Assignment 2: Wrapping up regression

Student Name DS303, SP25 Prof. Amber Camp

2025-02-28

Table of contents

Assignment 2: Wrapping up linear and logistic regression*

Assignment 2 covers linear and logistic regression models and includes many of the topics we have covered over the entire semester. You will be using `palmerpenguins` data, which includes a nice variety of continuous and categorical predictors.

To submit this assignment, render this file and save as a pdf. Upload the pdf to Canvas.

*This assignment does not include mixed effects models, but you will be seeing that on another assignment.

Load Packages

```
library(lme4)
library(tidyverse)
# install.packages("palmerpenguins") # install if needed
library(palmerpenguins)
```

Load Data

Load the `penguins` data and examine it below (use `summary()`, etc.)

```
penguins <- penguins

summary(penguins)
```

```
      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
```

Question 1: Describe the data

What data is contained in this data set? Describe at least four variables (excluding `year`), including what they represent and their data type. Lastly, describe whether you think `year` would be a useful predictor in this data.

Species, island, bill length, bill depth, flipper length, and sex.

- The species variable represents the species of the penguin: Adelie, Chinstrap, and Gentoo. This variable is categorical.
- The island variable identifies which island the penguin was observed on (Biscoe, Dream, or Torgersen). This variable is also categorical.
- Bill length measures the length of a penguins bill in millimeters. This variable is numerical.
- Flipper length, very similar to bill length, measures the length of a penguin's flipper in millimeters. This is also a numerical variable.
- Sex is just the number of male, female, or NA penguins out of the total number of penguins they observed on these islands over the years. This would also be a numerical value.

Would the variable "year" be a useful predictor?

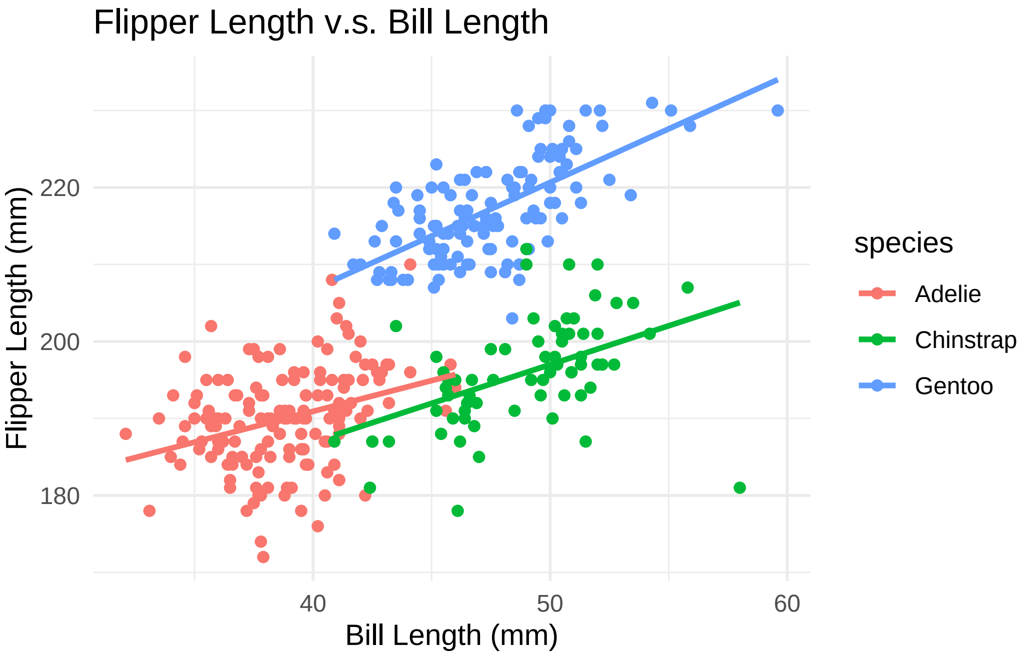
No, the year variable would not be a useful predictor because it represents the year the data was collected. Since penguins' physical characteristics and traits will typically not evolve or change significantly over such a short period of time, year may not be a strong predictor of traits like bill length or flipper length. However, if there was enough data dating back decades ago of the same species of penguins then it may be seen otherwise.

Question 2: EDA

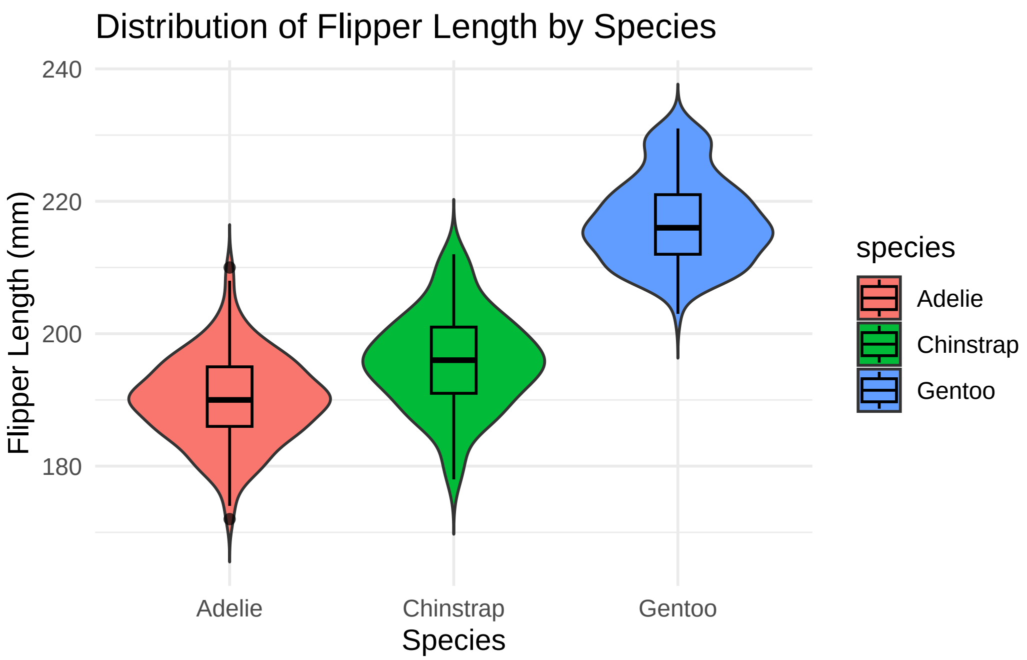
Explore your data visually. Create at least two visualizations that show the relationship between `flipper_length_mm` and its potential predictors.

```r
ggplot(penguins, aes(x = bill_length_mm, y = flipper_length_mm, color = species)) +
  geom_point()+
  geom_smooth(method = "lm", se = FALSE)+
  labs(title = "Flipper Length v.s. Bill Length", x = "Bill Length (mm)", y = "Flipper Lengt
  theme_minimal()
```

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



```r
ggplot(penguins, aes(x = species, y = flipper_length_mm, fill = species)) +
  geom_violin(trim = FALSE, aplha = 0.7) +
  geom_boxplot(width = 0.2, color = "black", alpha = 0.7)+
  labs(title = "Distribution of Flipper Length by Species",
       x = "Species",
       y = "Flipper Length (mm)") +
  theme_minimal()
```



Question 3: Apply a linear regression

Fit a simple linear regression model predicting `flipper_length_mm` from `body_mass_g`. Interpret the slope and intercept.

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

summary(flipper_body)
```

```

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

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
  (2 observations deleted due to missingness)
Multiple R-squared:  0.759, Adjusted R-squared:  0.7583 
F-statistic:  1071 on 1 and 340 DF,  p-value: < 2.2e-16
```

Interpret your model output in your own words below. Be sure to include a sentence explaining how `body_mass_g` impacts `flipper_length_mm` and whether or not the effect is significant.

Answer: Body mass significantly effects flipper length in these penguins. As shown in the summary, as the body mass increases, flipper length also increases.

Question 4: Apply a multiple linear regression

Fit a linear regression model predicting `flipper_length_mm` from both `body_mass_g` and `bill_length_mm`. Interpret the slopes and intercept.

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

summary(flipper_body_bill)
```

```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-21.0989  -4.5520   0.3379   4.8942  16.0953 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    1.220e+02  2.855e+00  42.715  < 2e-16 ***
body_mass_g    1.305e-02  5.452e-04  23.939  < 2e-16 ***
bill_length_mm 5.492e-01  8.008e-02   6.859 3.31e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.488 on 339 degrees of freedom
  (2 observations deleted due to missingness)
```

```
Multiple R-squared:  0.7884,    Adjusted R-squared:  0.7871
F-statistic: 631.4 on 2 and 339 DF,  p-value: < 2.2e-16
```

Similar to Question 3, interpret the model output in your own words here:

Answer: Body mass and bill length significantly effect and are both positively correlated with flipper length. Heavier body mass and longer bill are correlated with longer flippers in penguins and vice versa.

Question 5: Include an interaction

Fit a simple linear regression model predicting `flipper_length_mm` from `body_mass_g`, `bill_length_mm`, and the interaction of the two. Interpret the slopes and intercept.

```r
multi <- lm(flipper_length_mm ~ body_mass_g * bill_length_mm, data = penguins)

summary(multi)
```

```

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

Residuals:
     Min       1Q   Median       3Q      Max
-20.0160  -4.2070   0.3699   5.0793  16.6926

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 1.706e+02  1.612e+01  10.587  < 2e-16 ***
body_mass_g                 4.364e-04  4.149e-03   0.105  0.91629
bill_length_mm             -5.051e-01  3.528e-01  -1.432  0.15315
body_mass_g:bill_length_mm  2.707e-04  8.827e-05   3.066  0.00234 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.409 on 338 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7941,    Adjusted R-squared:  0.7923
F-statistic: 434.5 on 3 and 338 DF,  p-value: < 2.2e-16
```

Interpret the model output in your own words below. If there was a change in the pattern of significance, try to explain the logic below as well.

Answer: What this model is saying is individually, body mass and bill length have no significant correlation to flipper length; however, when looking at the interaction between the two variables, there is a significance with the p-value as it shows the body mass on flipper length depends on the bill length.

Question 6: Compare models

Compare the models you built in Questions 4 and 5 using `anova()`.

```
anova(flipper_body_bill)
```

```
Analysis of Variance Table

Response: flipper_length_mm
                Df Sum Sq Mean Sq  F value    Pr(>F)    
body_mass_g      1  51176   51176 1215.738 < 2.2e-16 ***
bill_length_mm   1   1980    1980   47.041 3.307e-11 ***
Residuals      339  14270      42                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(multi)
```

```
Analysis of Variance Table

Response: flipper_length_mm
                            Df Sum Sq Mean Sq  F value    Pr(>F)    
body_mass_g                  1  51176   51176 1245.873 < 2.2e-16 ***
bill_length_mm               1   1980    1980   48.207 1.972e-11 ***
body_mass_g:bill_length_mm   1    386     386    9.403  0.002341 ** 
Residuals                  338  13884      41                       
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Which is the better model? How do you know?
  - **Answer:** I honestly have no idea, but if I had to take a guess I would say that the flipper_body_bill model would be better, just because it seems more accurate because of the p-value and the *** next to them.

- Is it possible to compare the models from Questions 3 and 5 using the same method? Why or why not?
  - **Answer:** Question 3 is a simple linear regression with one predictor while Question 5 is a multi-linear regression. From my understanding, I don't think you could be able to compare the two and see which one would be "better" just because they are different and not both single or multi-linear regression models.

Question 7: Categorical predictors

Build a linear model that includes a categorical predictor of your choice. It is fine to stick with dummy coding. Optional: apply a different coding scheme AND interpret the output correctly for +1 extra credit.

```
flipper_species_body <- lm(flipper_length_mm ~ species + body_mass_g, data = penguins)
summary(flipper_species_body)
```

```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-14.5455  -3.1845   0.1307   3.3533  17.5313 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)      1.589e+02  2.387e+00  66.564  < 2e-16 ***
speciesChinstrap 5.597e+00  7.882e-01   7.101 7.33e-12 ***
speciesGentoo    1.568e+01  1.091e+00  14.374  < 2e-16 ***
body_mass_g      8.402e-03  6.339e-04  13.255  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.395 on 338 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8541,	Adjusted R-squared:  0.8528 
F-statistic: 659.4 on 3 and 338 DF,  p-value: < 2.2e-16
```

- What is the reference level of your categorical predictor?
  - **The reference level of our categorical predictor is Adelie.**

- What is your interpretation of this model output? Address all coefficients.
  - For every 1 gram increase in body mass, the flipper length is expected to increase by 0.0084 mm. Gentoo penguins have, on average, flipper lengths 15.68 mm longer than Adelie penguins and Chinstrap penguins have, on average, flipper lengths 5.6 mm longer than Adelie penguins also. They are all significant because of the p-value being so low.

Question 8: Relevel your categorical variable

Relevel your categorical variable so that a **different** level becomes the reference. Then, run the same model you did in Question 7 and interpret the output.

Relevel:

```
penguins$species <- relevel(penguins$species, ref = "Gentoo")
```

Apply model from Question 7:

```
relevel <- lm(flipper_length_mm ~ species + body_mass_g, data = penguins)
summary(relevel)
```

```

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

Residuals:
     Min       1Q   Median       3Q      Max 
-14.5455  -3.1845   0.1307   3.3533  17.5313 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       1.745e+02  3.254e+00  53.634  < 2e-16 ***
speciesAdelie    -1.568e+01  1.091e+00 -14.374  < 2e-16 ***
speciesChinstrap -1.008e+01  1.179e+00  -8.552 4.28e-16 ***
body_mass_g       8.402e-03  6.339e-04  13.255  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.395 on 338 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.8541,	Adjusted R-squared:  0.8528 
F-statistic: 659.4 on 3 and 338 DF,  p-value: < 2.2e-16
```

- What is the new reference level of your categorical predictor?
    - Gentoo penguins
- What is your interpretation of this new model output? Address all coefficients.
    - The coefficients from the species Adelie and Chinstrap tell us how much shorter their flippers are in length compared to Gentoo penguins. However, body mass does stay the same showing how flipper length increases with body mass.

Question 9: Apply a logistic regression

Apply a logistic regression. Include as many predictor variables as you'd like. Remember that your predicted outcome variable needs to be binary (or categorical with two levels).

Hint: You could use `sex` or create a binary variable of your own (e.g., Gentoo vs. non-Gentoo) to test your model.

```
penguins_clean <- penguins %>% drop_na(sex)

sex <- glm(sex ~ body_mass_g + flipper_length_mm + bill_length_mm,
                   data = penguins_clean,
                   family = binomial)

summary(sex)
```

```
Call:
glm(formula = sex ~ body_mass_g + flipper_length_mm + bill_length_mm,
    family = binomial, data = penguins_clean)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        8.2569324  2.7636399   2.988  0.00281 **
body_mass_g        0.0029604  0.0004188   7.069 1.56e-12 ***
flipper_length_mm -0.1348614  0.0234043  -5.762 8.30e-09 ***
bill_length_mm     0.1458945  0.0332279   4.391 1.13e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 461.61  on 332  degrees of freedom
Residual deviance: 350.73  on 329  degrees of freedom
AIC: 358.73

Number of Fisher Scoring iterations: 4
```

What are your key takeaways from this model?

Answer: This model predicts the probability of a penguin being male. Body mass coefficient shows us that because it is positive heavier penguins are more likely to be male and this is shown to be very significant. Similarly, there is a positive correlation in bill length where penguins with longer bills are more likely to be male also. However, when looking at flipper length, because the coefficient is negative this means longer flipper length is correlated with a lower probability of the penguin being a male.

Question 10: Synthesize the information

Imagine you're a biologist studying penguin populations. Which predictors do you think are most important to measure or record in the field to predict flipper length? Why?

Answer: I think if I were a biologist trying to study these penguin populations and predict flipper length, the best possible predictor would be species. Different penguin species have their own significantly different flipper lengths and ranges. Being able to identify the penguins in the field from their species, may be easier and take less time than having to weigh them individually. Also, we have already found out that Gentoo penguins generally have the longest flippers, followed by Chinstrap, and Adelie.

Bonus: Stepwise Regression

Perform stepwise regression to find the best model for an outcome of your choice. You will likely encounter an error – fixing that error and explaining your findings will earn you +1 extra credit. Show your work.

```r
library(MASS)
```

```

Attaching package: 'MASS'


The following object is masked from 'package:dplyr':

    select
```

```r
penguins_clean <- penguins %>% drop_na()

ALL <- lm(flipper_length_mm ~ body_mass_g + bill_length_mm + bill_depth_mm + species + sex,
                 data = penguins_clean)

stepwise <- stepAIC(ALL, direction = "both", trace = TRUE)
```

```
Start:  AIC=1109.22
flipper_length_mm ~ body_mass_g + bill_length_mm + bill_depth_mm + 
    species + sex

                 Df Sum of Sq     RSS    AIC
- sex             1     12.78  8942.2 1107.7
<none>                         8929.4 1109.2
- bill_depth_mm   1    144.39  9073.8 1112.6
- bill_length_mm  1    240.12  9169.5 1116.1
- body_mass_g     1    823.13  9752.5 1136.6
- species         2   2041.79 10971.2 1173.8

Step:  AIC=1107.7
flipper_length_mm ~ body_mass_g + bill_length_mm + bill_depth_mm + 
    species

                 Df Sum of Sq     RSS    AIC
<none>                         8942.2 1107.7
+ sex             1     12.78  8929.4 1109.2
- bill_depth_mm   1    199.71  9141.9 1113.1
- bill_length_mm  1    297.66  9239.8 1116.6
- body_mass_g     1   1138.28 10080.4 1145.6
- species         2   2036.43 10978.6 1172.0
```

According to this stepwise regression, explain how the final model was selected.

Answer: I think the stepwise regression model, using the AIC, removed the non significant predictors out of the model, leaving body mass, bill length, bill depth, and species, eliminating sex.