PSTAT 5LS Lab 7

YOUR NAME HERE

Week of June 3, 2024

Section 1

Announcements & Recap

Announcements

Insert any relevant announcements, important dates, things you want to remember here.

For instance, you might have them remember the date and time that homework and lab are due, and remind them that the deadlines are firm.

HW 6 Recap

Write down some common errors that you spotted while you graded/had office hours for HW 6.

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
4. Creating a linear model and finding the relevant values
5. Creating a plot of (x,y) quantitative values with the least-squares regression line.

Statistical Learning Objectives

1. Scatterplots with linear associations
2. Discussing the correlation coefficient
3. Discussing other important values in linear regression, such as R^2.
4. Discussing the least-squares regression line

Functions covered in this lab

1. `plot()`
2. `cor()`
3. `lm()`
4. `subset()`
5. `abline()`

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 in lecture that species may differ when it comes to physical measurements.

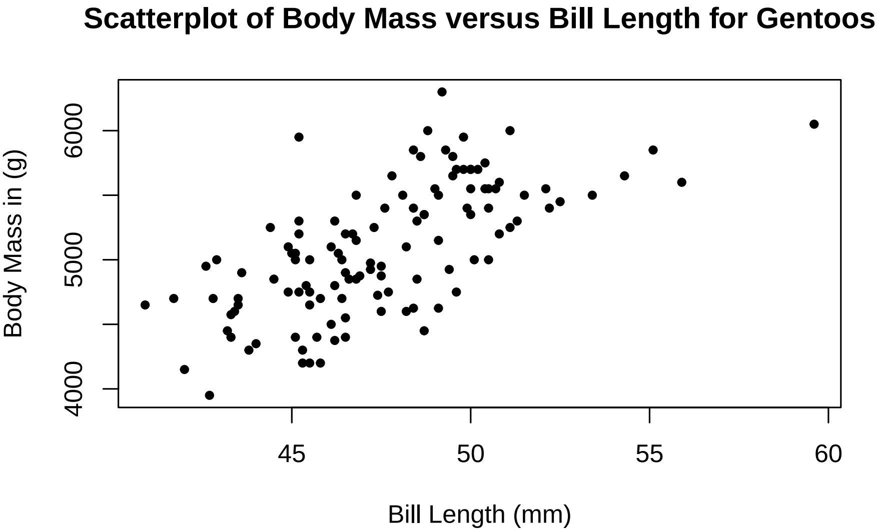
```
gentoo <- read.csv("gentoo.csv", stringsAsFactors = TRUE)
```

Go ahead and run the `loadGentoo` chunk in the `lab7-notes.Rmd` markdown file, and verify that the `gentoo` data is in your environment in the top right corner of your project.

Scatterplots with Linear Association

In lecture, we are focusing our attention on scatterplots 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`.

Scatterplot of Bill Length and Body Mass



Interpreting the Scatterplot

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

1. Direction (positive or negative)
2. Form (Linear or Nonlinear)
3. Strength (Weak, Moderate, Strong)
4. Outliers/Unusual Features (even if there are none, we should still comment this)

Let's interpret this scatterplot!

Interpret the Scatterplot of Bill Length and Body Mass

Let's start by commenting on the direction.

Interpret the Scatterplot of Bill Length and Body Mass

Next, comment on the form.

Interpret the Scatterplot of Bill Length and Body Mass

Then, discuss the strength.

Interpret the Scatterplot 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).

Scatterplot Code

Let's try this code out in the `tryit1` code chunk.

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

Be *very* careful setting up scatterplots!

- 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.

Strength and Correlation

In class, we have been observing scatterplots and commenting on the strength of the relationship. Earlier in our scatterplot, 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. Let's do so using the function `cor()`:

```
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

If we wanted to consider the correlation between multiple numeric variables, we could use `cor()` on every pair of them, but that's tedious. Instead, we'll compute a correlation *matrix*. In order to achieve this, we will have to make sure that the data we send to `cor()` is all numeric variables. It cannot contain categorical variables.

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.

```
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

We're going to perform a linear regression of body mass on flipper length. This means we're going to use the flipper length as the explanatory variable (x) and body mass as the response variable (y).

We'll use the function `lm()` (for **l**inear **m**odel), and provide it a formula (`y ~ x`) and a `data` argument. We'll store that as an object called `line1`. Then, to get detailed results, we'll use the `summary()` function.

Linear Regression Code

Let's try this code together in the `tryit5` code chunk.

```
line1 <- lm(body_mass_g ~ flipper_length_mm, data = gentoo)
summary(line1)
```

Linear Regression Output

Here's what the output looks like for our linear regression model.

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm, data = gentoo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -704.69 -244.29  -58.87  161.98 1003.65
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6674.204   1075.436  -6.206 8.51e-09 ***
## flipper_length_mm   54.165      4.948  10.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 354 on 117 degrees of freedom
## Multiple R-squared:  0.506,  Adjusted R-squared:  0.5017
## F-statistic: 119.8 on 1 and 117 DF,  p-value: < 2.2e-16
```

How to Read the Linear Regression Output

As we read this table:

- The first two lines are just the code we typed in being displayed.
- The next piece dealing with *residuals* can be skipped over for now.
- We want the piece dealing with the **coefficients**. In the *coefficients* portion of the output, there are two rows of information with four columns. The column we will be dealing with in this lab is the **Estimate** column.

How to Read the Linear Regression Output Continued

- The first row of information is called the **`(Intercept)`**. This represents information about the vertical (y) intercept of the regression line. So if we go to the `Estimate` column in the `(Intercept)` row, we will get the value of the vertical (y) intercept for the least-squares regression line.
- Notice that the next row of information is called **`flipper_length_mm`**, which is our explanatory (x) variable. This is a great way to verify that your logic of `y ~ x` was done correctly! This second row will always contain the name of the explanatory variable you chose. If we go to the `Estimate` column of the `flipper_length_mm` row, we will get the value of the slope for the least-squares regression line.

How to Read the Linear Regression Output Continued

- The next line has a value called the **residual standard error**, and this value is known as s.
- Then we will look at the line of output that has the **multiple R-squared** value – *ignore the adjusted R-squared value*.

What We Need from the Linear Regression Output

So again, the values we want to find from this output:

1. the vertical intercept of the least-squares regression line from our sample data
2. the slope of the least-squares regression line from our sample data
3. the residual standard error
4. the multiple r-squared value which is known as the *coefficient of determination*

What is the equation for the least-squares regression line?

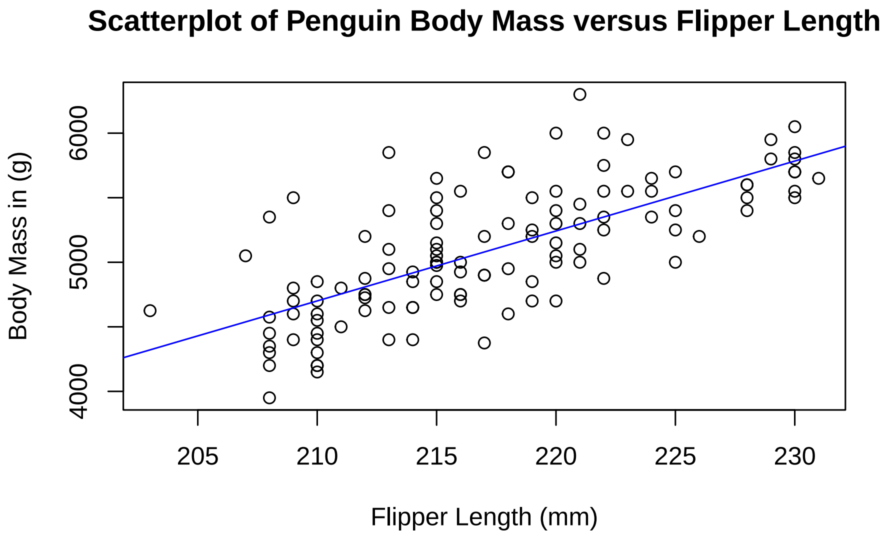
Adding the Regression Line to the Scatterplot

We can add the estimated least-squares regression line to our scatterplot by giving the model object to the `abline()` function.

Let's try this out in the `tryit6` in your notes.

```
plot(gentoo$body_mass_g ~ gentoo$flipper_length_mm,
     main = "Scatterplot of Body Mass versus Flipper Length fo
     xlab = "Flipper Length (mm)",
     ylab = "Body Mass in (g)")
abline(line1, col = "blue")
```

Scatterplot and Least-Squares Regression Line



Section 4

Questions

What Questions Do You Have?