FE590. Assignment #1 (Gang Ping Zhu)

2017-09-22

Question 1

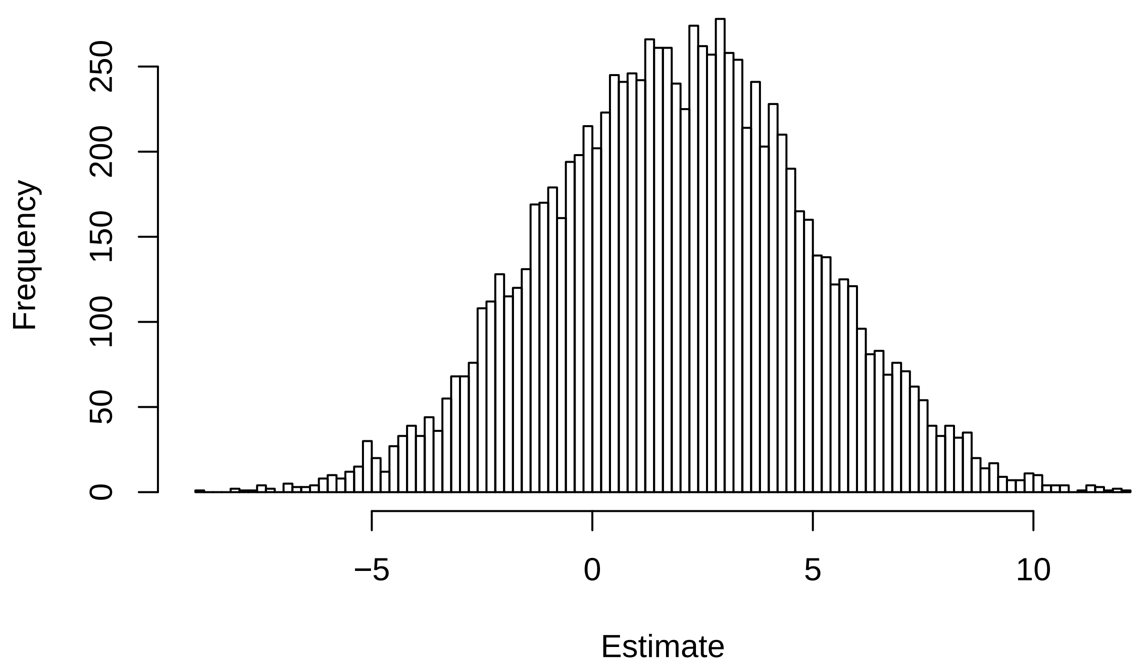
Question 1.1

Generate a vector `x` containing 10,000 realizations of a random normal variable with mean 2.0 and standard deviation 3.0, and plot a histogram of `x` using 100 bins. To get help generating the data, you can type `?rnorm` at the R prompt, and to get help with the histogram function, type `?hist` at the R prompt.

Solution:

```
x <- rnorm(10000, mean = 2.0, sd = 3)

hist(x, 100, main = "", xlab = "Estimate")
```



Question 1.2

Confirm that the mean and standard deviation are what you expected using the commands `mean` and `sd`.

Solution:

```r
m <- mean(x)
s <- sd(x)

c(m, s)
```

```
## [1] 1.948401 3.001232
```

The standard deivation and the mean are close to what is expected when creating the random data.

Question 1.3

Using the `sample` function, take out 10 random samples of 500 observations each. Calculate the mean of each sample. Then calculate the mean of the sample means and the standard deviation of the sample means.

Solution:

```r
randmatrix <- matrix(NA, 500, 10)
set.seed(10)

for (k in 1:10)
{
  rsample <- sample(rnorm(10000),500);
  randmatrix[,k] <- rsample;
}

s1 <- mean(randmatrix[,1])
s2 <- mean(randmatrix[,2])
s3 <- mean(randmatrix[,3])
s4 <- mean(randmatrix[,4])
s5 <- mean(randmatrix[,5])
s6 <- mean(randmatrix[,6])
s7 <- mean(randmatrix[,7])
s8 <- mean(randmatrix[,8])
s9 <- mean(randmatrix[,9])
s10 <- mean(randmatrix[,10])

mtotal <- c(s1, s2, s3, s4, s5, s6, s7, s8, s9, s10)

meanssample <- mean(mtotal)
meanssample
```

```
## [1] 0.01288664
```

```r
stdsample <-sd(mtotal)
stdsample
```

```
## [1] 0.05058528
```

Question 2

Sir Francis Galton was a controversial genius who discovered the phenomenon of "Regression to the Mean." In this problem, we will examine some of the data that illustrates the principle.

Question 2.1

First, install and load the library `HistData` that contains many famous historical data sets. Then load the Galton data using the command `data(Galton)`. Take a look at the first few rows of `Galton` data using the command `head(Galton)`.

Solution:

```
library('HistData')
data(Galton)

head(Galton)
```

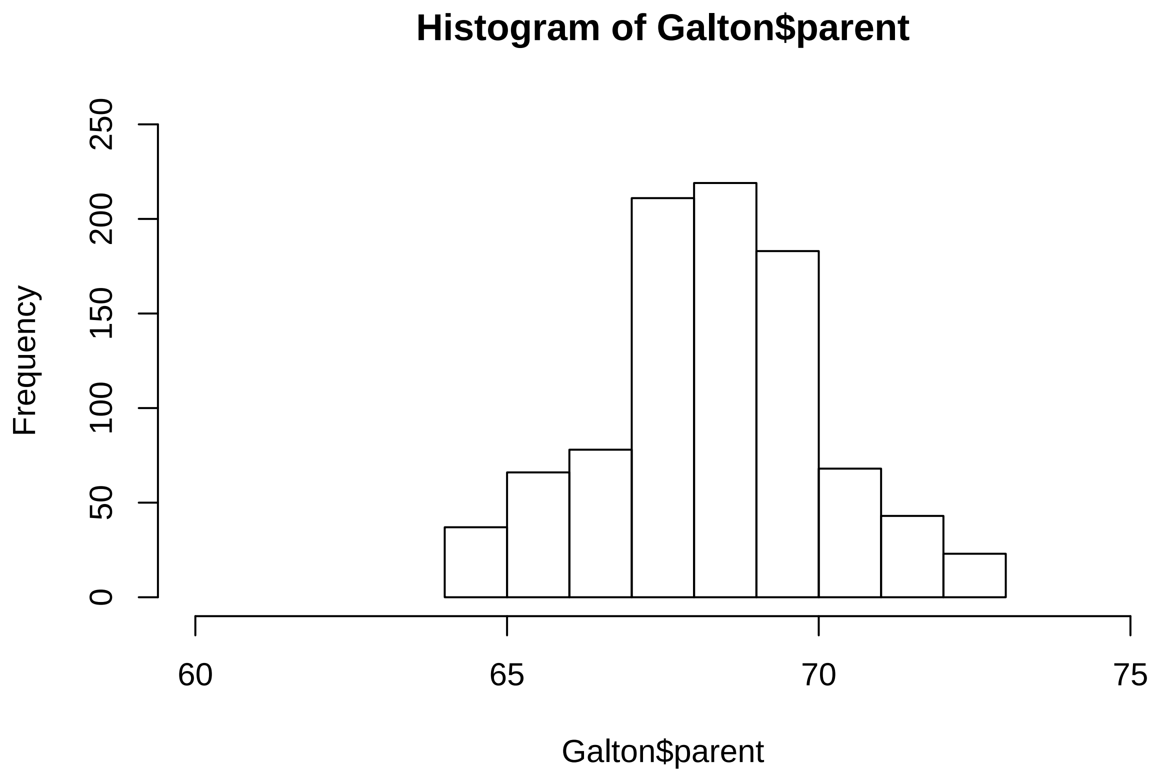
```
##   parent child
## 1   70.5  61.7
## 2   68.5  61.7
## 3   65.5  61.7
## 4   64.5  61.7
## 5   64.0  61.7
## 6   67.5  62.2
```

As you can see, the data consist of two columns. One is the height of a parent, and the second is the height of a child. Both heights are measured in inches.

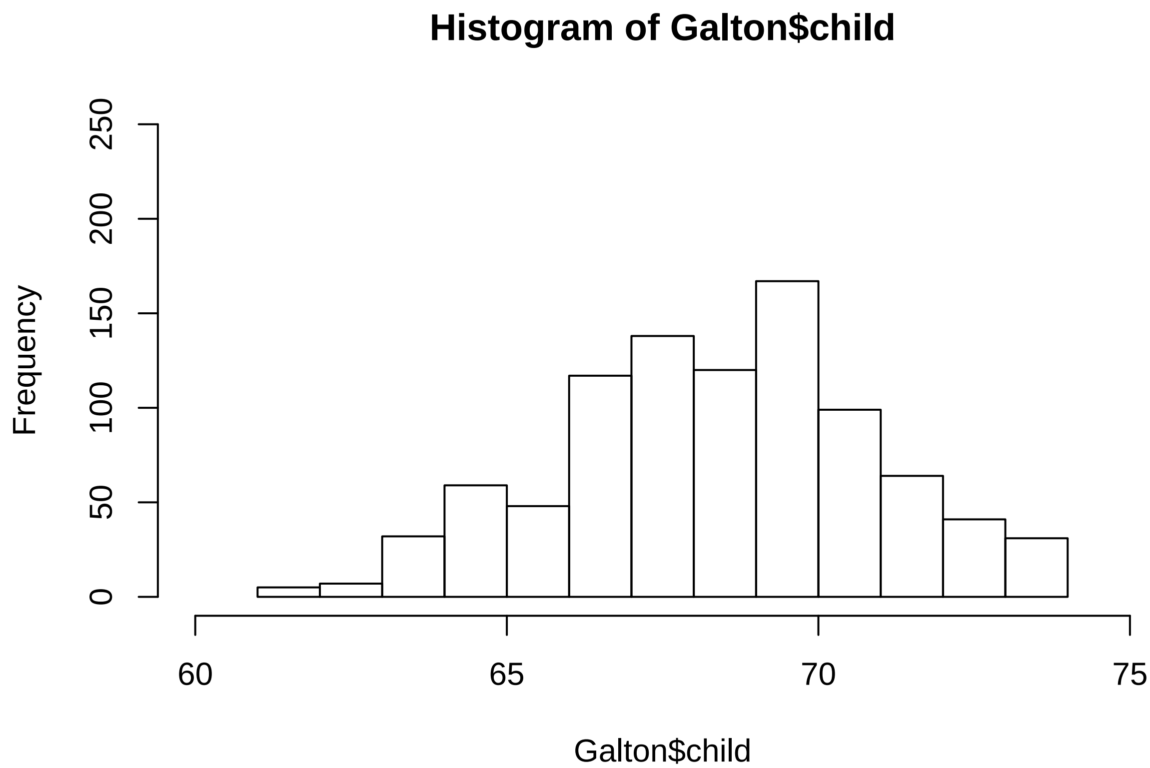
Plot one histogram of the heights of the children and one histogram of the heights of the parents. This histograms should use the same `x` and `y` scales.

Solution:

```
hist(Galton$parent, ylim = c(0,250), xlim = c(60, 75))
```




```
hist(Galton$child, ylim = c(0,250), xlim = c(60, 75))
```



Comment on the shapes of the histograms.

Solution:

The shape of the parent histogram is more narrow than the child histogram. This should be expected as adult heights shouldn't vary as much as childrens' heights. The child histogram is more spread out as there could be varying ages among the children that cause their heights to be different.

Question 2.2

Make a scatterplot the height of the child as a function of the height of the parent. Label the x-axis "Parent Height (inches)," and label the y-axis "Child Height (inches)." Give the plot a main tile of "Galton Data."

Perform a linear regression of the child's height onto the parent's height. Add the regression line to the scatter plot.

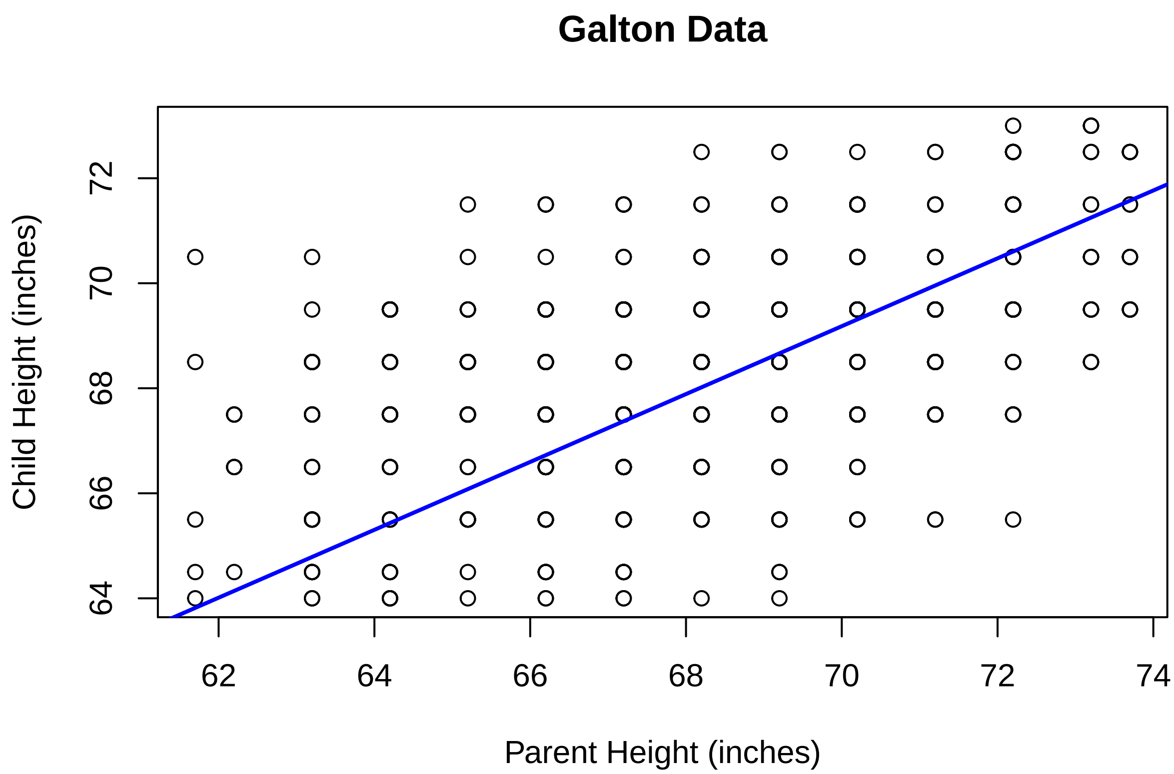
Using the `summary` command, print a summary of the linear regression results.

Solution:

```
plot(x = Galton$child, y = Galton$parent, type = "p", main = "Galton Data", xlab = "Parent Height (inche

linreg <- lm(Galton$child ~ Galton$parent, data = Galton)
```

```r
abline(linreg,col="blue",lwd=2);
```




```r
summary(linreg)
```

```
##
## Call:
## lm(formula = Galton$child ~ Galton$parent, data = Galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8050 -1.3661  0.0487  1.6339  5.9264
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   23.94153    2.81088   8.517   <2e-16 ***
## Galton$parent  0.64629    0.04114  15.711   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.239 on 926 degrees of freedom
## Multiple R-squared:  0.2105, Adjusted R-squared:  0.2096
## F-statistic: 246.8 on 1 and 926 DF,  p-value: < 2.2e-16
```

```r
# Enter your R code here!
```

What is the slope of the line relating a child's height to the parent's height? Can you guess why Galton says that there is a "regression to the mean"?

Solution:

The slope of the line is 0.6462906. With the values distributed, you can see a trend of the heights of both parents and children converging towards the mean. Even with the outliers, you can somewhat see that the mean of the heights is where most of the group of parents and children land. The slope gives that indication as well with its direction and angle.

Is there a significant relationship a child's height to the parent's height? If so, how can you tell from the regression summary?

Solution:

Yes, it appears there is a significant relationship between the two as the summary of the linear model has provided that coefficients are very significant.

Question 3

If necessary, install the `ISwR` package, and then `attach` the `bp.obese` data from the package. The data frame has 102 rows and 3 columns. It contains data from a random sample of Mexican-American adults in a small California town.

Question 3.1

The variable `sex` is an integer code with 0 representing male and 1 representing female. Use the `table` function operation on the variable 'sex' to display how many men and women are represented in the sample.

Solution:

```r
library('ISwR')
attach(bp.obese)
summary(bp.obese)
```

```
##       sex             obese             bp       
##  Min.   :0.0000   Min.   :0.810   Min.   : 94.0  
##  1st Qu.:0.0000   1st Qu.:1.143   1st Qu.:116.0  
##  Median :1.0000   Median :1.285   Median :124.0  
##  Mean   :0.5686   Mean   :1.313   Mean   :127.0  
##  3rd Qu.:1.0000   3rd Qu.:1.430   3rd Qu.:137.5  
##  Max.   :1.0000   Max.   :2.390   Max.   :208.0
```

```r
head(bp.obese)
```

```
##   sex obese  bp
## 1   0  1.31 130
## 2   0  1.31 148
## 3   0  1.19 146
## 4   0  1.11 122
## 5   0  1.34 140
## 6   0  1.17 146
```

```r
table(sex)
```

```
## sex
##  0  1
## 44 58
```

Question 3.2

The `cut` function can convert a continuous variable into a categorical one. Convert the blood pressure variable `bp` into a categorical variable called `bpc` with break points at 80, 120, and 240. Rename the levels of `bpc` using the command `levels(bpc) <- c("low", "high")`.

Solution:

```r
bpc <- cut(bp, breaks =c(80, 120, 240))
levels(bpc) <- c("low", "high")
```

Question 3.3

Use the `table` function to display a relationship between `sex` and `bpc`.

Solution:

```r
newtable <- table(sex, bpc)
newtable
```

```
##    bpc
## sex low high
##   0  16   28
##   1  28   30
```

Question 3.4

Now cut the `obese` variable into a categorical variable `obesec` with break points 0, 1.25, and 2.5. Rename the levels of `obesec` using the command `levels(obesec) <- c("low", "high")`.

Use the `ftable` function to display a 3-way relationship between `sex`, `bpc`, and `obesec`.

Solution:

```r
obesec <- cut(obese, breaks = c(0, 1.25, 2.5))

levels(obesec) <- c("low", "high")

ftable(sex, bpc, obesec)
```

```
##          obesec low high
## sex bpc
## 0   low          12    4
##     high         15   13
## 1   low          14   14
##     high          4   26
```

Which group do you think is most at risk of suffering from obesity?

Solution:

The high blood pressure females are likely to suffer from obseity. From the information below, we can see that this has the highest count across the categories.

```r
ftable(sex, bpc, obesec)[8]
```

```
## [1] 26
```

Question 4

Using the Boston data in the MASS library, run a linear regression fit to determine a predictive model for the median value of a home using the indicators of rooms per dwelling and the property tax.

```r
library('MASS')
data(Boston)
help(Boston)

linregfit <- lm(Boston$medv ~ Boston$rm + Boston$tax)
summary(linregfit)
```

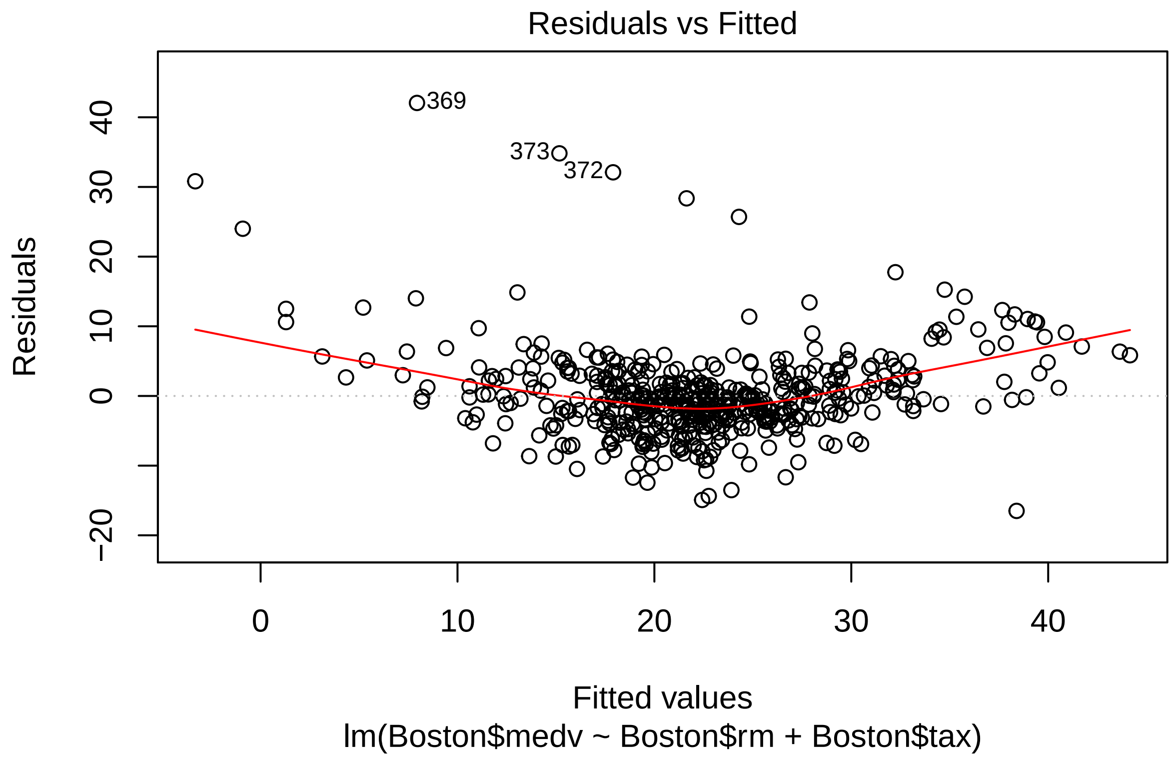
```
##
## Call:
## lm(formula = Boston$medv ~ Boston$rm + Boston$tax)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.495  -3.123  -0.548   2.384  42.057
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -21.233093   2.834371  -7.491 3.09e-13 ***
## Boston$rm     7.992681   0.404534  19.758  < 2e-16 ***
## Boston$tax   -0.015837   0.001686  -9.391  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.109 on 503 degrees of freedom
## Multiple R-squared:  0.5606, Adjusted R-squared:  0.5588
## F-statistic: 320.8 on 2 and 503 DF,  p-value: < 2.2e-16
```

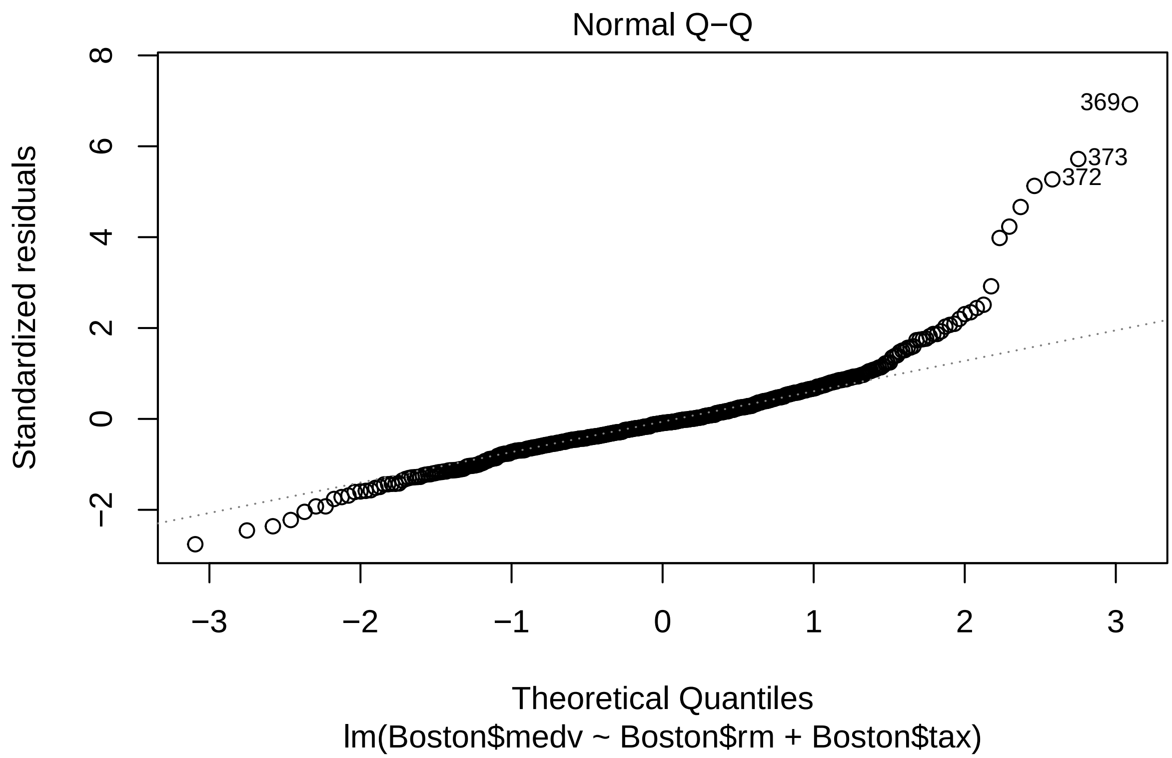
```r
dfpredict <- predict(linregfit, data.frame( rm=c(5,10,15), tax = c(5,10,15)),interval = "prediction")
head(dfpredict)
```

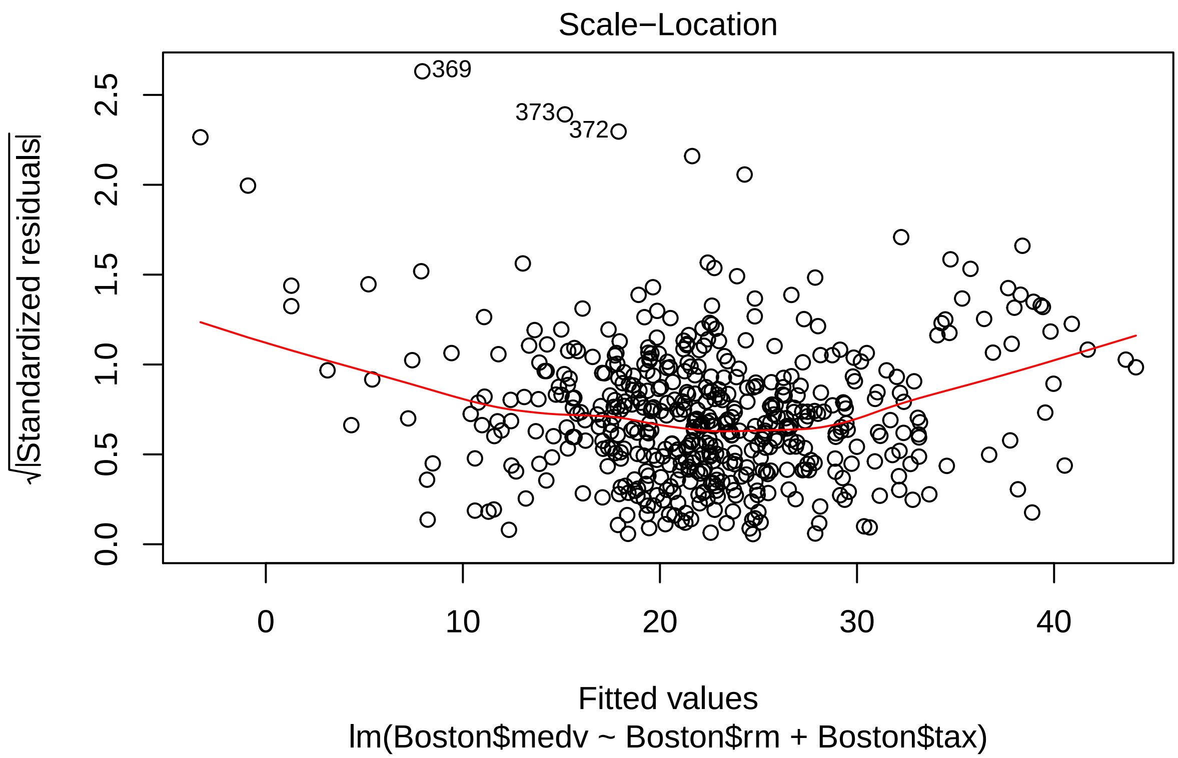
```
##        fit      lwr      upr
## 1 26.63109 14.61131 38.65086
```

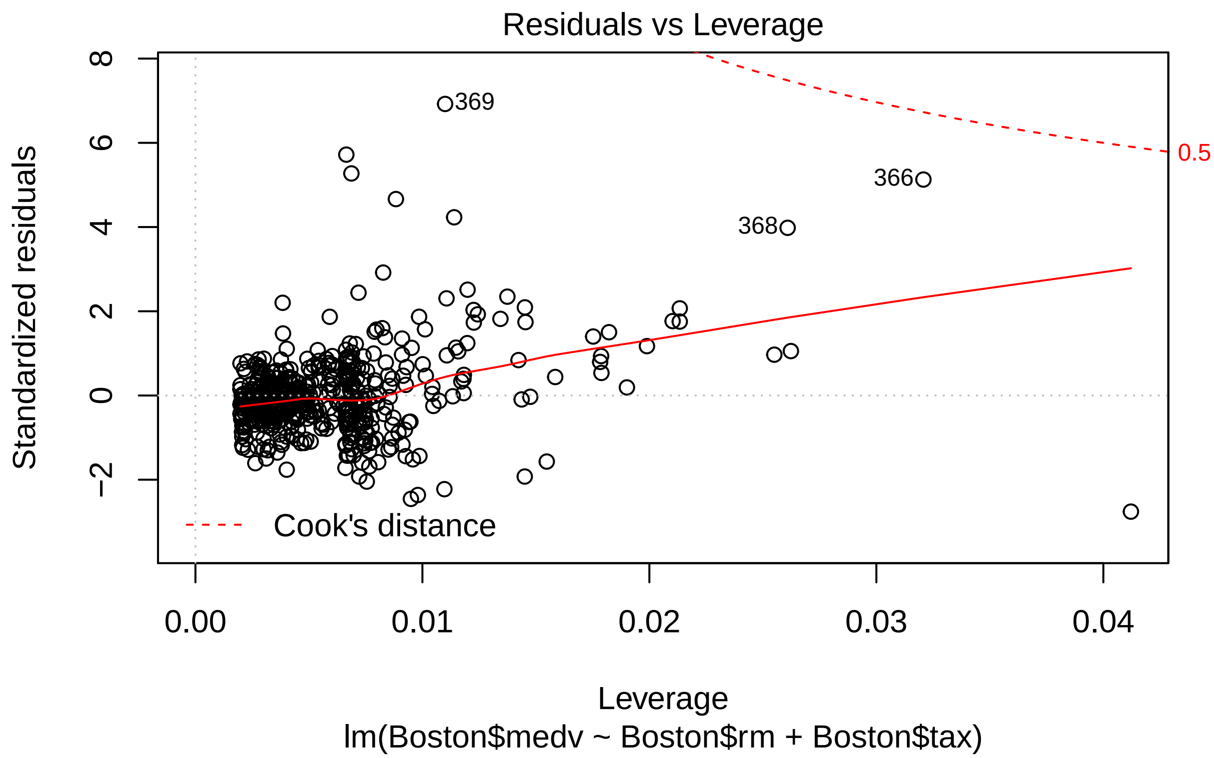
```
## 2 26.25540 14.22985 38.28095
## 3 32.36181 20.32359 44.40003
## 4 31.18392 19.14932 43.21851
## 5 32.37483 20.33584 44.41381
## 6 26.64407 14.61552 38.67263
```

```r
plot(linregfit)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Boston$medv ~ Boston$rm + Boston$tax)
369
373
372
−2
0
2
4
6
8
−3
−2
−1
0
1
2
3


Scale−Location
369
373
372
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
10
20
30
40
Fitted values
lm(Boston$medv ~ Boston$rm + Boston$tax)



Is there evidence that the indicators are useful(why or why not)?

Solution:

There is evidence that the indicators are useful. You can see in the our graphs that our fits are close to staight lines and aligning to what our data represents. Within the summary, we can see that both rooms per dwelling and property tax are both significant codes. We also see that the p-value of $< 2.2e\text{-}16$ determines that indicators should be used.