M03Guided

Alanna Hazlett

2024-02-11

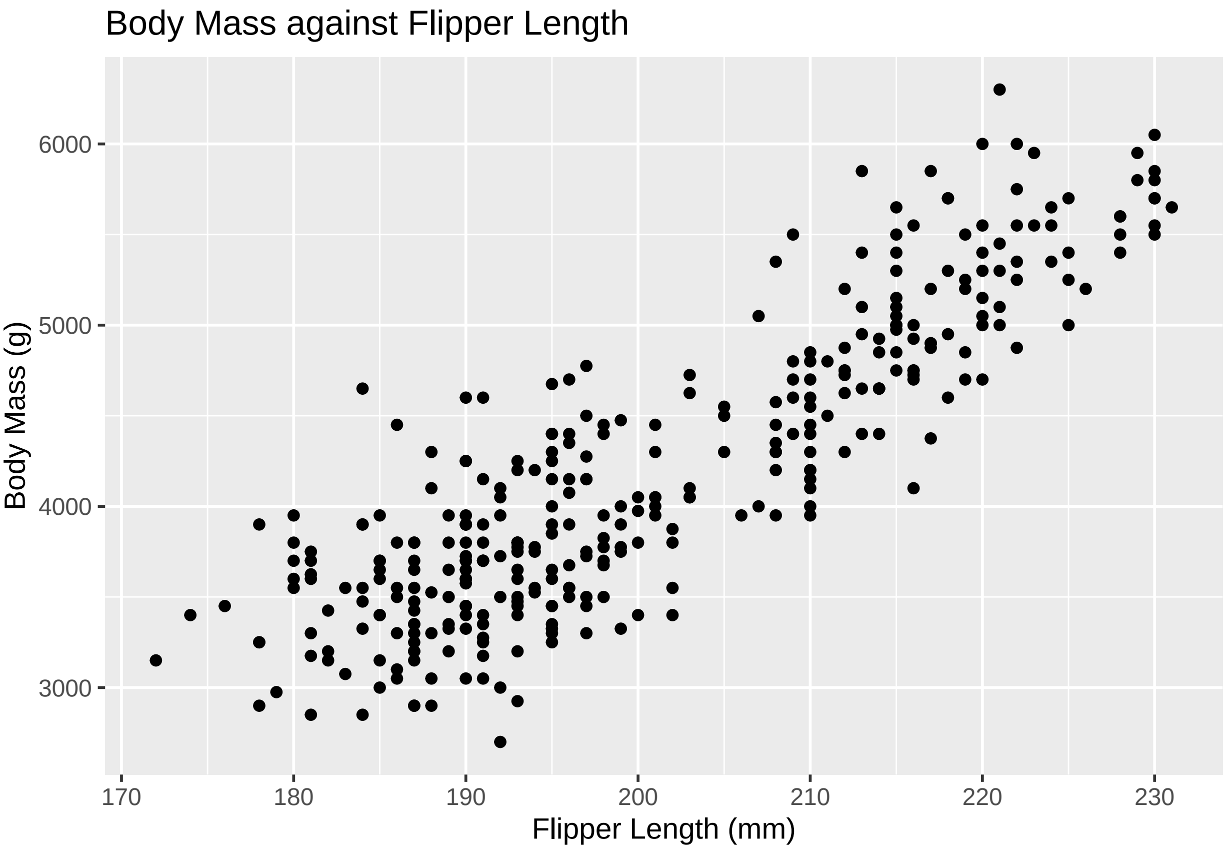
#Problem 1
We will explore the relationship between the response variable body mass (in grams), body_mass_g, and the predictor length of the flippers (in mm), flipper_length_mm.
Produce a scatterplot of the two variables. How would you describe the relationship between the two variables? Be sure to label the axes and give an appropriate title. Based on the appearance of the plot, does a simple linear regression appear reasonable for the data?

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(palmerpenguins)
Data<-penguins
ggplot2::ggplot(Data,aes(x=flipper_length_mm,y=body_mass_g))+
  geom_point()+
  #geom_smooth(method="lm",se=FALSE)+
  labs(x="Flipper Length (mm)",y="Body Mass (g)",title="Body Mass against Flipper Length")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```



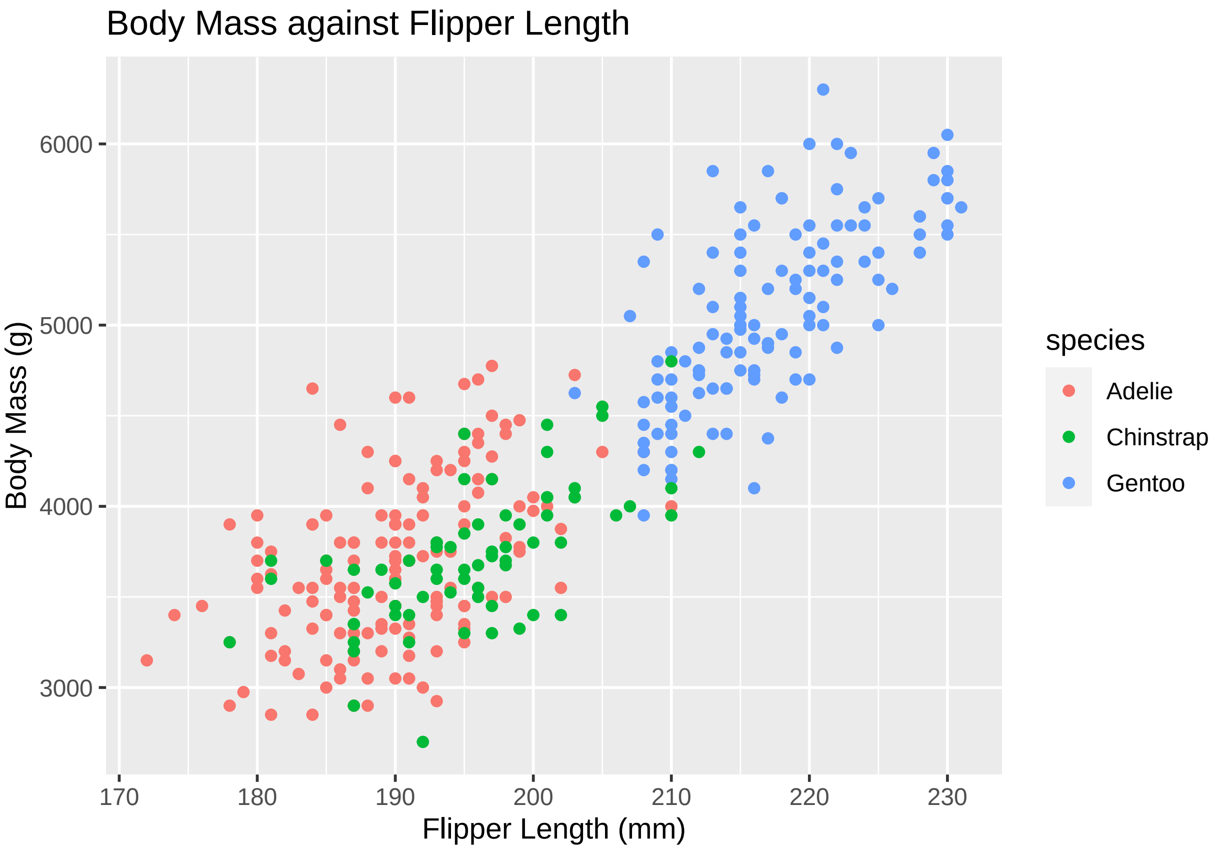
The relationship between the variables appear to be linear. A simple linear regression does seem appropriate for this data.

Problem 2

Produce a similar scatterplot, but with different colored plots for each species. How does this scatterplot influence your answer to the previous part?

```
ggplot2::ggplot(Data,aes(x=flipper_length_mm,y=body_mass_g,color=species))+
  geom_point()+
  #geom_smooth(method="lm",se=FALSE)+
  labs(x="Flipper Length (mm)",y="Body Mass (g)",title="Body Mass against Flipper Length")
```

```
## Warning: Removed 2 rows containing missing values (`geom_point()`).
```

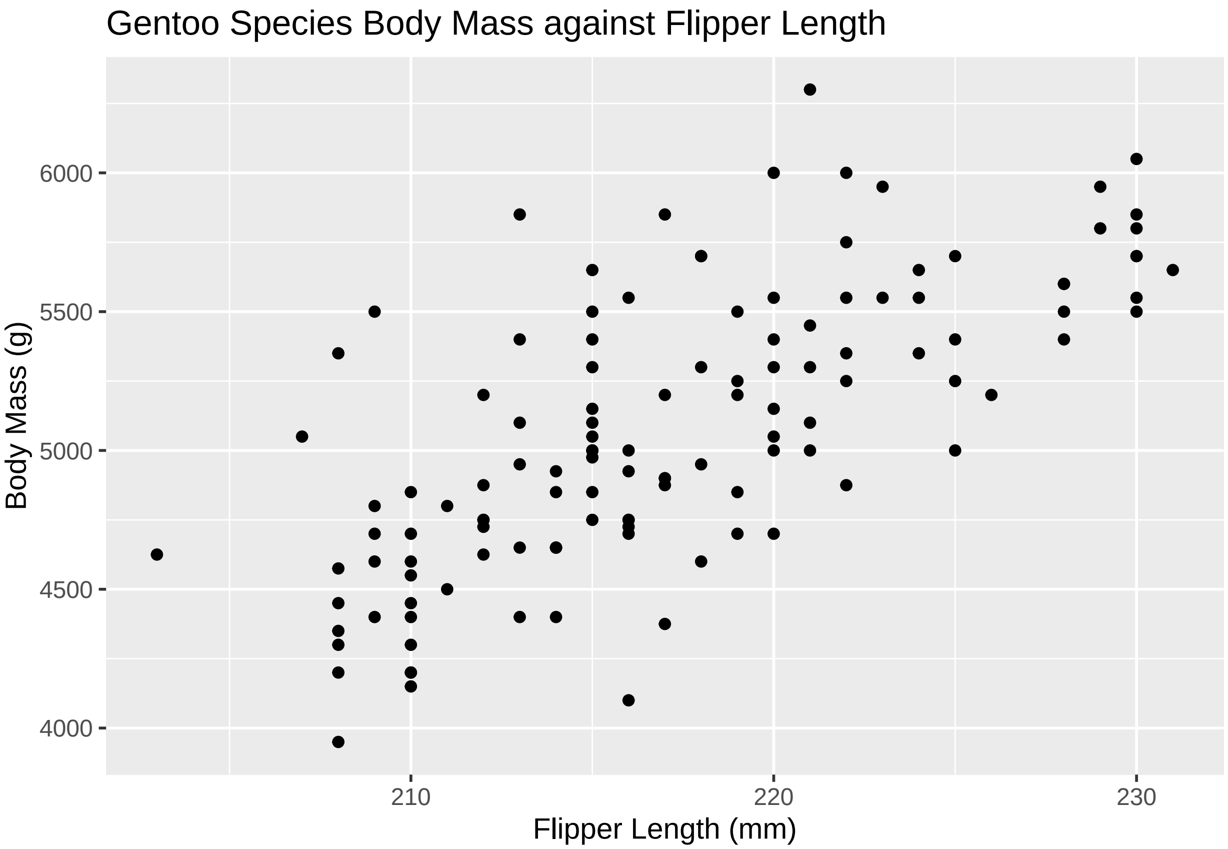


Within each species the linear relationship is maintained between flipper length and body mass. The Gentoo species is completely out of range of the other two species, Adelie and Chinstrap.

Problem 3

Regardless of your answer to the previous part, produce a scatterplot of body mass and flipper length for Gentoo penguins. Based on the appearance of the plot, does a simple linear regression appear reasonable for the data?

```
Gentoo<-Data %>%
  filter(species=="Gentoo") %>%
  filter(!is.na(flipper_length_mm)) %>%
  filter(!is.na(body_mass_g))
ggplot2::ggplot(Gentoo,aes(x=flipper_length_mm,y=body_mass_g))+
  geom_point()+
  #geom_smooth(method="lm",se=FALSE)+
  labs(x="Flipper Length (mm)",y="Body Mass (g)",title="Gentoo Species Body Mass against Flipper Length
```



A simple linear regression appears reasonable for this data, because as you scan from left to right the data points seem even placed on either side of the estimated linear regression.

Problem 4

What is the correlation between body mass and flipper length for Gentoo penguins. Interpret this correlation contextually. How reliable is this interpretation? For the rest of the questions, assume the assumptions to perform linear regression on Gentoo penguins are met.

```
cor(Gentoo$flipper_length_mm,Gentoo$body_mass_g, use = "complete.obs")
```

```
## [1] 0.7026665
```

The correlation is a postive value, which indicates that as the flipper length increases, so does the body mass. The correlation is moderate, with a value closer to 1 than to 0.

Problem 5

Use the lm() function to fit a linear regression for body mass and flipper length for Gentoo penguins. Write out the estimated linear regression equation.

```
result<-lm(body_mass_g~flipper_length_mm,data=Gentoo)
summary(result)
```

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

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -911.18 -235.76  -51.93  170.75 1015.71
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6787.281   1092.552  -6.212 7.65e-09 ***
## flipper_length_mm    54.623      5.028  10.863  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 360.2 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
```

The estimated linear regression equation is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i \hat{y} = -6787.281 + 54.623x$$

Problem 6

Interpret the estimated slope contextually.
As flipper length increases by 1mm, the body mass increases by 54.623g, on average.

Problem 7

Does the estimated intercept make sense contextually?
No, the estimated intercept does not make sense, as you can not have a negative value for body mass or a flipper length of zero.

Problem 8

Report the value of R2 from this linear regression, and interpret its value contextually. R2 is 0.4937, which means that about 49.37% of the variation of body mass can be explained by flipper length for the Gentoo penguins.

Problem 9

What is the estimated value for the standard deviation of the error terms for this regression model, sigma? s = 360.2

Problem 10

For a Gentoo penguin which has a flipper length of 220mm, what is its predicted body mass in grams? The predicted body mass of a Gentoo penguin with flipper length 220mm is 5229.67g.

```
yhat<-result$coefficients[1] + result$coefficients[2]*220
yhat
```

```
## (Intercept) 
##     5229.67
```

Problem 11

Produce the ANOVA table for this linear regression. Using only this table, calculate the value of R2.

```
anova.tab<-anova(result)
anova.tab
```

```
## Analysis of Variance Table
## 
## Response: body_mass_g
##                    Df   Sum Sq  Mean Sq F value    Pr(>F)    
## flipper_length_mm   1 15308045 15308045  118.01 < 2.2e-16 ***
## Residuals         121 15696203   129721                      
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
SST<-sum(anova.tab$"Sum Sq")
SST
```

```
## [1] 31004248
```

```
anova.tab$"Sum Sq"[1]/SST
```

```
## [1] 0.4937402
```

$$R^2 = \frac{SS_R}{SS_T} = \frac{15308045}{31004248} = 0.4937402$$

Problem 12

What are the null and alternative hypotheses for the ANOVA F test?

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

Problem 13

Explain how the F statistic of 118.01 is found.
The F statistic is:

$$F = \frac{MS_R}{MS_{res}} = \frac{15308045}{129721}$$

Problem 14

Write an appropriate conclusion for the ANOVA F test for this simple linear regression model.
Find the critical value:

```r
qf(1-0.05, 1, 123-2)
```

```
## [1] 3.919465
```

Our F statistic of 118 is larger than our critical value of 3.92, so we reject our null hypothesis of

$$\beta_1 = 0$$

our data supports the alternative hypothesis of the slope being different from 0, which indicates a linear association between our variables of flipper length and body mass.