DATA 605 Assignment 11

Bin Lin

2017-4-21

Using R’s lm function, perform regression analysis and measure the significance of the independent variables for the following two data sets. In the first case, you are evaluating the statement that we hear that Maximum Heart Rate of a person is related to their age by the following equation:

$$MaxHR = 220 - Age$$

You have been given the following sample:

Age 18 23 25 35 65 54 34 56 72 19 23 42 18 39 37

MaxHR 202 186 187 180 156 169 174 172 153 199 193 174 198 183 178

Perform a linear regression analysis fitting the Max Heart Rate to Age using the lm function in R.

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

model1 <- lm(maxhr ~ age)
summary(model1)
```

```
## 
## Call:
## lm(formula = maxhr ~ age)
## 
## 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
```

What is the resulting equation? Is the effect of Age on Max HR significant?

Based on the summary statistics of the linear regression model, the resulting formular is: $MaxHR = 210 - 0.798 * Age + \varepsilon$

The effect of Age on MaxHR is significant because the p-value is 3.848e-08, which is very closed to 0, which is less than the significant threshold (typically 0.01)

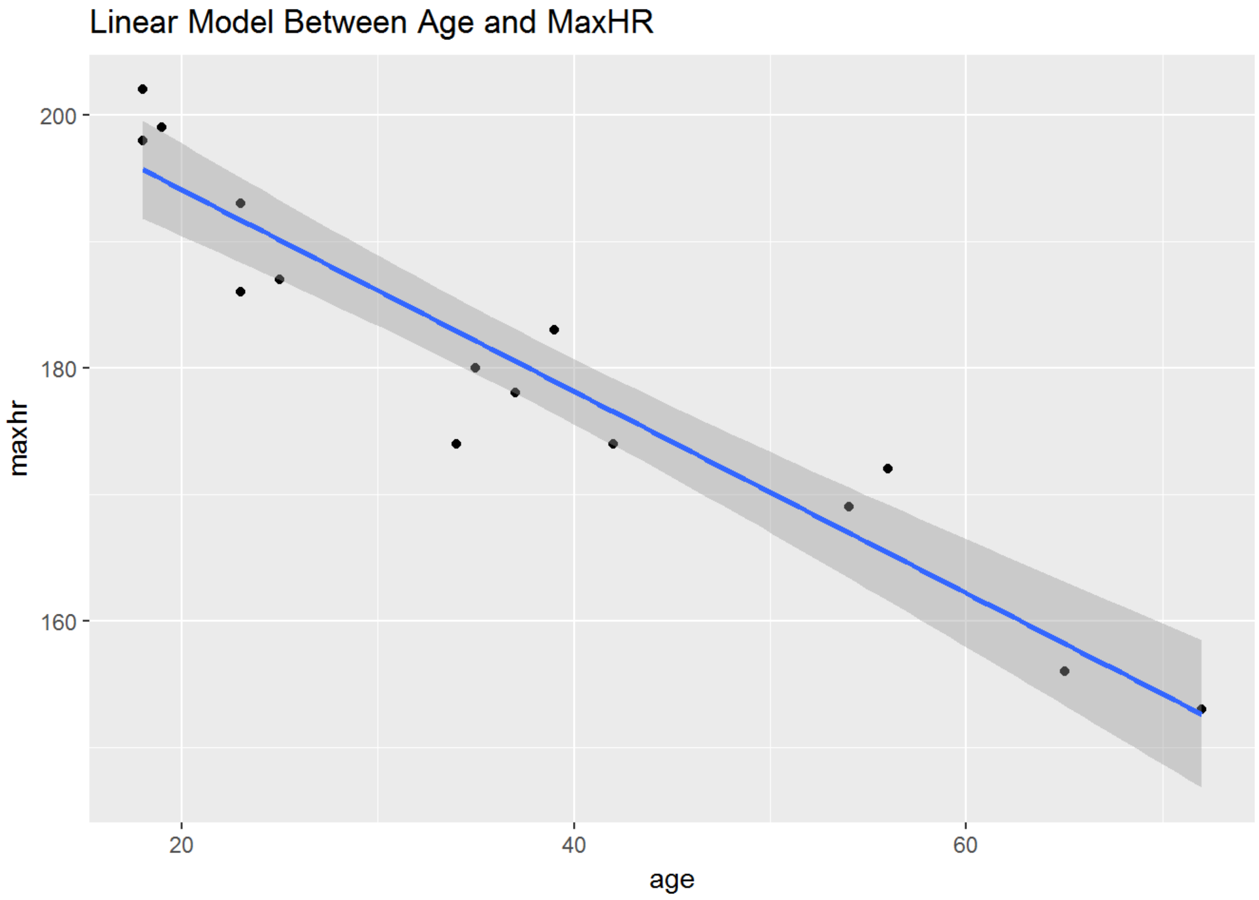
What is the significance level? Please also plot the fitted relationship between Max HR and Age.

Based on the summary statistics of the model, the significance level of this model can be as small as 0.001. Even at this level, the effect of Age is significant, because p-value is almost 0.

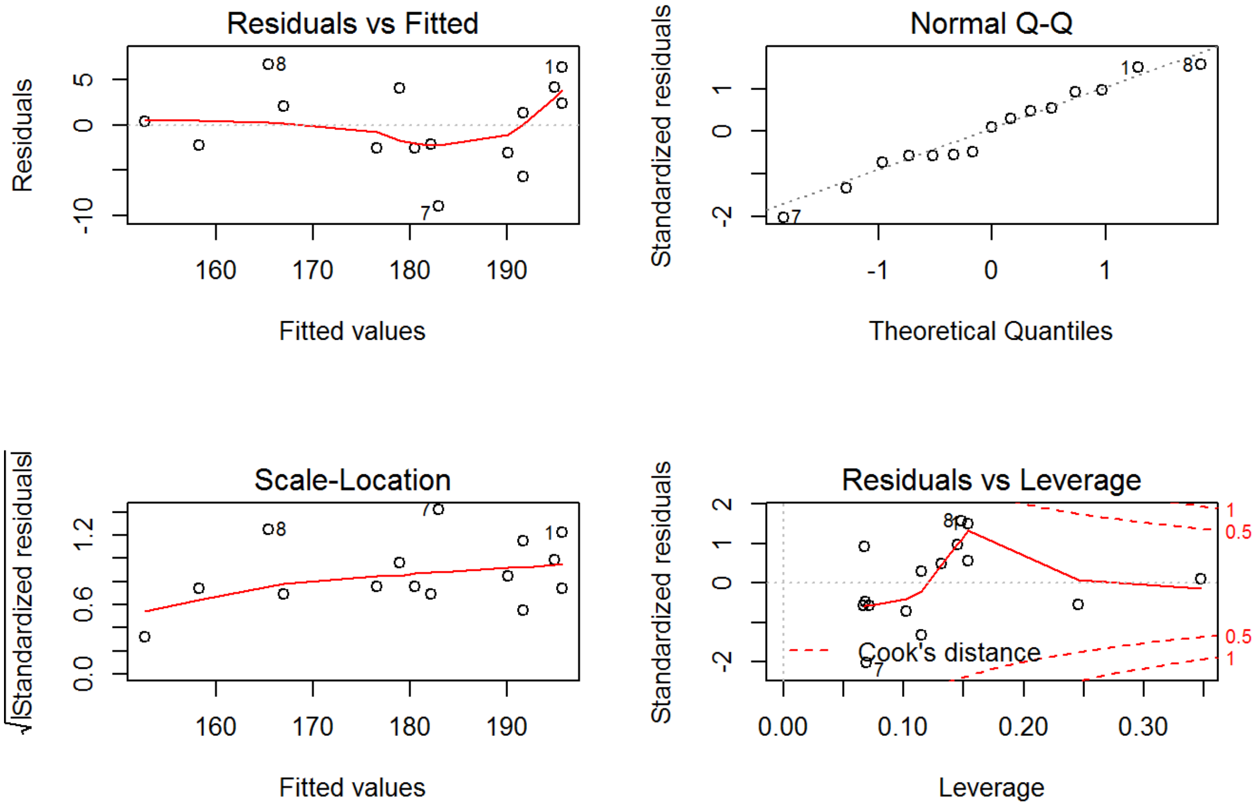
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.3.3
```

```
ggplot(data = data.frame(age, maxhr), aes(x = age, y = maxhr)) + geom_point() + geom_smooth(method =
"lm") + ggtitle("Linear Model Between Age and MaxHR")
```



```
par(mfrow=c(2,2))
plot(model1)
```



Using the Auto data set from Assignment 5 (also attached here) perform a Linear Regression analysis using mpg as the dependent variable and the other 4 (displacement, horsepower, weight, acceleration) as independent variables.

```
auto <- read.table("C:/Users/blin261/Desktop/DATA605/assign11/auto-mpg.data", stringsAsFactors = FALS
E)
colnames(auto) <- c("displacement", "horsepower", "weight", "acceleration", "mpg")
head(auto)
```

```
##   displacement horsepower weight acceleration mpg
## 1          307        130   3504         12.0  18
## 2          350        165   3693         11.5  15
## 3          318        150   3436         11.0  18
## 4          304        150   3433         12.0  16
## 5          302        140   3449         10.5  17
## 6          429        198   4341         10.0  15
```

```
model2 <- lm(mpg~., data = auto)
summary(model2)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.378  -2.793  -0.333   2.193  16.256
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  45.2511397  2.4560447  18.424  < 2e-16 ***
## displacement -0.0060009  0.0067093  -0.894  0.37166
## horsepower   -0.0436077  0.0165735  -2.631  0.00885 **
## weight       -0.0052805  0.0008109  -6.512  2.3e-10 ***
## acceleration -0.0231480  0.1256012  -0.184  0.85388
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.247 on 387 degrees of freedom
## Multiple R-squared:  0.707,  Adjusted R-squared:  0.704
## F-statistic: 233.4 on 4 and 387 DF,  p-value: < 2.2e-16
```

What is the final linear regression fit equation? Which of the 4 independent variables have a significant impact on mpg? What are their corresponding significance levels?

Tha final linear regression fit equation of raw data is:
$mpg = 45.251 - 0.006 * displacement - 0.044 * horsepower - 0.005 * weight - 0.023 * acceleration + \varepsilon$

```
#Significance level of corresponding variables.
summary(model2)$coefficients[,4]
```

```
##  (Intercept) displacement   horsepower       weight acceleration
## 7.072099e-55 3.716584e-01 8.848982e-03 2.302545e-10 8.538765e-01
```

horsepower and weight have significant impact on mpg. The significance level of horsepower is 0.01, while the significance level of weight is 0.001.

What are the standard errors on each of the coefficients?

```
summary(model2)$coefficients[,2]
```

```
##  (Intercept) displacement   horsepower       weight acceleration
## 2.4560446927 0.0067093055 0.0165734633 0.0008108541 0.1256011622
```

```
#Confidence interval of raw data for independent variables.
confint(model2, level = 0.95)
```

```
##                     2.5 %       97.5 %
## (Intercept)  40.422278855 50.080000544
## displacement -0.019192122  0.007190380
## horsepower   -0.076193029 -0.011022433
## weight       -0.006874738 -0.003686277
## acceleration -0.270094049  0.223798050
```

Please perform this experiment in two ways. First take any random 40 data points from the entire auto data sample and perform the linear regression fit and measure the 95% confidence intervals. Then, take the entire data set (all 392 points) and perform linear regression and measure the 95% confidence intervals. Please report the resulting fit equation, their signficance values and confidence intervals for each of the two runs.

```
set.seed(88)
random <- sample(1:nrow(auto), 40, replace = FALSE)
sample_data <- auto[random,]
model3 <- lm(mpg ~ ., data = sample_data)
summary(model3)
```

```
## 
## Call:
## lm(formula = mpg ~ ., data = sample_data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -5.0552 -2.3324 -0.1078  1.7459  5.5419 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  53.957999   7.876265   6.851 5.96e-08 ***
## displacement  0.009424   0.014643   0.644  0.52404    
## horsepower   -0.033924   0.054534  -0.622  0.53792    
## weight       -0.008093   0.002364  -3.423  0.00159 ** 
## acceleration -0.278362   0.459051  -0.606  0.54817    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 3.135 on 35 degrees of freedom
## Multiple R-squared:  0.831,  Adjusted R-squared:  0.8117 
## F-statistic: 43.03 on 4 and 35 DF,  p-value: 4.773e-13
```

```
#Significance level of corresponding variables.
summary(model3)$coefficients[,4]
```

```
##  (Intercept) displacement   horsepower       weight acceleration 
## 5.958780e-08 5.240352e-01 5.379242e-01 1.592553e-03 5.481721e-01
```

```
#Confidence interval of sample data for independent variables.
confint(model3, level = 0.95)
```

```
##                   2.5 %       97.5 %
## (Intercept)  37.96833162 69.947666538
## displacement -0.02030243  0.039150357
## horsepower   -0.14463381  0.076785488
## weight       -0.01289267 -0.003293761
## acceleration -1.21028386  0.653560696
```

Tha final linear regression fit equation of sample data is:

$$mpg = 53.958 + 0.009 * displacement - 0.034 * horsepower - 0.008 * weight - 0.278 * acceleration + \varepsilon$$

Conclusion: On the sample data, only weight has significant impact on mpg. However, on the full data, both horsepower and weight are significant. This means the larger the size of the sample, the easier we will detect the effects of independent variables. Also by comparing the confidence interval between two datasets, we know the larger samples size, the tighter the confidence interval. On the other words, the larger sample size, the smaller standard errors. (Because confidence interval is basically point estimate +/- standard error)