Week 9

Week 9

Body mass vs. Flipper length problems

Here is the regression model for Body Mass vs. Flipper Length:

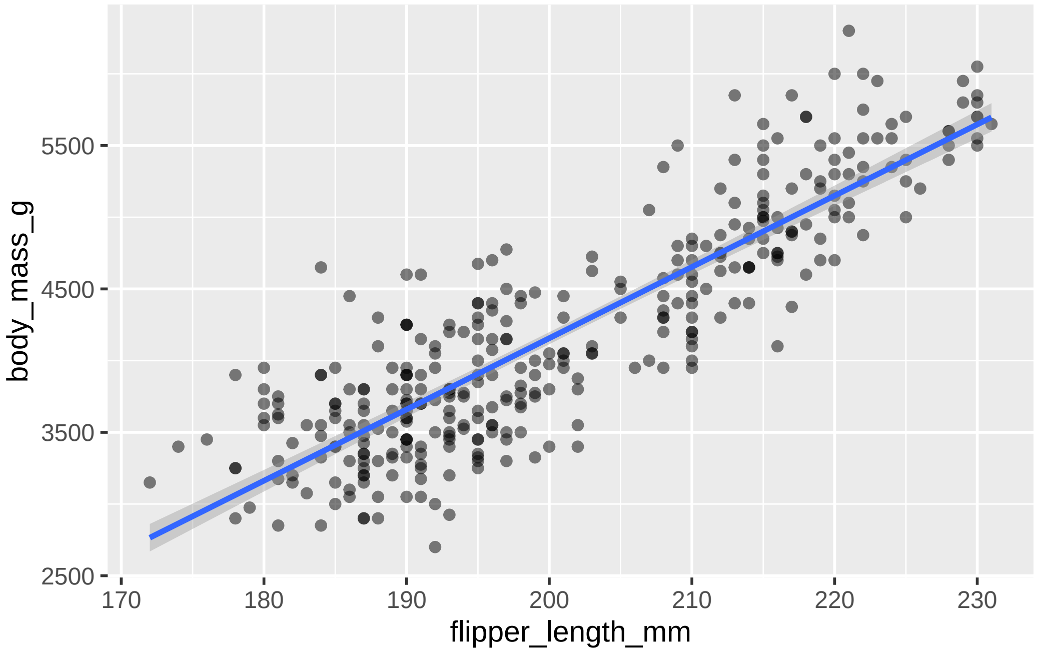
```
bm_fl_fit <- linear_reg() %>%
  fit(body_mass_g ~ flipper_length_mm, data = penguins)

tidy(bm_fl_fit)
```

```
# A tibble: 2 x 5
  term              estimate std.error statistic   p.value
  <chr>                <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)         -5781.     306.      -18.9 5.59e- 55
2 flipper_length_mm      49.7      1.52      32.7 4.37e-107
```

Here's the model visualized:

```
ggplot(penguins, aes(x = flipper_length_mm, y = body_mass_g)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm")
```



What is the estimated body mass for the penguin with a flipper length of 210?

```
penguin_210 <- tibble(flipper_length_mm = 210)

predict(bm_fl_fit, new_data = penguin_210)
```

```
# A tibble: 1 x 1
  .pred
  <dbl>
1 4653.
```

The estimated Body Mass for a penguin with a flipper length of 210 is 4653.138g

What is the estimated body mass for a penguin with a flipper length of 100?

```
penguin_100 <- tibble(flipper_length_mm = 100)

predict(bm_fl_fit, new_data = penguin_100)
```

```
# A tibble: 1 x 1
  .pred
  <dbl>
1 -812.
```

The estimated body mass for a penguin with a flipper length of 100 is -812.2747

Body mass vs. island

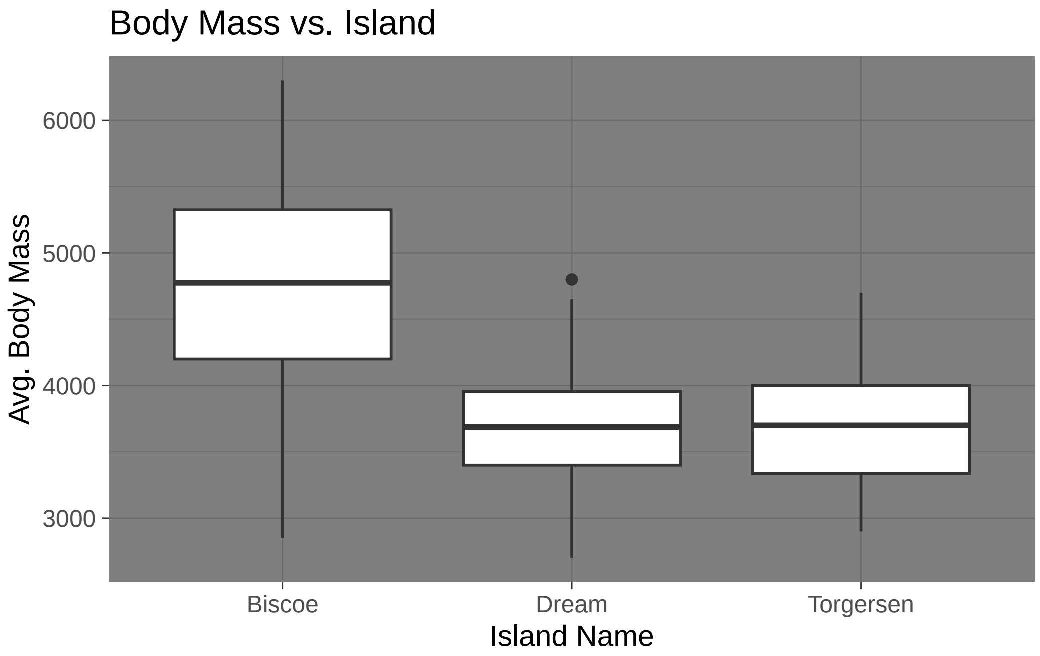
A different researcher wants to look at body weight of penguins based on the island they were recorded on. How are the variables involved in this analysis different?

The variables in this analysis are different because it involves one categorical variable, and one numerical variable instead of two numerical variables.

Make an appropriate visualization to investigate this relationship below. Additionally, calculate the mean body mass by island.

Mass vs. Island Visulaization

```r
penguindata <- penguins %>%
  group_by(island) %>%
  summarize(
    bdymass_island = mean(body_mass_g, na.rm = TRUE)
  )
penguins %>%
  ggplot(aes(x = island, y = body_mass_g)) +
  geom_boxplot() +
  labs(
    x = "Island Name",
    y = "Avg. Body Mass",
    title = "Body Mass vs. Island"
  ) + theme_dark()
```



Mean body Mass

Mean body mass grouped by *island*, NA results have been removed.

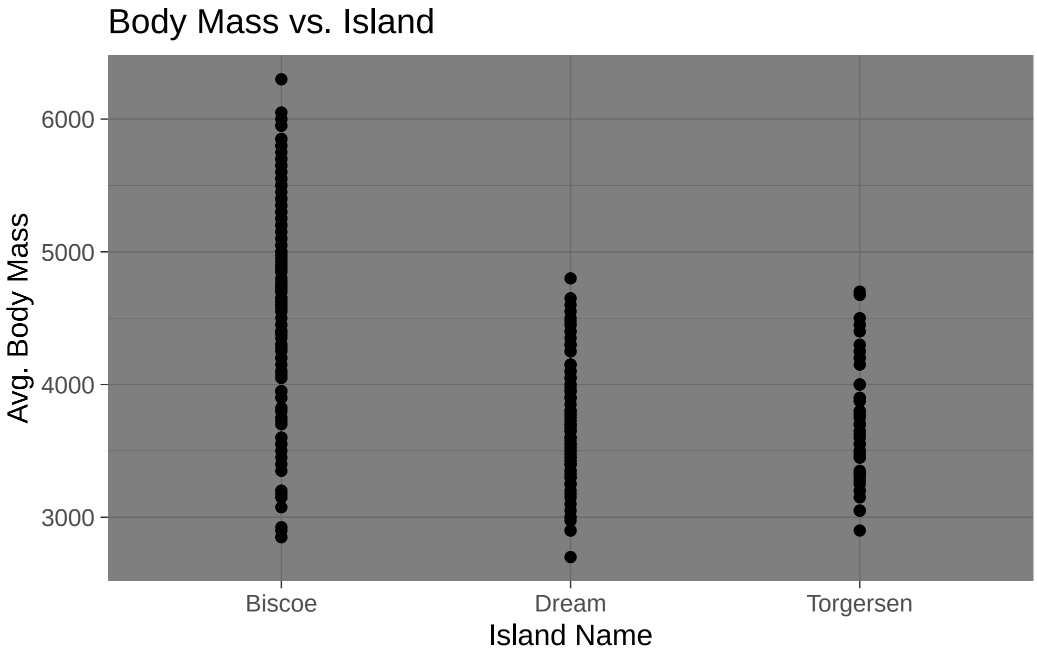
```
penguindata
```

```
# A tibble: 3 x 2
  island    bdymass_island
  <fct>              <dbl>
1 Biscoe             4716.
2 Dream              3713.
3 Torgersen          3706.
```

Change the geom of your previous plot to geom_point(). Use this plot to think about how R models these data.

```
penguindata <- penguins %>%
  group_by(island) %>%
  summarize(
    bdymass_island = mean(body_mass_g, na.rm = TRUE)
  )
```

```
penguins %>%
  ggplot(aes(x = island, y = body_mass_g)) +
  geom_point() +
  labs(
    x = "Island Name",
    y = "Avg. Body Mass",
    title = "Body Mass vs. Island"
  ) + theme_dark()
```



Since the x axis represents the categorical variable, and the y axis represents the numerical variable, the data will be directly on the line for said categorical variable. R models this data as it should, however, this is NOT the best way to visually depict this data. We can use bar plots, box plots, or histograms to best describe this data. In particularly, I would use a box plot to visualize this data as it emphasizes the summary statistics as the main focal point to the graph.

Fit the linear regression model and display the results. Write the estimated model output below.

```
penguin.beans <- linear_reg() %>%
  fit(body_mass_g ~ island, data = penguins)
tidy(penguin.beans)
```

```
# A tibble: 3 x 5
  term            estimate std.error statistic   p.value
  <chr>              <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)        4716.      48.5      97.3 8.93e-250
2 islandDream       -1003.      74.2     -13.5 1.42e- 33
3 islandTorgersen   -1010.     100.      -10.1 4.66e- 21
```

```
#Remember the statistic in this situation is the t-Statistic
```

Interpret each coefficient in context of the problem.

- **Biscoe (Intercept)**
  - Estimate
    - **Highest**
    - Positive correlation when compared with other islands.
  - Standard Error
    - **Lowest**
    - Highest Precision in accuracy.
  - Statistic
    - **Highest**
    - Predictor variable is most likely to be significant.
  - P Value
    - **Lowest**
    - Below 0.5, strongly suggests predictor variable is significant.
- **Dream**
  - Estimate
    - **Middle**
    - Negative correlation towards the reference level (Biscoe-Intercept).
  - Standard Error
    - **Middle**
    - Mediocre precision in accuracy.
  - Statistic
    - **Lowest**
    - Predictor variable is least likely to be statistically significant.
  - P Value
    - **Middle**
    - Below 0.5, strongly suggests predictor variable is significant.

- **Torgersen**
  - Estimate
    - **Lowest**
    - Negative correlation towards the reference level (Biscoe-Intercept).
  - Standard Error
    - **Lowest**
    - Lowest Precision in accuracy.
  - Statistic
    - **Middle**
    - Predictor variable is second least likely to be significant.
  - P Value
    - **Highest**
    - Below 0.5, strongly suggests predictor variable is significant.

What is the estimated body weight of a penguin on Biscoe island? What are the estimated body weights of penguins on Dream and Torgersen islands?

- The estimated body weight of a penguin on **Biscoe** island is…
  - **4716.018**
- The estimated body weight of a penguin on **Torgersen** island is…
  - **3712.903**
- The estimated body weight of a penguin on **Dream** island is…
  - **3706.373**

Body mass vs. flipper length and island

Fit a model to predict body mass from flipper length and island. Display the summary output and write out the estimate regression equation below.

Summary Output

```
bm_fl_island_fit <- linear_reg() %>%
  fit(body_mass_g ~ flipper_length_mm + island, data = penguins)
tidy(bm_fl_island_fit)
```

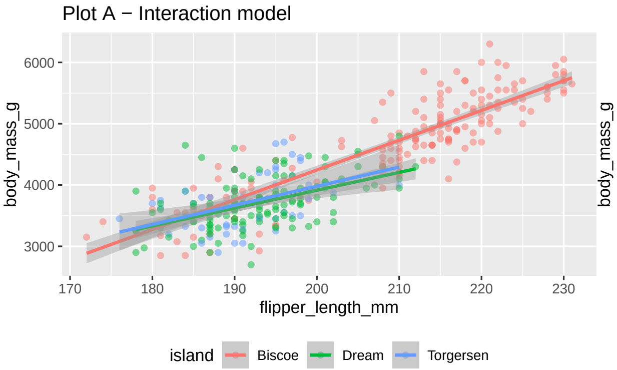
```
# A tibble: 4 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)        -4625.     392.       -11.8  4.29e-27
2 flipper_length_mm     44.5      1.87      23.9  1.65e-74
3 islandDream         -262.      55.0       -4.77 2.75e- 6
4 islandTorgersen     -185.      70.3       -2.63 8.84e- 3
```

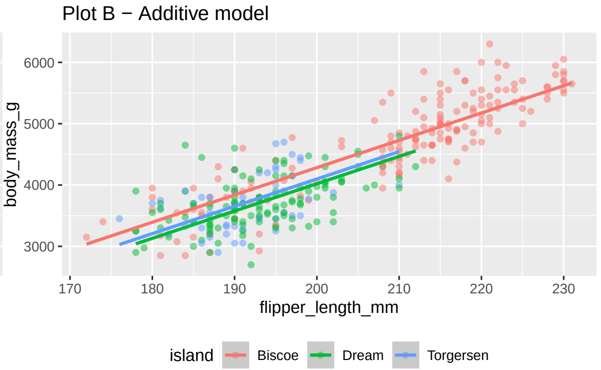
Estimate Regression Equation

$$y = -4624.98 + 44.54(FlipperLength) - 262.18(DreamIsland) - 185.13(TorgensenIsland)$$

Additive vs. interaction models

Run the two chunks of code below and create two separate plots. How are the two plots different than each other? Which plot does the model we fit above represent?





- The two plots are different because one has straight best fit lines, while the other plot (Plot A) has lines with slight curves, as well as an error envelope encompassing said lines.
- The plot that represents the model we fit above is **Plot B - Additive model**.

Interpret the slope coefficient for flipper length in the context of the data and the research question.

The slope in this context is as follows: As flipper length **(flipper_length_mm)** increases, so does body mass **(body_mass_g)**. In particular, the slope is 44.5g. As fipper length increases by 1mm, body mass increases by 44.5g.

Predict the body mass of a Dream island penguin with a flipper length of 200 mm.

```
 penguin <- data.frame(
  island = "Dream",
  flipper_length_mm = 200
  )
#Can use tibble instead of data.frame
# Could also do:
#   penguin <- penguins %>%
#     filter(flipper_length_mm == 200, island == "Dream") %>%
#     slice(1:1)


predict(bm_fl_island_fit, new_data = penguin)
```

```
# A tibble: 1 x 1
  .pred
  <dbl>
1 4021.
```

The predicted body mass of a Dream island penguin with a flipper length of 200mm given this model is **4021.473g**.

Look back at Plot B. What assumption does the additive model make about the slopes between flipper length and body mass for each of the three islands?

The assumption that the *Additive Model* makes about the slopes between flipper length and body mass for each of the three islands is that the change in body mass is linear, and doesn't include fluctuations as seen in the *Interactive Model*.

Now fit the interaction model represented in Plot A and write the estimated regression model.

```
mass.fl.island <- linear_reg() %>%
fit(body_mass_g ~ flipper_length_mm * island, data = penguins)

# Can also do:
# mass.fl.island <- lm(body_mass_g ~ flipper_length_mm * island, data = penguins)
#tidy(mass.fl.island)
#NOTE: won't give accurate prediction when needing to predict dependent value.
```

Estimate Regresison Equation

$$y = -5463.9 + 48.5(FL) + 3550.7(Dream) + 3217.8(Torgensen) - 19.4(FL*Dream) - 17.4(FL*Torgensen)$$

What does modeling body mass with an interaction effect get us that without doing so does not?

Modeling body mass with an interaction effect gives us more accurate predictions than an additive model. While a additive model is more linear, it does not account for the "randomness," or unusual observations when examining data in real life. For example, if this graph where to have extreme observations, where the standard deviation was extreme, but the mean was normal, we might get a straight line. Whereas the interactive model would display this data as a graph of that similar to **f(x)=sin(x)** or **f(x)=cos(x)**.

Predict the body mass of a Dream island penguin with a flipper length of 200 mm.

```r
penguin_200 <- tibble(flipper_length_mm = 200, island = "Dream")
#Can use 'data.frame()' instead of tibble
predict(mass.fl.island, new_data = penguin_200)
```

```
# A tibble: 1 x 1
  .pred
  <dbl>
1 3915.
```

Choosing a model

The model I'm choosig is the Body Mass vs Flipper Length by Island fitted Model.

```r
glance(bm_fl_island_fit)
```

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic   p.value    df logLik   AIC   BIC
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>
1     0.774         0.772  383.      386. 7.60e-109     3 -2517. 5045. 5064.
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

What is R-squared? What is adjusted R-squared?

R-squared is roughly…

- **0.7742334**

Adjusted R-squared is roughly…

- **0.7722296**

Your turn

Now, explore body mass, and it's relationship to bill length and flipper length. Brainstorm: How could we visualize this?

We could visualize body mass and it's relationship to bill & flipper length by using an additive model. Since all variables in this equation are numerical, we are better off using an additive model rather then an interactive model.

Fit the additive model. Interpret the slope for flipper in context of the data and the research question.

Fitted Model

```
bm_bl_fl_fit <- linear_reg() %>%
  fit(body_mass_g ~ flipper_length_mm + bill_length_mm, data = penguins)
tidy(bm_bl_fl_fit)
```

```
# A tibble: 3 x 5
  term              estimate std.error statistic  p.value
  <chr>                <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)       -5737.      308.       -18.6  7.80e-54
2 flipper_length_mm    48.1       2.01      23.9  7.56e-75
3 bill_length_mm        6.05      5.18       1.17 2.44e- 1
```

Slope Interpretation - Flipper Length

The slope for *Flipper Length (flipper_length_mm)*, is 48.1. This means as flipper length increases from millimeter to millimeter, *Body Mass (body_mass_g)* will also increase by 48.1g.