) Flipper Length (mm)
##  Min.   :2007-11-09   Min.   :32.10      Min.   :13.10     Min.   :172.0
##  1st Qu.:2007-11-28   1st Qu.:39.23      1st Qu.:15.60     1st Qu.:190.0
##  Median :2008-11-09   Median :44.45      Median :17.30     Median :197.0
##  Mean   :2008-11-27   Mean   :43.92      Mean   :17.15     Mean   :200.9
##  3rd Qu.:2009-11-16   3rd Qu.:48.50      3rd Qu.:18.70     3rd Qu.:213.0
##  Max.   :2009-12-01   Max.   :59.60      Max.   :21.50     Max.   :231.0
##                       NA's   :2          NA's   :2         NA's   :2
##  Body Mass (g)      Sex            Delta 15 N (o/oo) Delta 13 C (o/oo)
##  Min.   :2700   Length:344         Min.   : 7.632    Min.   :-27.02
##  1st Qu.:3550   Class :character   1st Qu.: 8.300    1st Qu.:-26.32
##  Median :4050   Mode  :character   Median : 8.652    Median :-25.83
##  Mean   :4202                      Mean   : 8.733    Mean   :-25.69
##  3rd Qu.:4750                      3rd Qu.: 9.172    3rd Qu.:-25.06
##  Max.   :6300                      Max.   :10.025    Max.   :-23.79
##  NA's   :2                         NA's   :14        NA's   :13
##    Comments
##  Length:344
##  Class :character
##  Mode  :character
##
##
##
##
```

```
head(penguins_raw)
```

```
## # A tibble: 6 x 17
##   study~1 Sampl~2 Species Region Island Stage Indiv~3 Clutc~4 `Date Egg` Culme~5
##   <chr>     <dbl> <chr>   <chr>  <chr>  <chr> <chr>   <chr>   <date>       <dbl>
## 1 PAL0708       1 Adelie~ Anvers Torge~ Adul~ N1A1    Yes     2007-11-11    39.1
## 2 PAL0708       2 Adelie~ Anvers Torge~ Adul~ N1A2    Yes     2007-11-11    39.5
## 3 PAL0708       3 Adelie~ Anvers Torge~ Adul~ N2A1    Yes     2007-11-16    40.3
## 4 PAL0708       4 Adelie~ Anvers Torge~ Adul~ N2A2    Yes     2007-11-16    NA
## 5 PAL0708       5 Adelie~ Anvers Torge~ Adul~ N3A1    Yes     2007-11-16    36.7
## 6 PAL0708       6 Adelie~ Anvers Torge~ Adul~ N3A2    Yes     2007-11-16    39.3
## # ... with 7 more variables: `Culmen Depth (mm)` <dbl>,
## #   `Flipper Length (mm)` <dbl>, `Body Mass (g)` <dbl>, Sex <chr>,
## #   `Delta 15 N (o/oo)` <dbl>, `Delta 13 C (o/oo)` <dbl>, Comments <chr>, and
## #   abbreviated variable names 1: studyName, 2: `Sample Number`,
## #   3: `Individual ID`, 4: `Clutch Completion`, 5: `Culmen Length (mm)`
```

```
names(penguins_raw)
```

```
##  [1] "studyName"           "Sample Number"       "Species"
##  [4] "Region"              "Island"              "Stage"
##  [7] "Individual ID"       "Clutch Completion"   "Date Egg"
## [10] "Culmen Length (mm)"  "Culmen Depth (mm)"   "Flipper Length (mm)"
## [13] "Body Mass (g)"       "Sex"                 "Delta 15 N (o/oo)"
## [16] "Delta 13 C (o/oo)"   "Comments"
```

Cleaning the data

I first need to make a safe copy of the raw data by making a new folder within the working directory.

```
write.csv(penguins_raw, paste0("data_raw/penguins_raw.csv"))
```

To avoid overwriting the code, I need to re-read in the raw data:

```
penguins_raw <- read.csv("data_raw/penguins_raw.csv")
```

Use piping to avoid overwriting the code.

I will now make a function called cleaning (and save this to a separate r script):

```
cleaning <- function(penguins_raw){
  penguins_raw %>%
  select(-starts_with("Delta")) %>%
select(-Comments) %>%
  clean_names()}

remove_empty_flipper_length_and_mass <- function(penguins_clean){
  penguins_clean %>%
    filter(!is.na(flipper_length_mm)) %>%
    filter(!is.na(body_mass_g)) %>%
    select(body_mass_g, flipper_length_mm, species)
}
```

I can now now clean my data using the following code:

```
source("functions/cleaning.r")
```

I can now apply these functions to the penguins_raw dataset:

```
penguins_clean <- cleaning(penguins_raw)
penguins_now_clean <- remove_empty_flipper_length_and_mass(penguins_clean)

write.csv(penguins_clean, paste0("data_clean/penguins_clean.csv"))
write.csv(penguins_now_clean, paste0("data_clean/penguins_now_clean.csv"))
```

```
names(penguins_now_clean)
```

```
## [1] "body_mass_g"       "flipper_length_mm" "species"
```

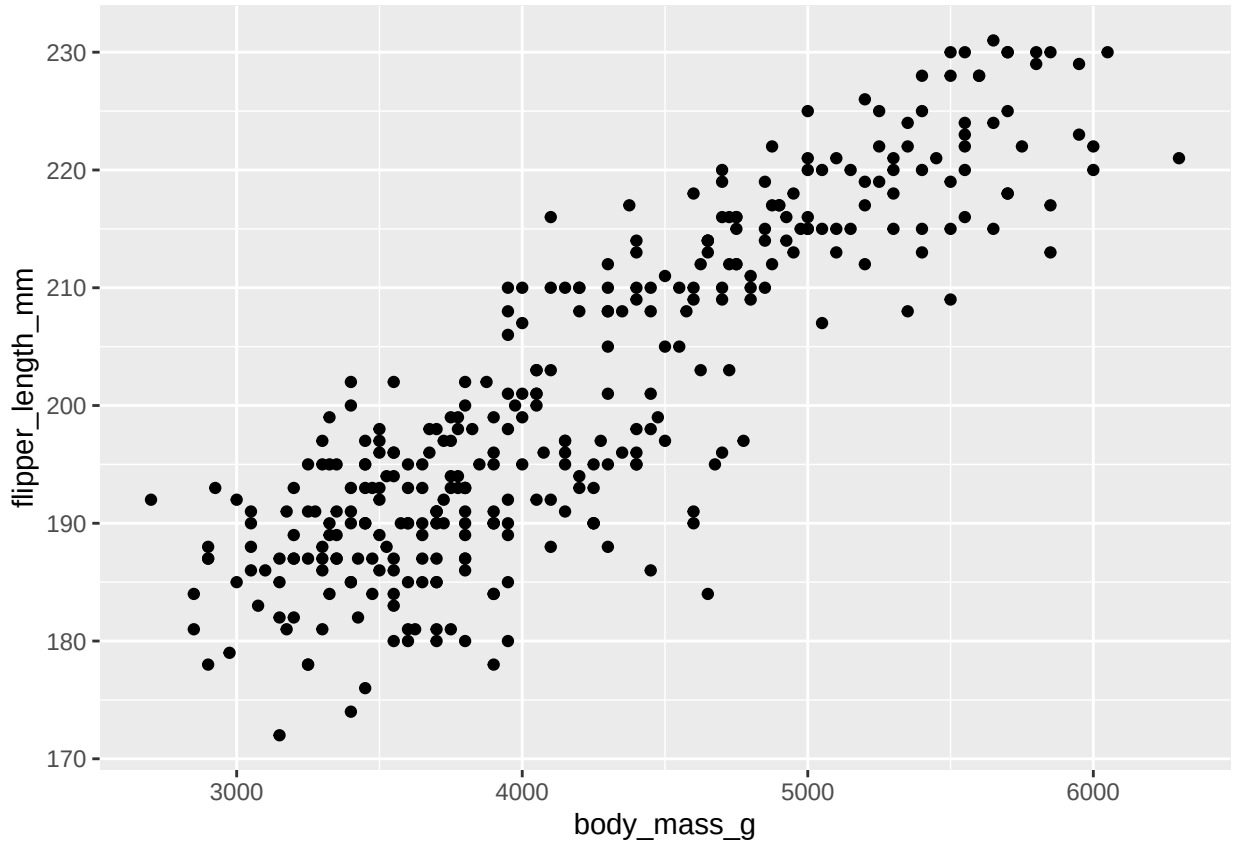
Make a statistical test

I am going to test whether body mass predicts flipper length in penguins.

H0: body mass does not predict flipper length in penguins H1: body mass does predict flipper length in penguins

To visualise my data, first I am going to plot a scatter graph:

```
ggplot(data=penguins_now_clean, aes(x=body_mass_g, y=flipper_length_mm)) + geom_point()
```



Next, I am going to make a linear regression

```
penguins_model <- lm(flipper_length_mm ~ body_mass_g, data=penguins_now_clean)
```

```
summary(penguins_model)
```

```
##
## Call:
```

```
## lm(formula = flipper_length_mm ~ body_mass_g, data = penguins_now_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.7626  -4.9138   0.9891   5.1166  16.6392
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.367e+02  1.997e+00   68.47   <2e-16 ***
## body_mass_g 1.528e-02  4.668e-04   32.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.913 on 340 degrees of freedom
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.7583
## F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

The R2 value shows us that 75.83% of variance in flipper length is explained by body mass. Furthermore, we can see that the y-intercept for the slope is 1.367e+02, whilst the slope is 1.528e-02.

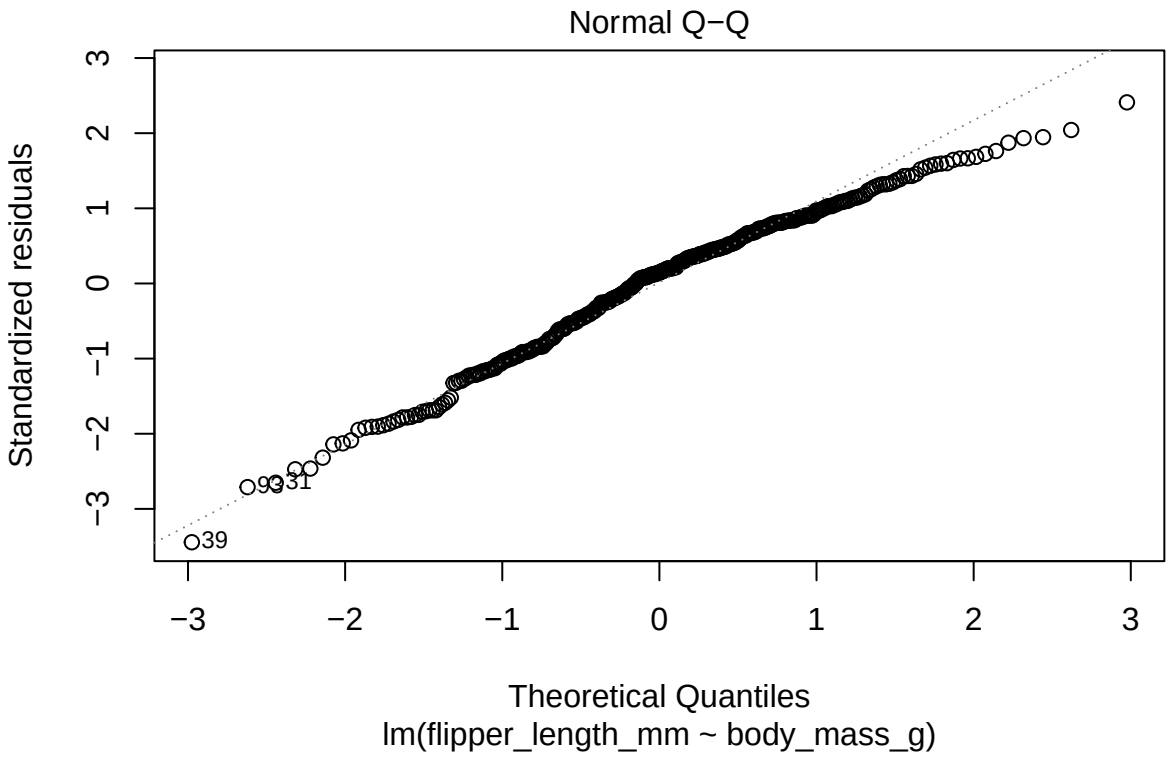
I will then find the confidence intervals of these estimates using the following code:

```
confint(penguins_model)
```

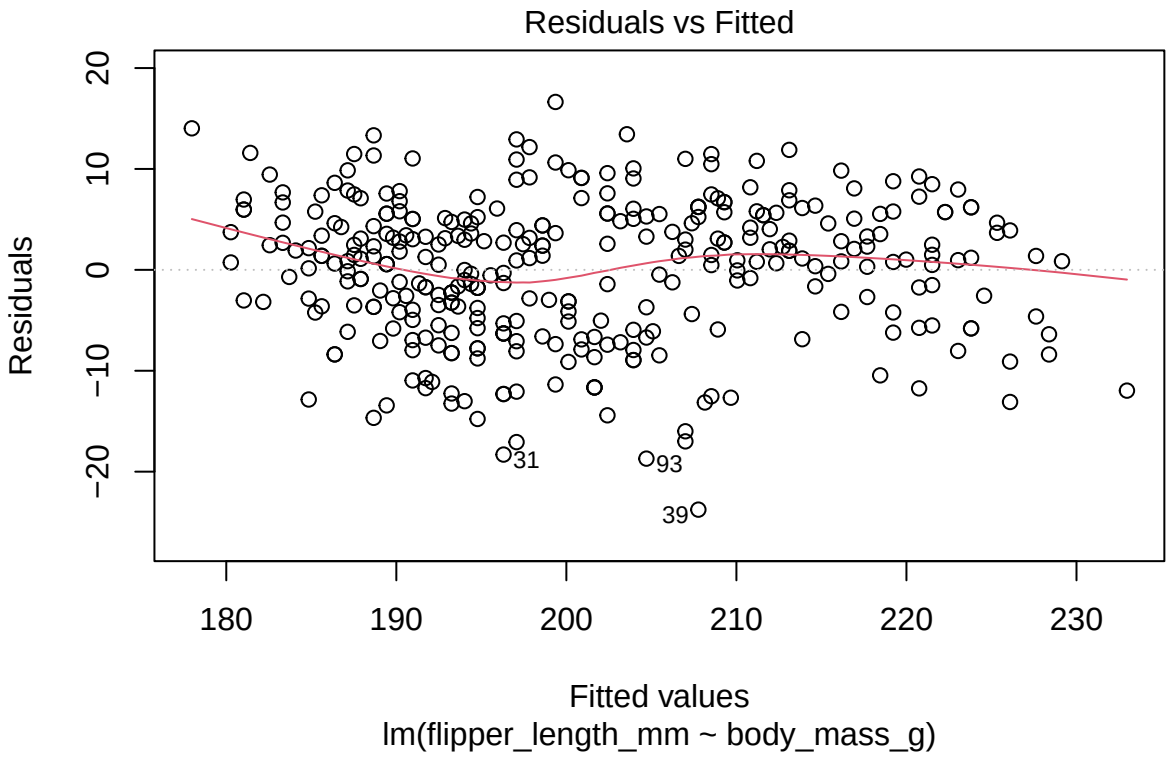
```
##                    2.5 %       97.5 %
## (Intercept) 132.80185245 140.65726609
## body_mass_g   0.01435767   0.01619417
```

Next, I am going to test my assumptions of normality and homogeneity of variance using a qqplot and a residuals vs fitted plot:

```
plot(penguins_model, which=2)
```




```
plot(penguins_model, which=1)
```



Using visual assessment, points in the qq plot fall mostly along the dashed line, therefore the assumption of normality is well met. In the residuals vs fitted plot, the red line falls mostly along the dashed line and the residuals are arranged randomly around the line, therefore the assumption of homogeneity of variance is well met. It is therefore appropriate to conduct an anova statistical test.

Next, I want to test the statistical significance of my regression line using ANOVA.

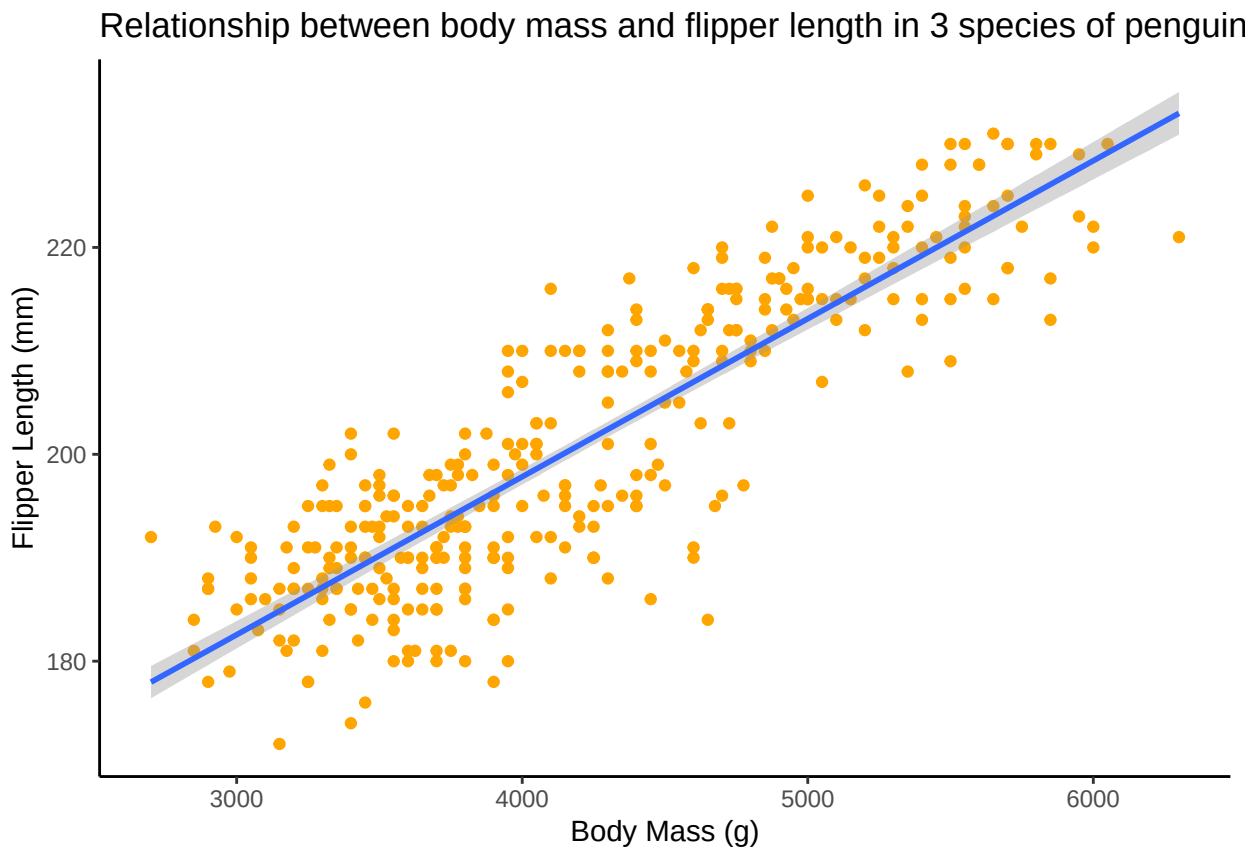
```
anova(penguins_model)
```

```
## Analysis of Variance Table
##
## Response: flipper_length_mm
##              Df Sum Sq Mean Sq F value    Pr(>F)
## body_mass_g   1  51176   51176  1070.7 < 2.2e-16 ***
## Residuals   340  16250      48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of 2.2e-16 is less than 0.05, and so I can reject the null hypothesis. Therefore, there is a significant effect of penguin species on body mass. I will now create a figure to illustrate this.

```
ggplot(data = penguins_now_clean, aes(x = body_mass_g, y = flipper_length_mm)) +
  geom_point(colour = "orange") + geom_smooth(method="lm") +
  labs(x="Body Mass (g)", y="Flipper Length (mm)",
       title = "Relationship between body mass and flipper length in 3 species of penguin") +
  theme_classic()
```

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



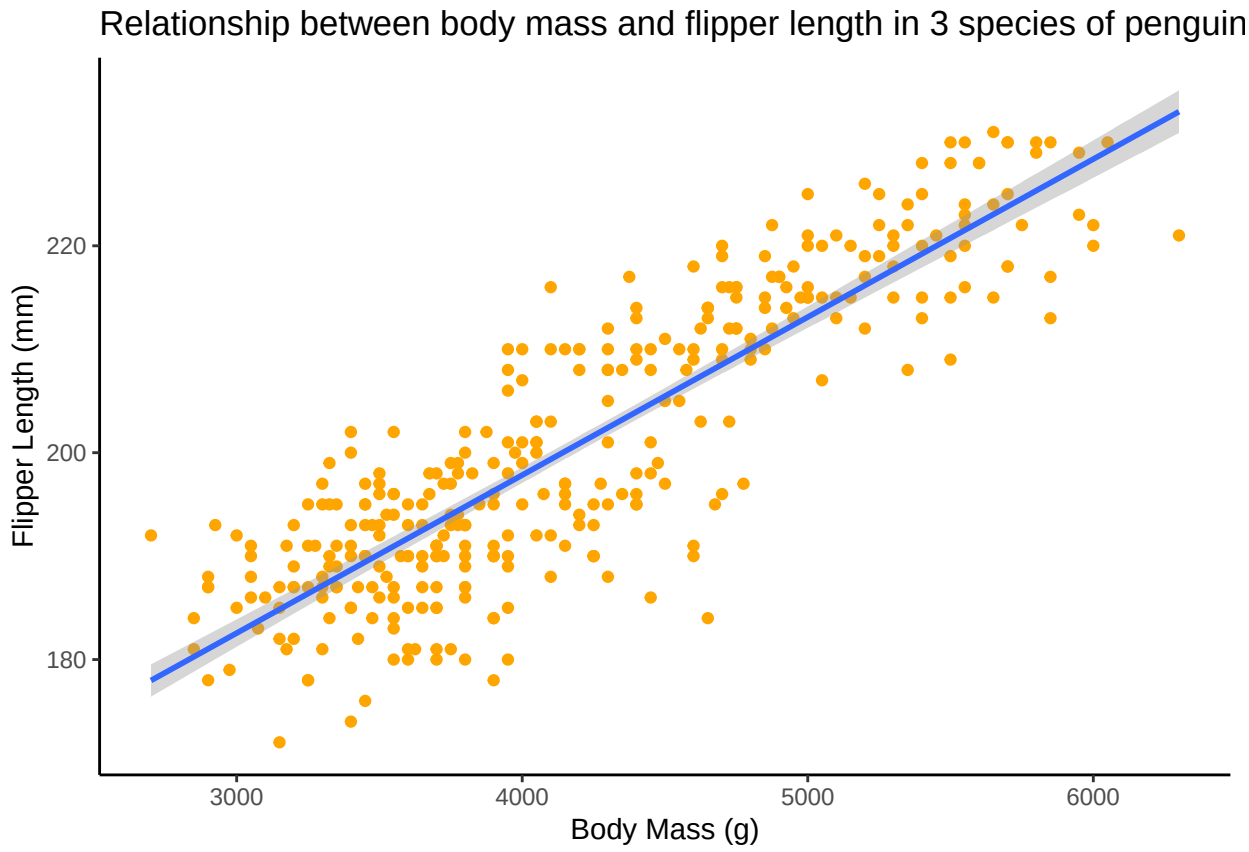
In this graph we can see that there is a strong positive correlation between body mass and flipper length in penguins - body mass is a good predictor of flipper length. To improve this graph, it would be useful to colour points by different species type, as this would help to visualise whether the correlation differs between species.

I have also saved this code as a seperate function called plot_penguin_regression. This means I can call on this graph using the following code:

```
source("functions/plotting.r")

penguin_regression1 <- plot_penguin_regression1(penguins_now_clean)
penguin_regression1
```

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



Simpson's paradox occurs when a trend that appears in several combined groups is not the same as in several different groups. To check this not the case, I am going to also individually test the regression between flipper length and body mass for each species of penguin.

First, I will look at adelie penguins:

```
adelie <- filter(penguins_now_clean, species == "Adelie Penguin (Pygoscelis adeliae)")
```

```
adelie_model <- lm(flipper_length_mm ~ body_mass_g, adelie)
summary(adelie_model)
```

```
##
## Call:
## lm(formula = flipper_length_mm ~ body_mass_g, data = adelie)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2769  -3.6192   0.0569   3.4696  18.0477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.652e+02  3.849e+00  42.929  < 2e-16 ***
## body_mass_g 6.677e-03  1.032e-03   6.468 1.34e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5.798 on 149 degrees of freedom
## Multiple R-squared:  0.2192, Adjusted R-squared:  0.214
## F-statistic: 41.83 on 1 and 149 DF,  p-value: 1.343e-09
```

This shows that for adelie penguins, variation in body mass only explains 21.4% of variation in flipper length, which is less than when the three penguin species were combined.

```
anova(adelie_model)
```

```
## Analysis of Variance Table
##
## Response: flipper_length_mm
##              Df Sum Sq Mean Sq F value    Pr(>F)
## body_mass_g   1 1406.2 1406.18  41.833 1.343e-09 ***
## Residuals   149 5008.5   33.61
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, the p-value is 1.343e-09 which is still significantly less than 0.05, therefore I can reject the null hypothesis - body mass is a good predictor of flipper length in adelie penguins.

Next, I will do the same for chinstrap penguins.

```
chinstrap <- filter(penguins_now_clean, species == "Chinstrap penguin (Pygoscelis antarctica)")
```

```
chinstrap_model <- lm(flipper_length_mm ~ body_mass_g, chinstrap)
summary(chinstrap_model)
```

```
##
## Call:
## lm(formula = flipper_length_mm ~ body_mass_g, data = chinstrap)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.4296  -3.3315   0.4097   2.8889  11.5941
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.514e+02  6.575e+00  23.024  < 2e-16 ***
## body_mass_g 1.191e-02  1.752e-03   6.795 3.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.512 on 66 degrees of freedom
## Multiple R-squared:  0.4116, Adjusted R-squared:  0.4027
## F-statistic: 46.17 on 1 and 66 DF,  p-value: 3.748e-09
```

This shows that for adelie penguins, variation in body mass explains 40.27% of variation in flipper length, which is also less than when the three penguin species were combined.

```
anova(chinstrap_model)
```

```
## Analysis of Variance Table
##
## Response: flipper_length_mm
##             Df Sum Sq Mean Sq F value    Pr(>F)
## body_mass_g  1 1402.7 1402.68  46.168 3.748e-09 ***
## Residuals   66 2005.2   30.38
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

However, the p-value is 3.748e-09 which is still significantly less than 0.05, therefore I can reject the null hypothesis - body mass is a good predictor of flipper length in chinstrap penguins.

Finally, I will look at gentoo penguins:

```
gentoo <- filter(penguins_now_clean, species == "Gentoo penguin (Pygoscelis papua)")
```

```
gentoo_model <- lm(flipper_length_mm ~ body_mass_g, gentoo)
summary(gentoo_model)
```

```
##
## Call:
## lm(formula = flipper_length_mm ~ body_mass_g, data = gentoo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.0194  -2.7401   0.1781   2.9859   8.9806
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.713e+02  4.244e+00   40.36   <2e-16 ***
## body_mass_g 9.039e-03  8.321e-04   10.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.633 on 121 degrees of freedom
## Multiple R-squared:  0.4937, Adjusted R-squared:  0.4896
## F-statistic:   118 on 1 and 121 DF,  p-value: < 2.2e-16
```

This shows that for gentoo penguins, variation in body mass only explains 48.96% of variation in flipper length, which is still less than when the three penguin species were combined.

```
anova(gentoo_model)
```

```
## Analysis of Variance Table
##
## Response: flipper_length_mm
##              Df Sum Sq Mean Sq F value    Pr(>F)
## body_mass_g   1 2533.2 2533.23  118.01 < 2.2e-16 ***
## Residuals   121 2597.5   21.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

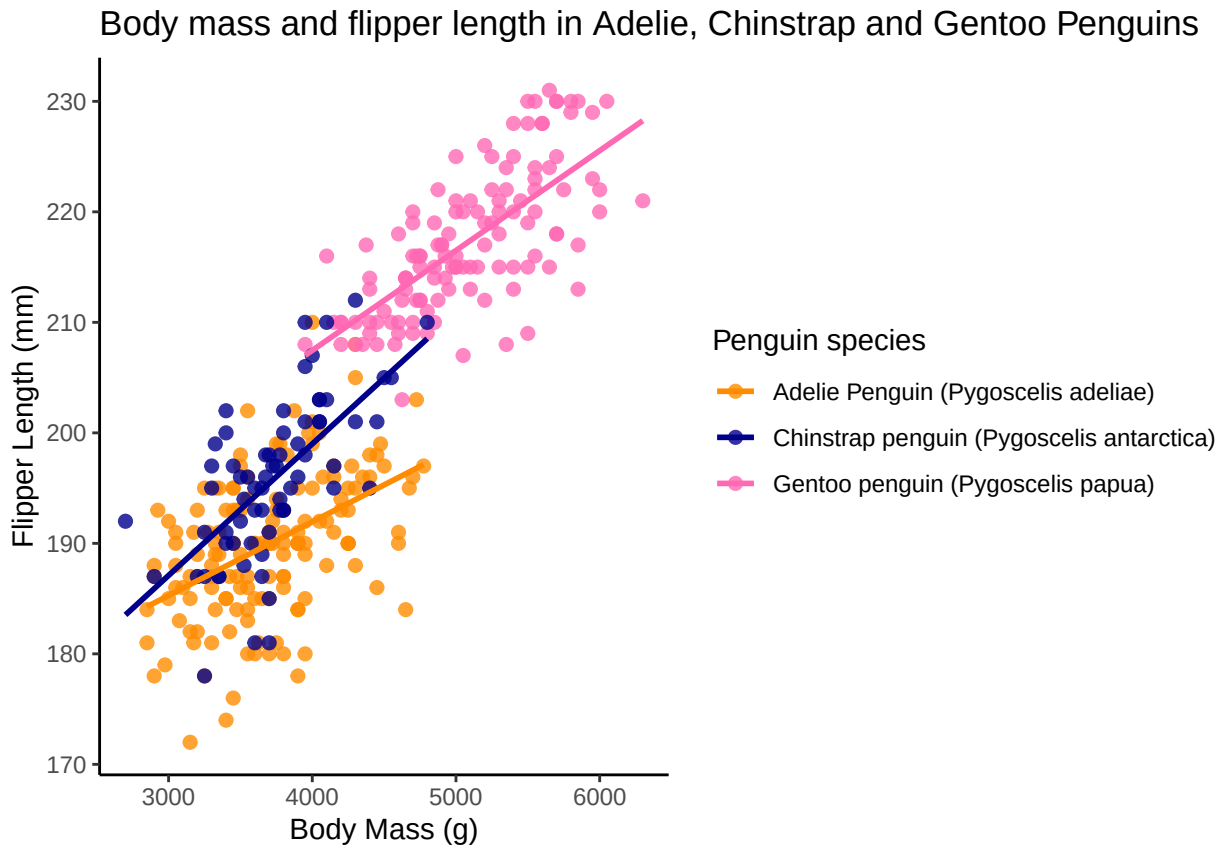
However, the p-value is 2.2e-16 which is still significantly less than 0.05, therefore I can reject the null hypothesis - body mass is a good predictor of flipper length in gentoo penguins.

Therefore, whilst variation in body mass explains less of the variation in flipper length when each species is tested seperately, for all three species, variation in body mass is still a good predictor of variation in flipper length.

To show this visually, I am now going to plot these three regressions on the same figure. I will also save this code as a function in a seperate r script called 'plotting.r'.

```
ggplot(penguins_now_clean, aes(x = body_mass_g, y = flipper_length_mm, colour = species)) +
  geom_point(size = 2, alpha = 0.8) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_classic() +
  scale_colour_manual(values = c("darkorange","darkblue","hotpink")) +
  labs(title = "Body mass and flipper length in Adelie, Chinstrap and Gentoo Penguins",
       x = "Body Mass (g)", y = "Flipper Length (mm)", colour = "Penguin species")
```

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



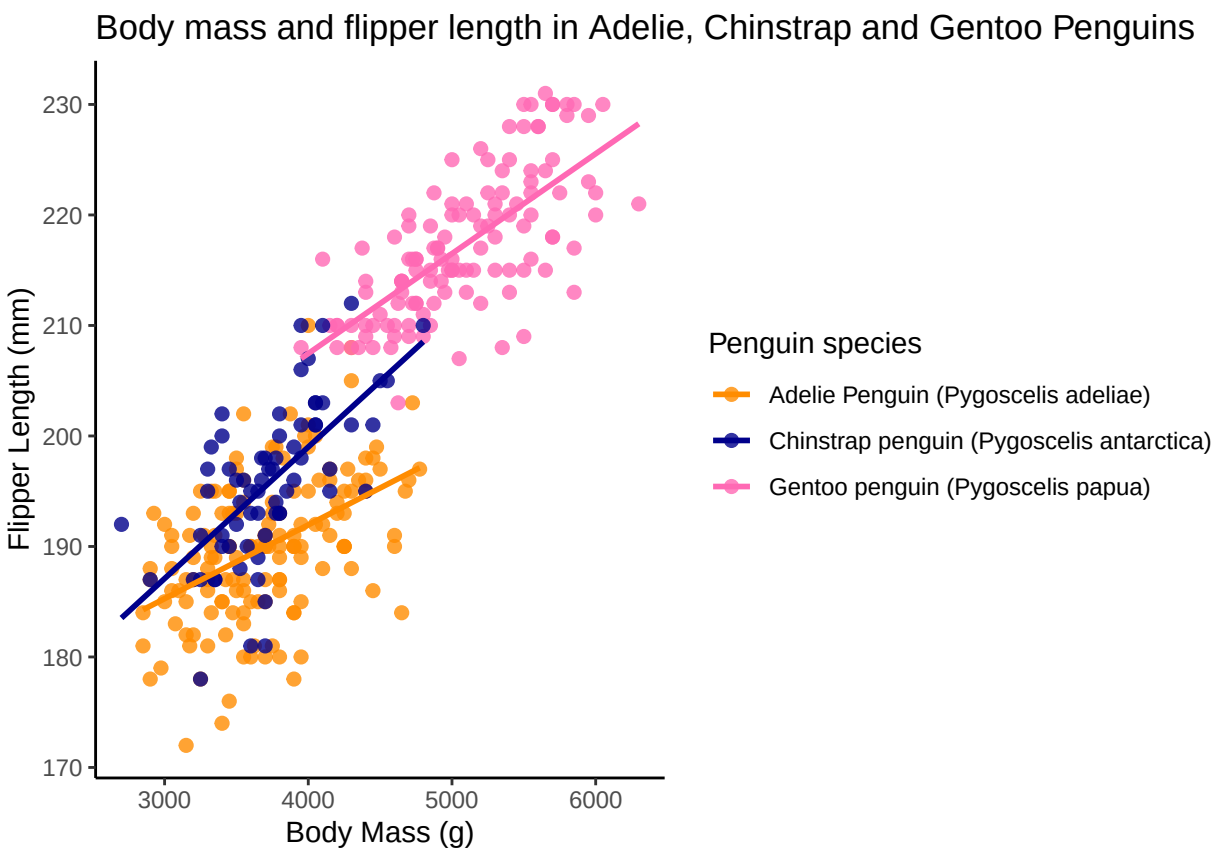
From this figure, we can see that for each penguin species, when considered individually, there is still a positive correlation between body mass and flipper length. Therefore, body mass is a good predictor of flipper length in adelie, chinstrap and gentoo penguins, and penguins with a greater body mass are likely to have longer flippers

Because I have saved this plot as a separate function, I can also call on this graph using the following code:

```
source("functions/plotting.r")

penguin_regression2 <- plot_penguin_regression2(penguins_now_clean)
penguin_regression2
```

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



I will now save my images as a png:

```
library(ragg)
```

Saving figure 1 as a png:

```
agg_png("figures/penguins_regression1_25x15.png",
        width=25, height=15, units="cm", res=600, scaling=1.4)
penguin_regression1
```

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

```
dev.off()
```

```
## pdf
##   2
```

Saving figure 2 as a png:

```
agg_png("figures/penguins_regression2_30x15.png",
        width=30, height=15, units="cm", res=600, scaling=1.4)
penguin_regression2
```

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

```
dev.off()
```

```
## pdf
##   2
```

I can also save the following code as a function inside a separate r script:

```
save_graph1_png <- function(penguins_now_clean,
                                  filename, size, res, scaling){
    agg_png("figures/penguins_regression1_25x15.png",
            width   =  25,
            height  =  15,
            units   =  "cm",
            res     =  600,
            scaling =  1.4)
    penguins_regression1 <- plot_penguin_regression1(penguins_now_clean)
    print(penguins_regression1)
    dev.off()
}

save_graph2_png <- function(penguins_now_clean,
                            filename, size, res, scaling){
  agg_png("figures/penguins_regression2_30x15.png",
          width   =  30,
          height  =  15,
          units   =  "cm",
          res     =  600,
          scaling =  1.4)
  penguins_regression2 <- plot_penguin_regression2(penguins_now_clean)
  print(penguins_regression2)
  dev.off()
}
```

I can now call on this function:

```
source("functions/figures.r")
penguin_regression1.png <- save_graph1_png(penguins_now_clean)
```

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

```
penguin_regression2.png <- save_graph2_png(penguins_now_clean)
```

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