Stat 6021: Addressing Guided Question Set for Module 3: Simple Linear Regression

Tom Lever

09/10/22

We will look at a data set concerning adult penguins near Palmer Station, Antarctica. The data set, `penguins`, comes from the `palmerpenguins` package. Be sure to install and load the `palmerpenguins` package. It is recommended to read the documentation for this data set by typing `?penguins`.

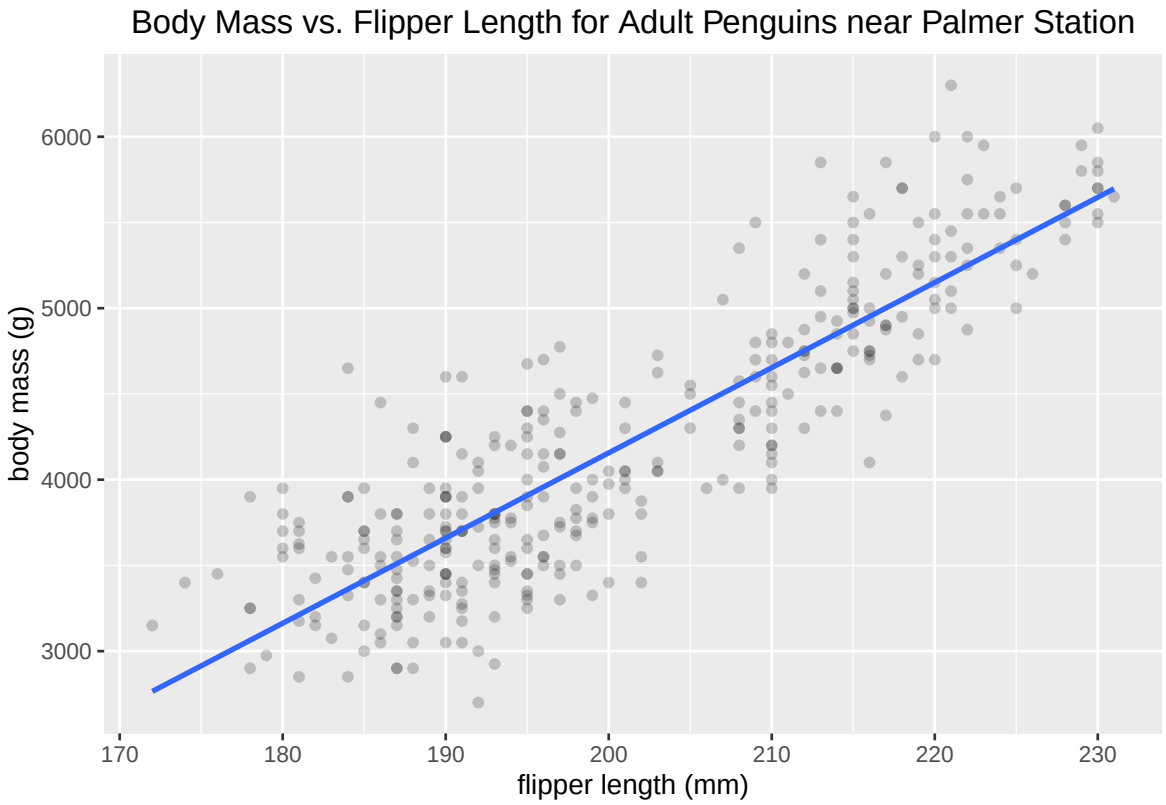
We will explore the relationship between the response variable body mass (in grams), `body_mass_g`, and the predictor length of the flippers (in mm), `flipper_length_mm`.

```
library(palmerpenguins)
library(dplyr)
species_flipper_length_and_body_mass <-
    palmerpenguins::penguins %>%
        select(species, flipper_length_mm, body_mass_g) %>%
        filter(!is.na(flipper_length_mm))
head(species_flipper_length_and_body_mass, n = 3)
```

```
## # A tibble: 3 x 3
##   species flipper_length_mm body_mass_g
##   <fct>               <int>       <int>
## 1 Adelie                181        3750
## 2 Adelie                186        3800
## 3 Adelie                195        3250
```

1. 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?

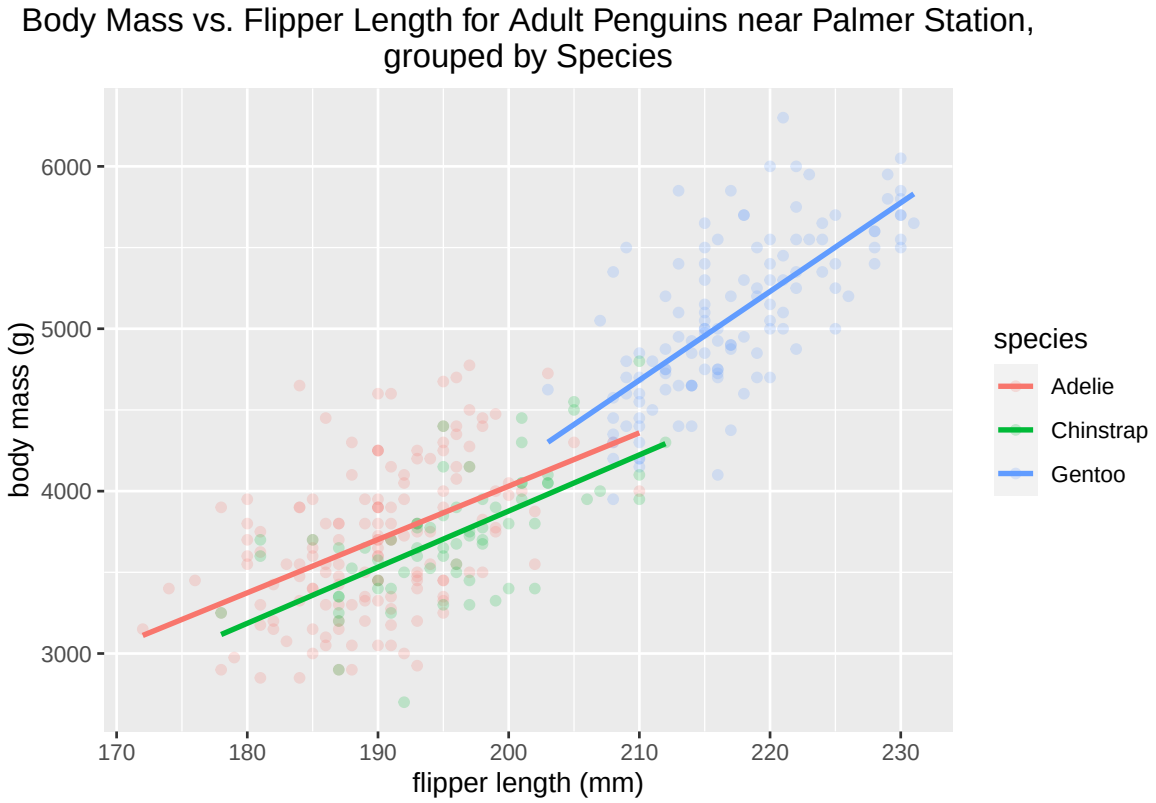
```
library(ggplot2)
ggplot(
    species_flipper_length_and_body_mass,
    aes(x = flipper_length_mm, y = body_mass_g)
) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "flipper length (mm)",
        y = "body mass (g)",
        title = "Body Mass vs. Flipper Length for Adult Penguins near Palmer Station"
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```



The relationship between flipper length and body mass for adult penguins near Palmer Station, Antarctica appears linear. A line of best fit has been rendered to aid in this determination. A simple linear regression model appears reasonable for flipper-length and body-mass data.

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

```
ggplot(
    species_flipper_length_and_body_mass,
    aes(x = flipper_length_mm, y = body_mass_g, color = species)
) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "flipper length (mm)",
        y = "body mass (g)",
        title = paste(
            "Body Mass vs. Flipper Length for Adult Penguins near Palmer Station,\n",
            "grouped by Species",
            sep = ""
        )
    ) +
    theme(
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```

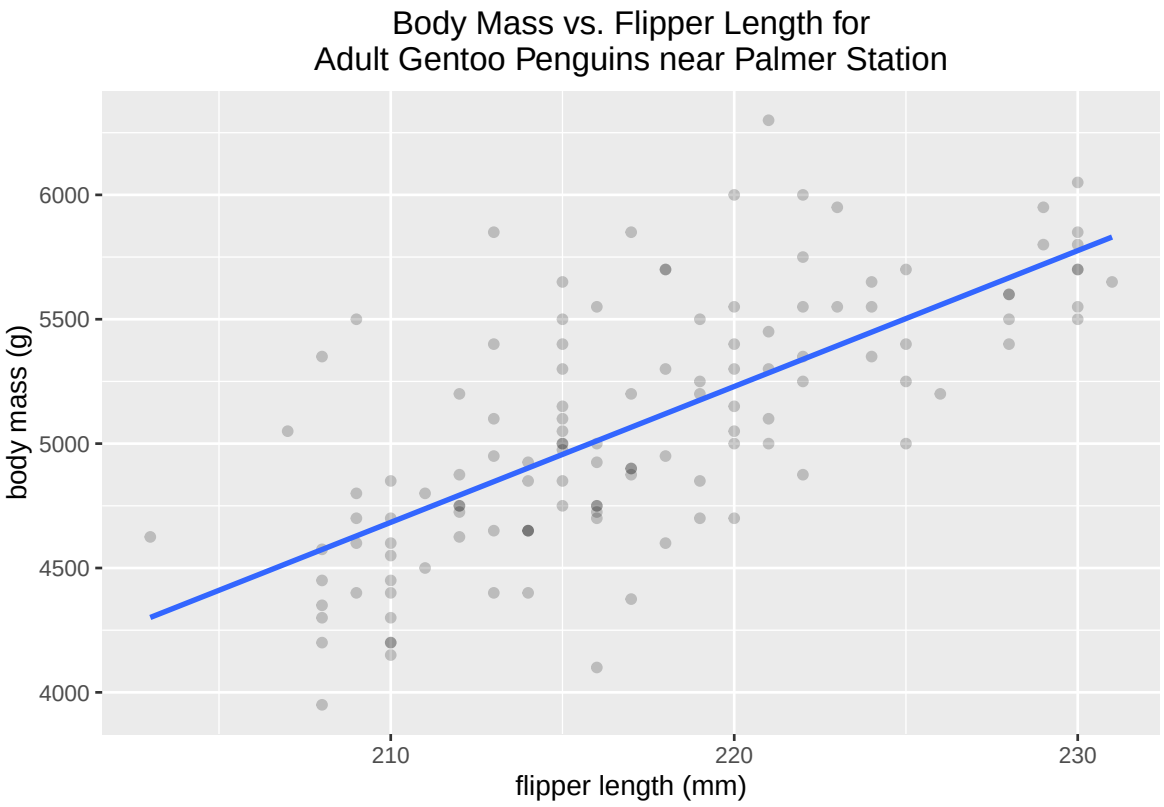


The relationship between flipper length and body mass for adult penguins near Palmer Station, Antarctica appears linear for each observed species. Lines of best fit have been rendered to aid these determinations. Simple linear regression appears reasonable for flipper-length and body-mass data for each observed species. Since adult Gentoo penguins on average have greater flipper lengths and body masses than adult Adelie and Chinstrap penguins, it may be worth considering separate linear regression models for each species.

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?

```
flipper_length_length_and_body_mass_for_adult_gentoo_penguins <-
    species_flipper_length_and_body_mass %>% filter(species == "Gentoo")
ggplot(
    flipper_length_length_and_body_mass_for_adult_gentoo_penguins,
    aes(x = flipper_length_mm, y = body_mass_g)
) +
    geom_point(alpha = 0.2) +
    geom_smooth(method = "lm", se = FALSE) +
    labs(
        x = "flipper length (mm)",
        y = "body mass (g)",
        title = paste(
            "Body Mass vs. Flipper Length for\n",
            "Adult Gentoo Penguins near Palmer Station",
            sep = ""
        )
    ) +
    theme(
```

```
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(angle = 0)
    )
```



The relationship between flipper length and body mass for adult Gentoo penguins near Palmer Station, Antarctica appears linear. A line of best fit has been rendered to aid in this determination. A simple linear regression appears reasonable for flipper-length and body-mass data.

4. What is the correlation between body mass and flipper length for Gentoo penguins? Interpret this correlation contextually. How reliable is this interpretation?

```
library(TomLeversRPackage)
data_set <- flipper_length_length_and_body_mass_for_adult_gentoo_penguins
linear_model_for_body_mass_vs_flipper_length <-
    lm(body_mass_g ~ flipper_length_mm, data = data_set)
the_summary <- summarize_linear_model(linear_model_for_body_mass_vs_flipper_length)
print(the_summary)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = data_set)
##
## 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
##
## E(y | x) = B_0 + B_1 * x = -6787.281 + 54.623 * x
## Number of observations: 123
## Estimated variance of errors: 129744.04
## Multiple R:  0.702666524357519   Adjusted R:  0.699714227381436
```

We assume that the sample of adult Gentoo penguins near Palmer Station is simple random. We assume that (flipper length, body mass) matched pairs of data for the adult Gentoo penguins near Palmer Station are independent. We assume that errors between actual and predicted body masses have mean 0 and constant variance. The linear model for body mass versus flipper length for Gentoo penguins has an adjusted sample linear Pearson correlation coefficient R of 0.703. The adjusted sample linear Pearson correlation coefficient measures the strength of the linear relationship between (flipper length, body mass) matched pairs of data for the adult Gentoo penguins near Palmer Station. The linear model has a positive linear Pearson correlation. The value of the linear Pearson correlation coefficient lies between -1 and 1 inclusive. Since the scatterplot shows a reasonable linear association, the correlation coefficient is reliable. Since the above probability 2e-16 is less than a significance level 0.05, we can reject the null hypothesis of a linear regression t test that there is no correlation between flipper length and body mass for adult Gentoo penguins near Palmer Station, and conclude that there is a correlation between flipper length and body mass for adult Gentoo penguins near Palmer Station. The linear model has an adjusted sample coefficient of determination of 0.490. The adjusted coefficient of determination R^2 is the proportion of the variation in body mass for adult Gentoo penguins that is explained by the linear relationship / flipper length. The adjusted correlation of determination lies between 0 and 1. Since the adjusted coefficient of determination is less than 0.8, the linear model is neither precise nor good for prediction.

For the rest of the questions, assume the assumptions to perform linear regression on the data set for adult Gentoo penguins near Palmer Station are met.

5. Use the `lm` function to fit a linear regression model for flipper length and body mass for adult Gentoo penguins near Palmer Station. Write out the estimated linear regression equation.

   See above.

6. Interpret the estimated slope contextually.

   The estimated slope 54.623 g/mm indicates that for every change in flipper length of 1 mm, predicted body mass will increase by 54.623 g.

7. Does the estimated intercept make sense contextually?

   The estimated intercept $-6,787.281$ g makes sense as an intercept / offset / bias that allows the estimated body mass to be $4,683.55$ g, $5,229.779$ g, and $5,776.009$ g for flipper lengths 210 mm, 220 mm, and 230 mm. That being said, a penguin cannot have a negative body mass, and a penguin cannot have a flipper length of 0 mm.

8. Report the value of R^2 from this linear regression and interpret its value contextually.

   See above.

9. What is the estimated value for the standard deviation of the error terms for this regression model, $\hat{\sigma}$?

   Errors are assumed to have mean 0 and unknown constant variance σ^2. An estimated variance is the residual mean square $\hat{\sigma}^2$. The residual standard error is $\hat{\sigma}$. The estimated value for the standard

deviation of the error terms for the regression model is also $\hat{\sigma}$. $\hat{\sigma} = 360.2\ g$.

10. For an adult Gentoo penguin which has a flipper length of 220 mm, what is its predicted body mass in grams?

    See above.

11. Produce the ANOVA table for this linear regression model. Using only this table, calculate the value of R^2.

```
analyze_variance(linear_model_for_body_mass_vs_flipper_length)
```

```
## 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
## DFT: 122, SST: 31004248
## R2: 0.493740244885153
## Number of observations: 123
```

12. What are the null and alternate hypotheses for the ANOVA F test?

    The null hypothesis for an ANOVA F test is that the slope β_1 of a linear model is equal to 0. The alternate hypothesis for an ANOVA F test is that the slope β_1 of a linear model is not equal to 0.

13. Explain how the F statistic of 118.01 is found.

$$F_0 = \frac{SS_R/df_R}{SS_{Res}/df_{Res}} = \frac{SS_R/1}{SS_{Res}/(n-2)} = (n-2)\frac{SS_R}{SS_{Res}} = (123-2)\frac{15,308,045\ g^2}{15,696,203\ g^2} = 118.01$$

$$F_0 = \frac{MS_R}{MS_{Res}} = \frac{15,308,045\ g^2}{129,721\ g^2} = 118.01$$

14. Write an appropriate conclusion for the ANOVA F test for this simple linear regression model.

    Per "Percentage Points of the F-Distribution",

$$F_{\alpha,df_R,df_{Res}} = F_{\alpha,1,n-2} = F_{0.05,1,123-2} = F_{0.05,1,121} = 3.84$$

    Since $F_0 > F_{\alpha,df_R,df_{Res}}$, we reject the null hypothesis that the slope of our linear model is equal to 0.

    Since the above probability 2.2×10^{-16} is les than 0.05, we reject the null hypothesis. We have sufficient evidence to conclude that there is a linear relationship between body mass and flipper length for adult Gentoo penguins.