STAT 8020 R Lab 1: Simple Linear Regression

Whitney Huang

8/20/2020

Lab Objective

- To gain experience with R, a programming language and free software environment for statistical computing and graphics.
- To learn how to use R to conduct a simple linear regression analysis

Example: Maximum Heart Rate vs. Age

The maximum heart rate (HR_{max}) of a person is often said to be related to age (Age) by the equation:

$$\text{HR}_{max} = 220 - \text{Age}$$

Let's use a dataset to assess this statement.

Setup

- You should have R installed, if not, open a web browser and go to (http://cran.r-project.org) and download and install R. It also helpful to install RStudo (http://rstudio.com).
- Create a folder for this R lab. Download the Maximum Heart Rate dataset at (http://whitneyhuang83.github.io/maxHeartRate.csv) and save it in the folder you just created.

Load the dataset

There are several ways to load a dataset into R:

- Importing Data over the Internet

```
dat <-read.csv('http://whitneyhuang83.github.io/STAT8010/Data/maxHeartRate.csv', header = T)
```

Let's take a look at the data

```
dat
```

```
##    Age MaxHeartRate
## 1   18          202
## 2   23          186
## 3   25          187
## 4   35          180
## 5   65          156
## 6   54          169
## 7   34          174
## 8   56          172
## 9   72          153
## 10  19          199
## 11  23          193
## 12  42          174
## 13  18          198
## 14  39          183
## 15  37          178
```

- Read the dataset from you computer

```
dat <- read.csv('maxHeartRate.csv', header = T)
```

- If the data is not too big, you can type the data into R

```
age <- c(18, 23, 25, 35, 65, 54, 34, 56, 72, 19, 23, 42, 18, 39, 37)
maxHeartRate <- c(202, 186, 187, 180, 156, 169, 174, 172, 153,
                  199, 193, 174, 198, 183, 178)
dat <- data.frame(cbind(age, maxHeartRate))
```

Examine the data before fitting models

```
summary(dat)
```

```
##       age         maxHeartRate
##  Min.   :18.00   Min.   :153.0
##  1st Qu.:23.00   1st Qu.:173.0
##  Median :35.00   Median :180.0
##  Mean   :37.33   Mean   :180.3
##  3rd Qu.:48.00   3rd Qu.:190.0
##  Max.   :72.00   Max.   :202.0
```

```
var(dat$age); var(dat$maxHeartRate)
```

```
## [1] 305.8095
```

```
## [1] 214.0667
```

```
cov(dat$age, dat$maxHeartRate)
```

```
## [1] -243.9524
```

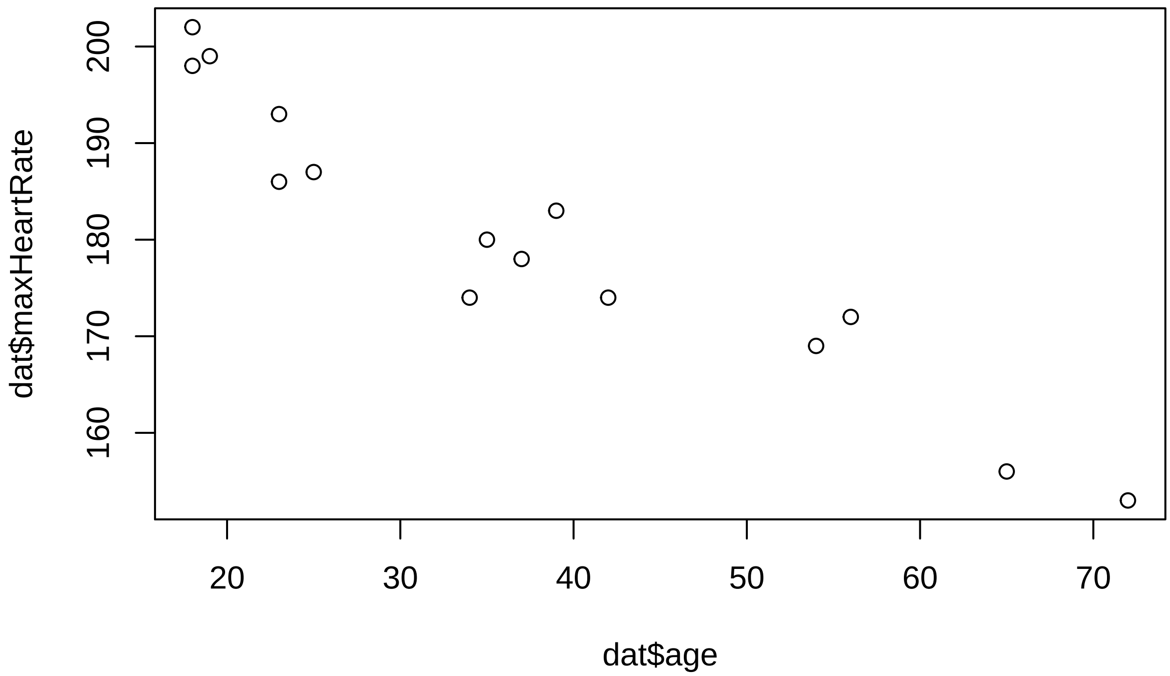
```
cor(dat$age, dat$maxHeartRate)
```

```
## [1] -0.9534656
```

Plot the data before fitting models

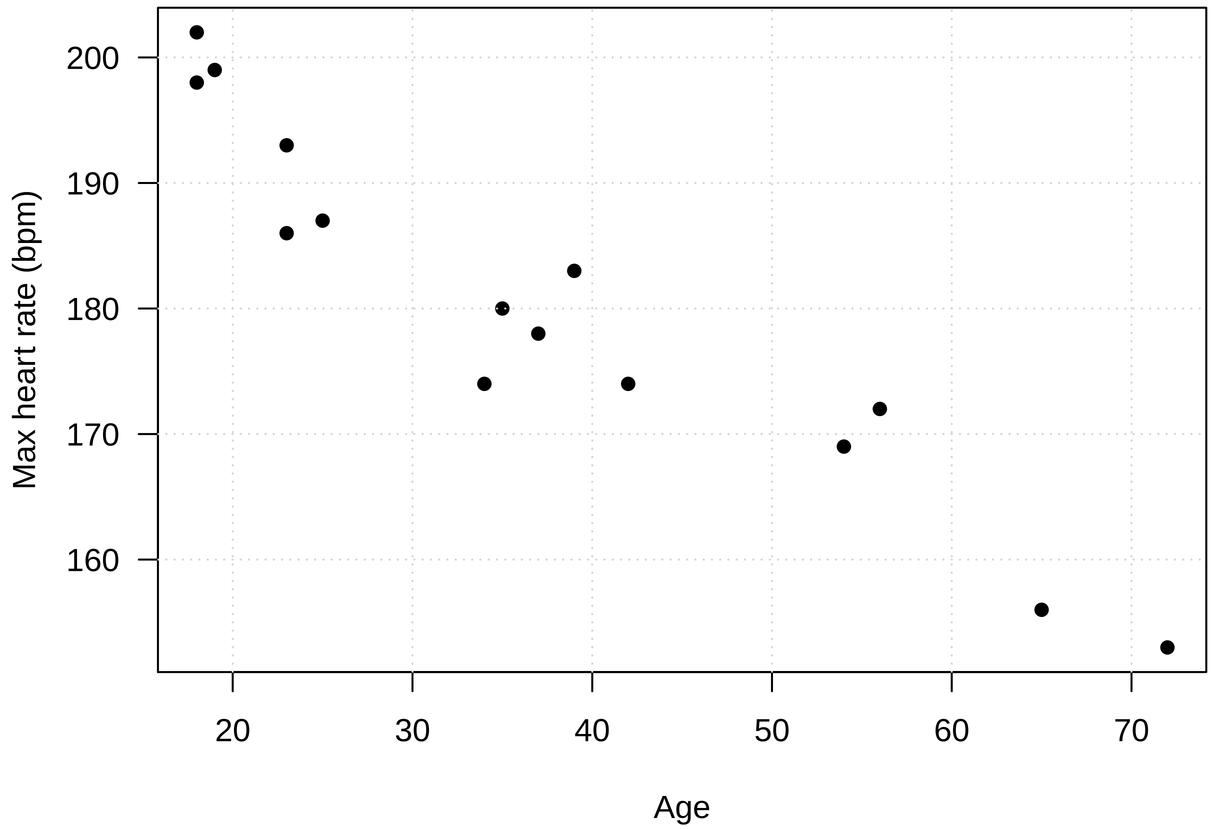
This is what the scatterplot would look like by default. Put predictor (age) to the first argument and response (maxHeartRate) to the second argument.

```
plot(dat$age, dat$maxHeartRate)
```



Let's make the plot look nicer (type ?plot to learn more).

```
par(las = 1, mar = c(4.1, 4.1, 1.1, 1.1))
plot(dat$age, dat$maxHeartRate,
     pch = 16, xlab = "Age",
     ylab = "Max heart rate (bpm)")
grid()
```



Question: Describe the direction, strength, and the form of the relationship.

Simple linear regression

Let's do the calculations to figure out the regression coefficients as well as the standard deviation of the random error.

- Slope

```
X <- dat$age; Y <- dat$maxHeartRate
Y_diff <- Y - mean(Y)
X_diff <- X - mean(X)
beta_1 <- sum(Y_diff * X_diff) / sum((X_diff)^2)
beta_1
```

```
## [1] -0.7977266
```

- Intercept

```
beta_0 <- mean(Y) - mean(X) * beta_1
beta_0
```

```
## [1] 210.0485
```

- Fitted values

```
Y_hat <- beta_0 + beta_1 * X
Y_hat
```

```
##  [1] 195.6894 191.7007 190.1053 182.1280 158.1962 166.9712 182.9258 165.3758
##  [9] 152.6121 194.8917 191.7007 176.5439 195.6894 178.9371 180.5326
```

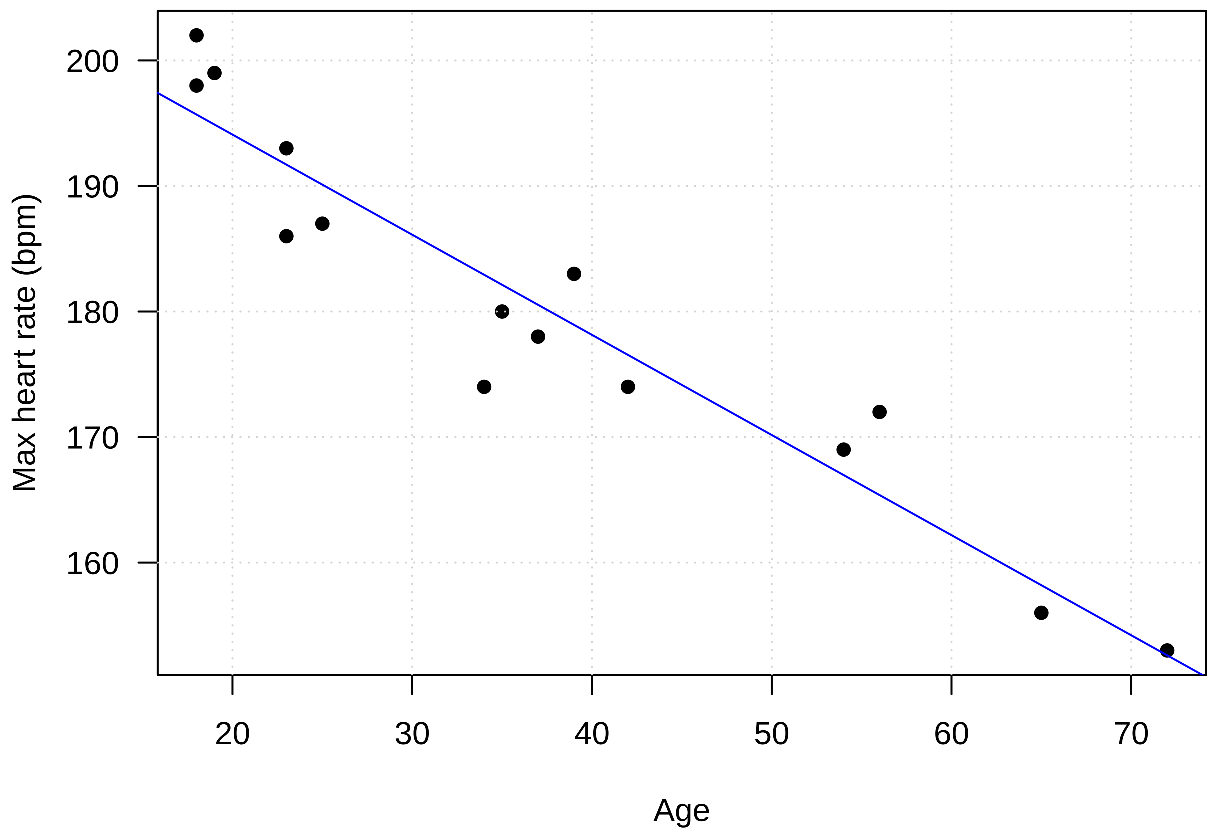
- $\hat{\sigma}$

```
sigma2 <- sum((Y - Y_hat)^2) / (length(Y) - 2)
sqrt(sigma2)
```

```
## [1] 4.577799
```

Add the fitted regression line to the scatterplot

```
par(las = 1, mar = c(4.1, 4.1, 1.1, 1.1))
plot(dat$age, dat$maxHeartRate,
     pch = 16, xlab = "Age",
     ylab = "Max heart rate (bpm)")
grid()
abline(a = beta_0, b = beta_1,
       col = "blue")
```



Let R do all the work

```
fit <- lm(maxHeartRate ~ age,
          data = dat)
summary(fit)
```

```
##
## Call:
## lm(formula = maxHeartRate ~ age, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -8.9258 -2.5383  0.3879  3.1867  6.6242
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
## age          -0.79773    0.06996  -11.40 3.85e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.578 on 13 degrees of freedom
## Multiple R-squared:  0.9091, Adjusted R-squared:  0.9021
## F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
```

- Regression coefficients

```r
fit$coefficients
```

```
## (Intercept)         age
## 210.0484584  -0.7977266
```

- Fitted values

```r
fit$fitted.values
```

```
##        1        2        3        4        5        6        7        8
## 195.6894 191.7007 190.1053 182.1280 158.1962 166.9712 182.9258 165.3758
##        9       10       11       12       13       14       15
## 152.6121 194.8917 191.7007 176.5439 195.6894 178.9371 180.5326
```

- $\hat{\sigma}$

```r
summary(fit)$sigma
```

```
## [1] 4.577799
```