Chapter 3 - Linear Regression

Thalles Quinaglia Liduares

02/03/2022

Applied Exercise 3.8

Upload packages

```
library(lmreg)
library(tibble)
library(dplyr)
library(readxl)
```

Upload database

```
setwd("C:\\Program Files\\R\\Machine Learning")

data<-readxl::read_excel("auto-mpg.xlsx")

str(data)
```

```
## tibble [398 x 9] (S3: tbl_df/tbl/data.frame)
##  $ mpg         : num [1:398] 18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num [1:398] 8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num [1:398] 307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num [1:398] 130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num [1:398] 3504 3693 3436 3433 3449 ...
##  $ acceleration: num [1:398] 12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ model year  : num [1:398] 70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num [1:398] 1 1 1 1 1 1 1 1 1 1 ...
##  $ car name    : chr [1:398] "chevrolet chevelle malibu" "buick skylark 320" "plymouth sat
ellite" "amc rebel sst" ...
```

8. This question involves the use of simple linear regression on the Auto data set.

(a) Use the `lm()` function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the `summary()` function to print the results.

```
lm1<-lm(data$mpg~data$horsepower)

summary(lm1)
```

```
##
## Call:
## lm(formula = data$mpg ~ data$horsepower)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     39.935861   0.717499   55.66   <2e-16 ***
## data$horsepower -0.157845   0.006446  -24.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 2.2e-16
```

The estimated equation is given by

$$\widehat{mpg} = 39.9 - 0.15mpg$$

Hence, there's a negative relationship between these two variables. Both the intercept and slope coefficient are statistically significant at the p-value <0.001.

For each additional `hp` in the car, the `mpg` diminishes 0.15

If `hp` =98, then $mpg = 39.9 - 0.15(98) = 25.2$

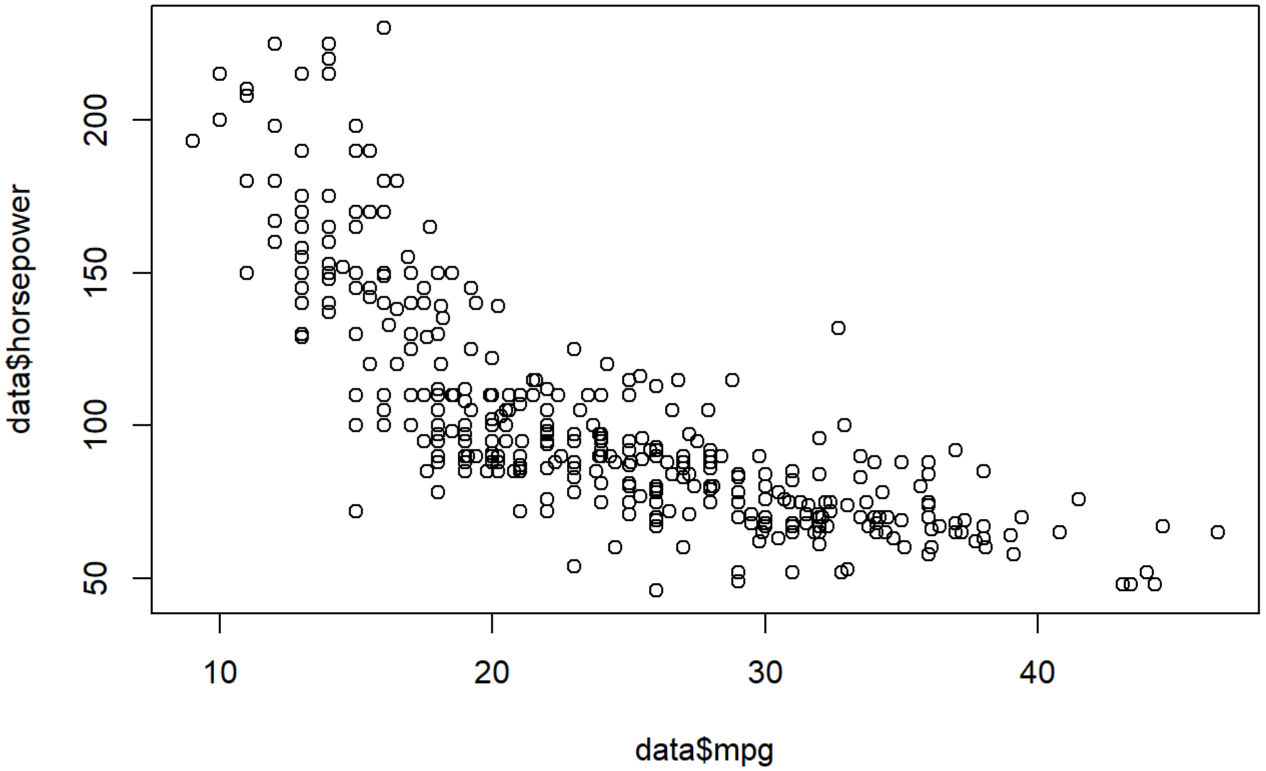
The confidence interval is equal to

```
confint(lm1)
```

```
##                     2.5 %     97.5 %
## (Intercept)     38.525212 41.3465103
## data$horsepower -0.170517 -0.1451725
```

(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.

```
plot(data$mpg, data$horsepower)
```



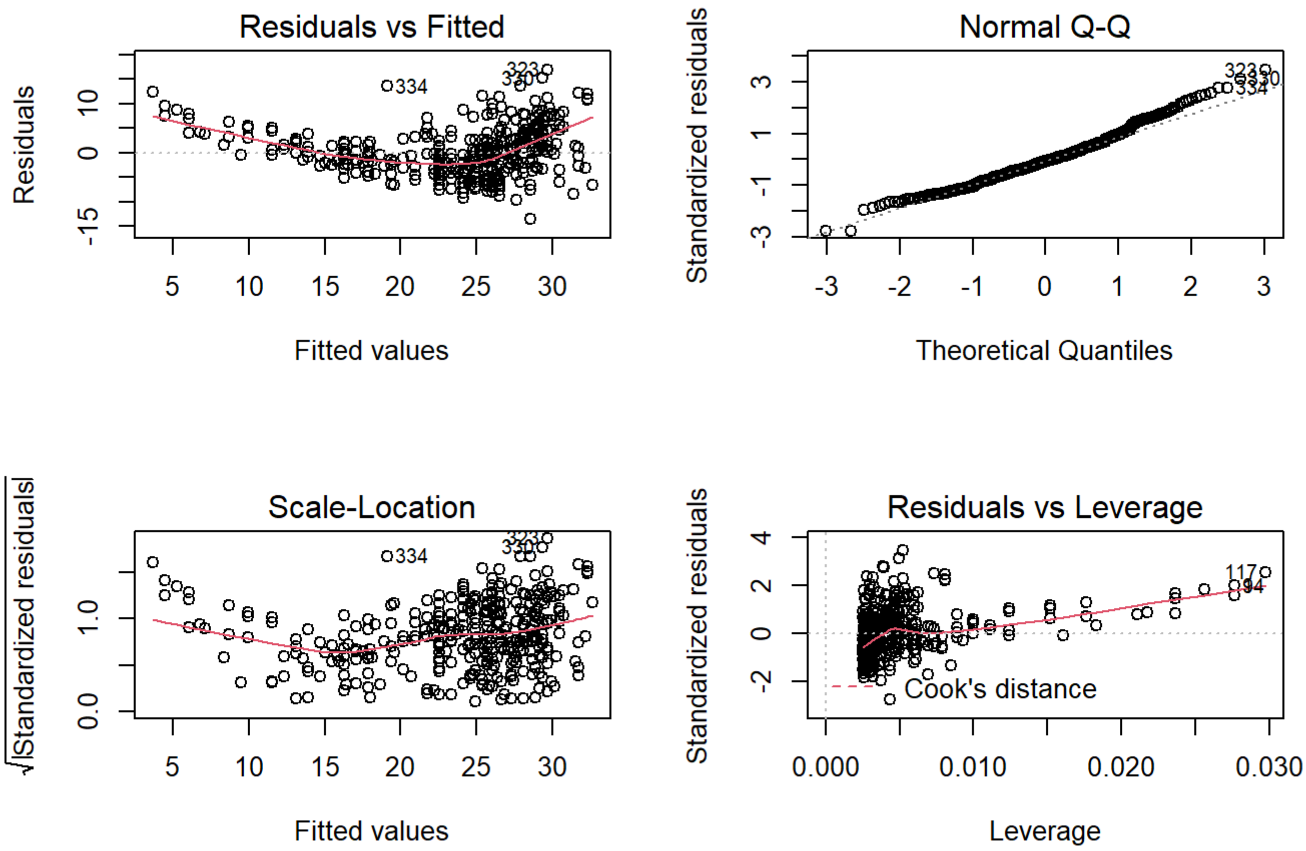
```
# The abline function is giving error

#abline(lm1)
```

(c) Use the `plot()` function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

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

plot(lm1)
```



Based on `Residual vs Fitted` plot, might exist a linear relationship between the `mpg` and `hp` , cause the points are well distributed along the red line.

The `Normal Q-Q` plot confirms the normal distribution of residuals, cause the points are well adjusted to the straigth line.

The `Residuals vs Leverage` analyses the influence of outliers. Besides the possible influence of observations 117 and 84, as showed in the plot, the distribution is well-behaved, as measured by the Cook Distance.