Lecture 1- A Bracing Tour of Simple Linear Regression with R

Peter Shaffery

12/31/2020

Example: Penguins!

```
library(palmerpenguins)
```

Say that we are researchers studying penguins in the Palmer Archipelago of Antarctica. We have collected data on three different penguin species (Adelie, Chinstrap, and Gentoo) on three different islands (Torgersen, Biscoe, and Dream). For each subject in our dataset, we have recorded the following:

- length and depth of the bill (mm)
- length of the flippers (mm)
- body mass (g)
- sex
- year of the measurement

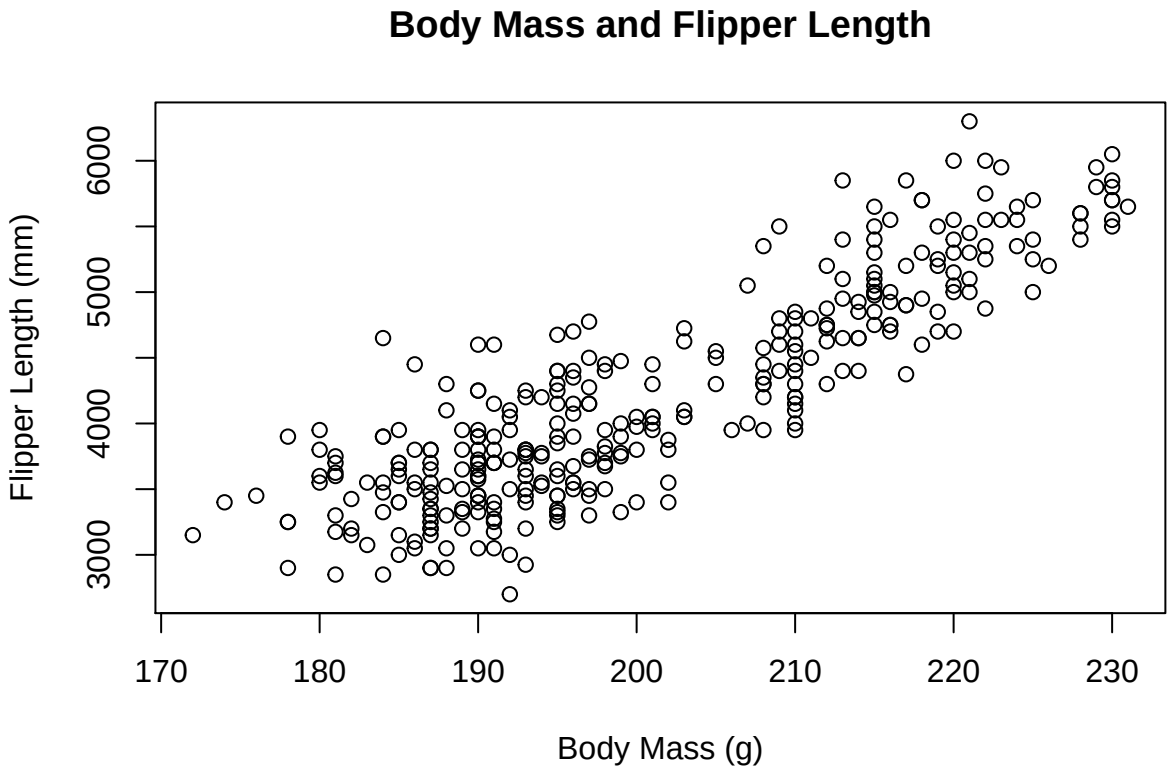
The data looks like this:

```
penguins
```

```
## # A tibble: 344 x 8
##    species island bill_length_mm bill_depth_mm flipper_length_~ body_mass_g
##    <fct>   <fct>           <dbl>         <dbl>            <int>       <int>
##  1 Adelie  Torge~           39.1          18.7              181        3750
##  2 Adelie  Torge~           39.5          17.4              186        3800
##  3 Adelie  Torge~           40.3          18                195        3250
##  4 Adelie  Torge~           NA            NA                 NA          NA
##  5 Adelie  Torge~           36.7          19.3              193        3450
##  6 Adelie  Torge~           39.3          20.6              190        3650
##  7 Adelie  Torge~           38.9          17.8              181        3625
##  8 Adelie  Torge~           39.2          19.6              195        4675
##  9 Adelie  Torge~           34.1          18.1              193        3475
## 10 Adelie  Torge~           42            20.2              190        4250
## # ... with 334 more rows, and 2 more variables: sex <fct>, year <int>
```

We are interested in the relationship between body mass (x) and flipper length (y)

```
dat = tidyr::drop_na(penguins)
x = dat$flipper_length_mm
y = dat$body_mass_g
plot(x,y,
     main='Body Mass and Flipper Length',
     xlab='Body Mass (g)',
     ylab='Flipper Length (mm)')
```

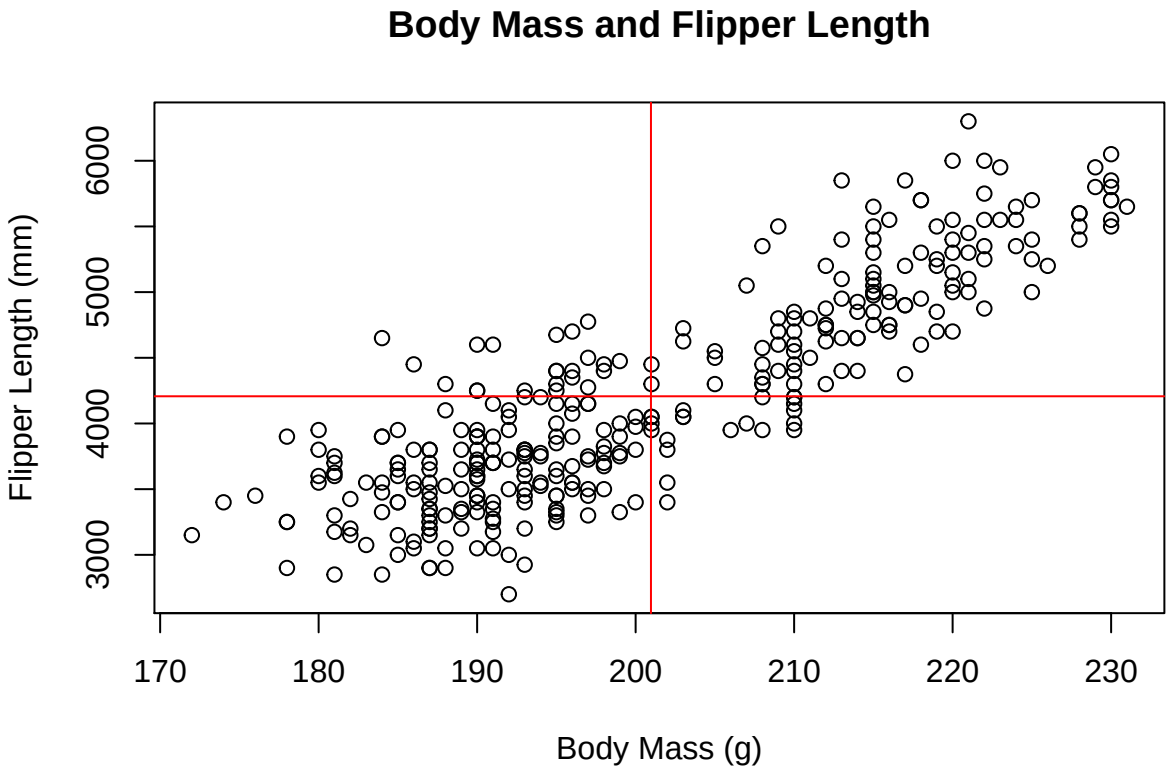


Correlation

They certainly *look* related, but how can we quantify this? One way is by using *correlation*. The idea here is to divide the scatter plot into quadrants, based on $\bar{x} = \frac{1}{n}\sum x_i$ and $\bar{y} = \frac{1}{n}\sum y_i$

```
x.bar = mean(x)
y.bar = mean(y)

plot(x,y,
     main='Body Mass and Flipper Length',
     xlab='Body Mass (g)',
     ylab='Flipper Length (mm)')
abline(v=x.bar, col='red')
abline(h=y.bar, col='red')
```



Notice that if flipper length *increases* with body mass than most points will in the upper right and lower left quadrants. That is to say, when the relationship is increasing then for most observations $x_i > \bar{x}$ and $y_i > \bar{y}$, or $x_i < \bar{x}$ and $y_i < \bar{y}$. Thus the product $(x_i - \bar{x})(y_i - \bar{y})$ will usually be *positive*. This leads us to define **covariance**:

$$\text{Cov}(x_i, y_i) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

When covariance is positive then *on average* points fall into either the upper right or lower left quadrants. When covariance is negative then the opposite must hold. **Cov(x, y) measures whether an increasing or decreasing relationship holds between x and y, on average**

```
n = length(x)
cov1 = (1/(n-1)) * sum((x-x.bar)*(y-y.bar))
cov2 = cov(x,y)

c(cov1,cov2)
```

```
## [1] 9852.192 9852.192
```

Say our collaborators use pounds (lbs) instead of grams (g). 1 pound is equal to about 454 grams, let's convert x to pounds, and calculate covariance again:

```
x.lbs = x/454
cov(x.lbs,y)
```

```
## [1] 21.70086
```

We got a totally different value! Note that the sign is still positive, but the value is two orders of magnitude smaller. It's therefore usually more convenient to work with *correlation*, which is scale invariant:

$$\text{Cor}(x_i, y_i) = \frac{\text{Cov}(x, y)}{s_x s_y}$$

Where s_x and s_y are the standard deviations of x and y:

$$s_x = \sqrt{\text{Var}[x]} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

```
cor(x,y)
```

```
## [1] 0.8729789
```

```
cor(x.lbs,y)
```

```
## [1] 0.8729789
```

Like covariance, correlation can be positive (when a relationship is increasing) or negative (when a relationship is decreasing), however unlike covariance correlations is bounded to the interval $[-1, 1]$.