AE 12: Modeling penguins with multiple predictors

In this application exercise we will be studying penguins. The data can be found in the **palmerpenguins** package and we will use **tidyverse** and **tidymodels** for data exploration and modeling, respectively.

```r
library(tidyverse)
library(tidymodels)
library(palmerpenguins)
```

Please read the following context and take a `glimpse` at the data set before we get started.

> This data set comprising various measurements of three different penguin species, namely Adelie, Gentoo, and Chinstrap. The rigorous study was conducted in the islands of the Palmer Archipelago, Antarctica. These data were collected from 2007 to 2009 by Dr. Kristen Gorman with the Palmer Station Long Term Ecological Research Program, part of the US Long Term Ecological Research Network. The data set is called `penguins`.

```r
glimpse(penguins)
```

```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adelie, Adelie, Adel~
$ island            <fct> Torgersen, Torgersen, Torgersen, Torgersen, Torgerse~
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34.1, ~
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18.1, ~
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190, 186~
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 3475, ~
$ sex               <fct> male, female, female, NA, female, male, female, male~
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007, 2007~
```

Our goal is to understand better how various body measurements and attributes of penguins relate to their body mass.

Body mass vs. flipper length

The regression model for body mass vs. flipper length is as follows.

```r
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
```

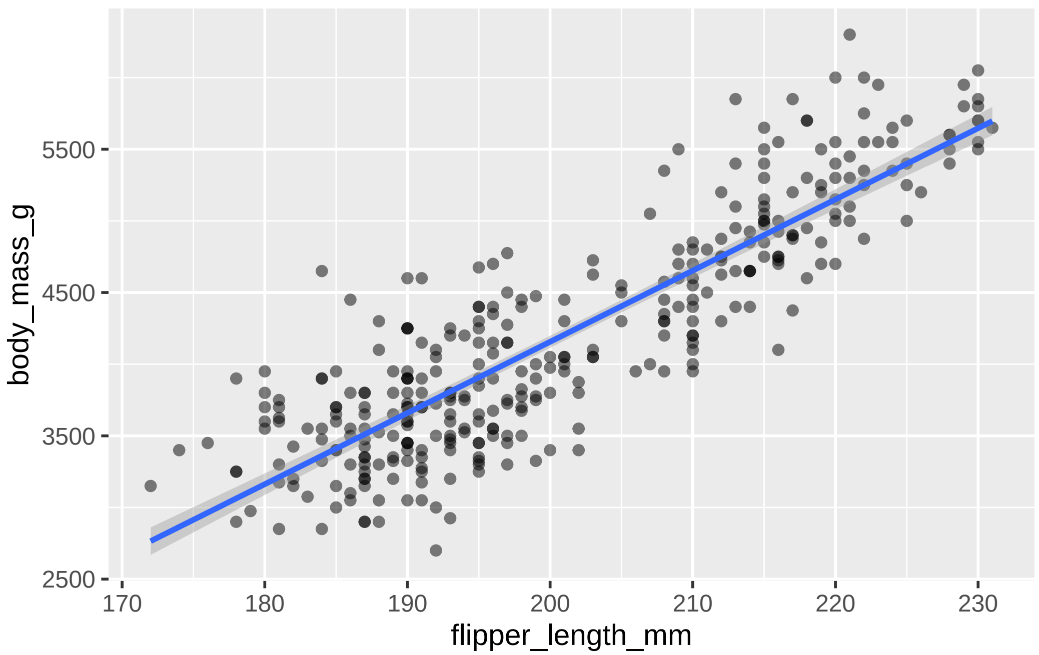
And here is the model visualized:

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

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

```
Warning: Removed 2 rows containing non-finite values (`stat_smooth()`).
```

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



- **Demo:** What is the estimated body mass for a penguin with a flipper length of 210?

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

predict(___, new_data = ___)
```

- **Your turn:** What is the estimated body mass for a penguin with a flipper length of 100?

```
# add code here
```

Body mass vs. island

- **Demo:** 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?

Add response here.

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

```
# add code here
```

```
# add code here
```

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

```
# add code here
```

- **Your turn:** Fit the linear regression model and display the results. Write the estimated model output below.

```
# add code here
```

- **Demo:** Interpret each coefficient in context of the problem.

Add response here.

- **Demo:** 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?

```
# add code here
```

Body mass vs. flipper length and island

Next, we will expand our understanding of models by continuing to learn about penguins. So far, we modeled body mass by flipper length, and in a separate model, modeled body mass by island. Could it be possible that the estimated body mass of a penguin changes by both their flipper length AND by the island they are on?

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

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

$$\textit{add math text here}$$

Additive vs. interaction models

- **Your turn:** 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?

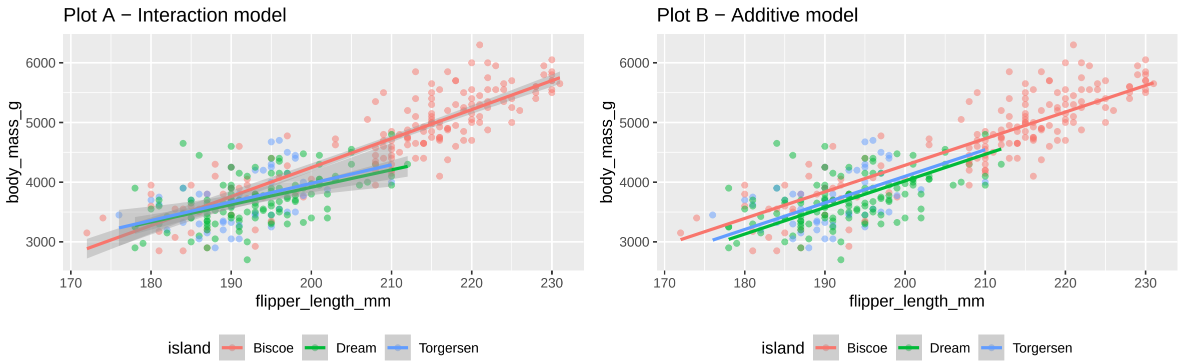
```
`geom_smooth()` using formula = 'y ~ x'
```

```
Warning: Removed 2 rows containing non-finite values (`stat_smooth()`).
```

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

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

```
Warning: Removed 2 rows containing non-finite values (`stat_smooth()`).
Removed 2 rows containing missing values (`geom_point()`).
```


Plot A − Interaction model

Plot B − Additive model

Add response here.

- **Your turn:** Interpret the slope coefficient for flipper length in the context of the data and the research question.

Add response here.

- **Demo:** Predict the body mass of a Dream island penguin with a flipper length of 200 mm.

```
# add code here
```

- **Review:** 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 additive model assumes the same slope between body mass and flipper length for all three islands.

- **Demo:** Now fit the interaction model represented in Plot A and write the estimated regression model.

```
# add code here
```

$$\textit{add math text here}$$

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

The interaction effect allows us to model the rate of change in estimated body mass as flipper length increases as different in the three islands.

- **Your turn:** Predict the body mass of a Dream island penguin with a flipper length of 200 mm.

```
# add code here
```

Choosing a model

Rule of thumb: **Occam's Razor** - Don't overcomplicate the situation! We prefer the *simplest* best model.

```
# add code here
```

- **Review:** What is R-squared? What is adjusted R-squared?

R-squared is the percent variability in the response that is explained by our model. (Can use when models have same number of variables for model selection)

Adjusted R-squared is similar, but has a penalty for the number of variables in the model. (Should use for model selection when models have different numbers of variables).

Your turn

- Now, explore body mass, and it's relationship to bill length and flipper length. Brainstorm: How could we visualize this?
- Fit the additive model. Interpret the slope for flipper in context of the data and the research question.

```
# add code here
```

Add response here.