PSTAT 5LS Lab 6

YOUR NAME HERE

Winter 2025

Section 1

Announcements & Recap

Section 2

Learning Objectives

R Learning Objectives

1. Creating a plot of (x,y) quantitative values.
2. Finding the correlation coefficient between two quantitative variables.
3. Creating a subset containing only selected variables.

Statistical Learning Objectives

1. Describe the relationship between two quantitative variables in a scatter plot, including direction, form, strength, and outliers.
2. Explain and interpret the correlation coefficient, focusing on its relationship to the strength and direction of the relationship between two quantitative variables.

Functions covered in this lab

1. `plot()`
2. `cor()`

Section 3

Lab Tutorial

Gentoo Penguins Data Set

We're back to hanging out with our penguin friends. This time, we will work with only the Gentoo penguins because we saw earlier that species may differ when it comes to physical measurements.

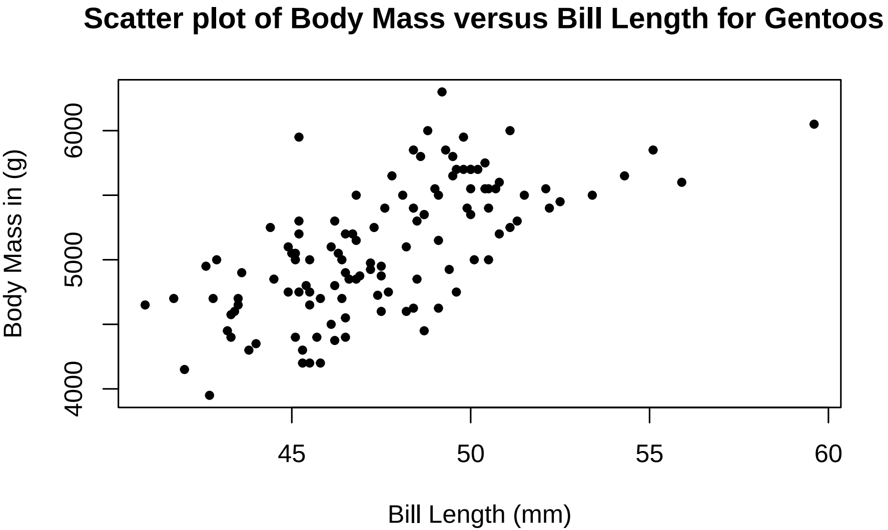
```
gentoo <- read.csv("gentoo.csv", stringsAsFactors = TRUE)
```

Go ahead and run the `loadGentoo` chunk in the `lab6-notes.Rmd` markdown file, and verify that the `gentoo` data is in your environment in the top right corner of your project.

Scatter Plots with Linear Association

In lecture, we are focusing our attention on scatter plots that appear to show a **linear** association between two numeric variables. Let's see if there is a linear association between `bill_length_mm` and `body_mass_g`.

Scatter Plot of Bill Length and Body Mass



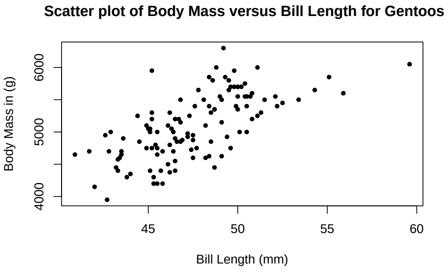
Interpreting the Scatter Plot

When interpreting a scatter plot, we want to comment on four key aspects.

1. Direction (positive or negative)
2. Form (Linear or clearly Nonlinear)
3. Strength (Weak, Moderate, Strong)
4. Outliers/Unusual Features (if there are no outliers or other unusual features, we should be sure to state that)

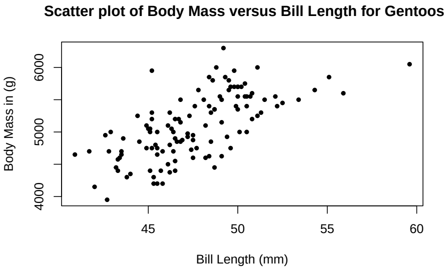
Let's interpret this scatter plot!

Interpret the Scatter Plot of Bill Length and Body Mass



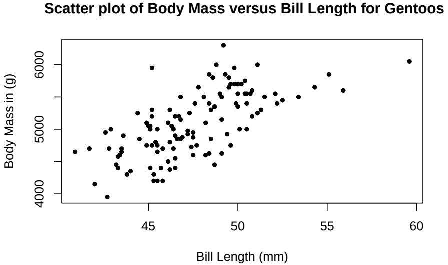
Let's start by commenting on the direction.

Interpret the Scatter Plot of Bill Length and Body Mass



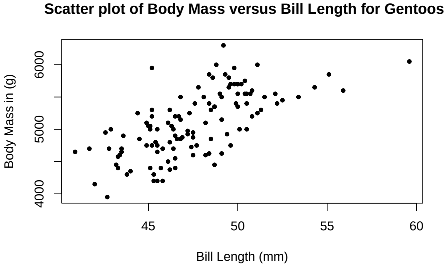
Next, comment on the form.

Interpret the Scatter Plot of Bill Length and Body Mass



Then, discuss the strength.

Interpret the Scatter Plot of Bill Length and Body Mass



Finally, discuss whether or not there are outliers or any other unusual features (e.g., groups). If you do notice anything unusual, be sure to point out where in the graph (give approximate numeric values).

Scatter Plot Code

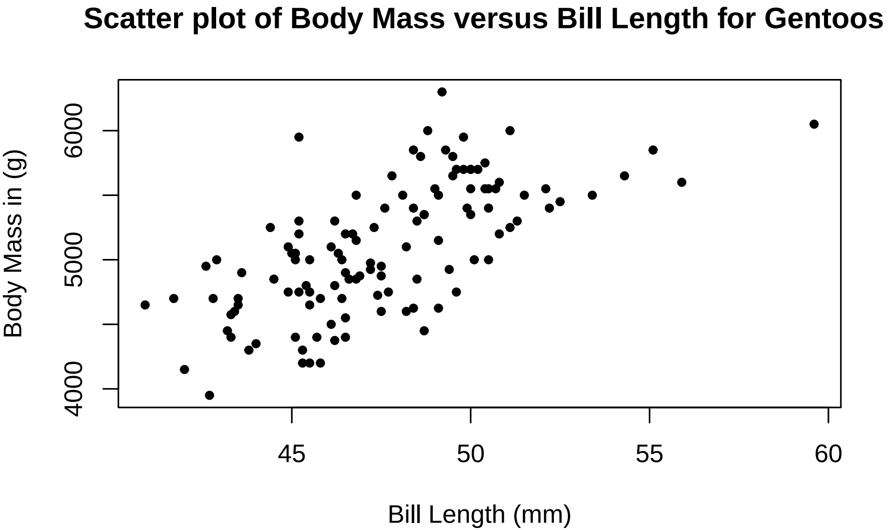
The `plot()` function in R allows us to create a scatter plot. Run the `tryIt1` chunk in your notes to create a scatter plot of body mass and bill length.

```
plot(body_mass_g ~ bill_length_mm,
     data = gentoo,
     main = " Scatter plot of Body Mass versus Bill Length for
     xlab = "Bill Length (mm)",
     ylab = "Body Mass in (g)",
     pch = 20)
```

Be *very* careful setting up scatter plots!

- Notice that the format of typing the variables in is the y ~ x format, where x is the explanatory variable and y is the response variable.
- Also notice that we can use the `data = data_name` argument in order to simplify what we write in the first line of code.

Scatter Plot of Bill Length and Body Mass



Strength and Correlation

In class, we have been observing scatter plots and commenting on the strength of the linear relationship. Earlier in our scatter plot, we observed a moderately-strong positive linear relationship, with no obvious outliers or clustering.

We can quantify the strength by computing a value called the correlation coefficient, r (or R). Let's do so using the function `cor()`:

```r
cor(gentoo$bill_length_mm, gentoo$body_mass_g)
```

```
## [1] 0.6667302
```

Let's try this code together in the `tryIt2` code chunk.

Correlation Matrix

To consider the correlation between multiple numeric variables, we could use `cor()` on every pair, but that would be tedious. Instead, we can compute a correlation *matrix*. To do this, we need to ensure that the data passed to `cor()` contains only numeric variables, as categorical variables cannot be included.

Unfortunately, this is not the case for the `gentoo` data. So we will need to subset the data to only include numeric variables.

To make this subset, we'll use the `subset()` function and the `select` argument. `select` is a vector of variable names in `gentoo`. Then, we can find the correlation of this subset that we will call `numericGentoo`.

Subsetting Data

This code has been provided to you in the `tryIt3` code chunk. Go ahead, take a peek, and hit the green run arrow to run this chunk.

```r
numericGentoo <- subset(gentoo,
                          select = c("bill_length_mm",
                                     "bill_depth_mm",
                                     "flipper_length_mm",
                                     "body_mass_g")
                          )
```

Correlation Matrix

Let's try this code in the `tryIt4` code chunk to see the correlation matrix of the numeric variables contained in the `gentoo` data. Don't forget to first run the `tryIt3` code chunk and verify that `numericGentoo` is in your environment!

```
cor(numericGentoo)
```

```
##                   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## bill_length_mm         1.0000000     0.6540233         0.6642052   0.6667302
## bill_depth_mm          0.6540233     1.0000000         0.7106422   0.7229672
## flipper_length_mm      0.6642052     0.7106422         1.0000000   0.7113053
## body_mass_g            0.6667302     0.7229672         0.7113053   1.0000000
```

Each "entry" in the correlation matrix is the correlation between the variables labeling that entry's row and column. So for example, the correlation between bill depth and bill length is about +0.6540.

Using the Correlation Matrix

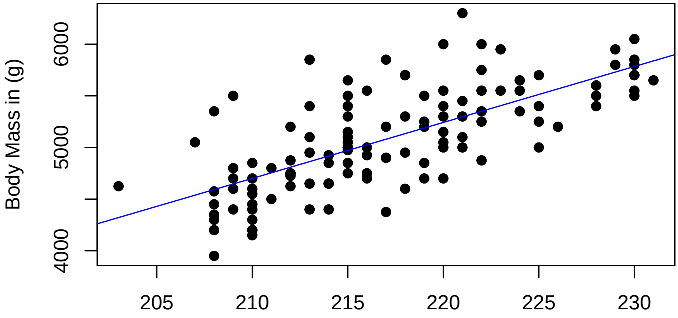
Which pair of variables has the strongest correlation (assuming that each pair in fact has a linear relationship)? The output is provided again below.

```
##                   bill_length_mm bill_depth_mm flipper_length_mm body_mass_g
## bill_length_mm         1.0000000     0.6540233         0.6642052   0.6667302
## bill_depth_mm          0.6540233     1.0000000         0.7106422   0.7229672
## flipper_length_mm      0.6642052     0.7106422         1.0000000   0.7113053
## body_mass_g            0.6667302     0.7229672         0.7113053   1.0000000
```

Linear Regression

Next time we will use R to find the equation of the line that best summarizes the relationship between flipper length and body mass. We will also talk about how to use this line to estimate flipper length for penguins of a particular body mass.

Scatter plot of Body Mass versus Flipper Length for Gentoos



What Questions Do You Have?