Chapter 3 - Linear Regression

Thalles Quinaglia Liduares

04/03/2022

Applied Exercise 3.9

Upload packages

```r
library(dplyr)
library(lmreg)
library(readxl)
library(corrplot)

source("http://www.sthda.com/upload/rquery_cormat.r")
```

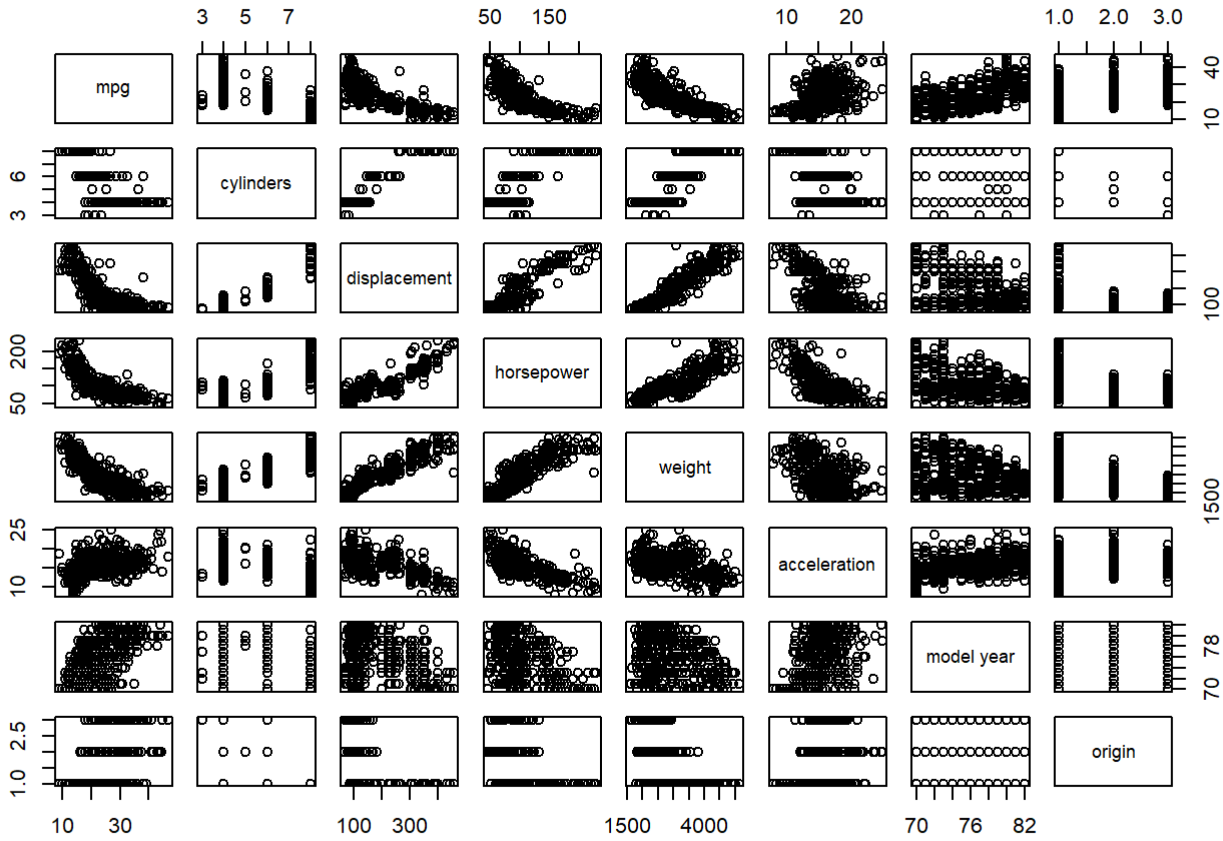
Upload Database

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

data<-readxl::read_excel("C:\\Program Files\\R\\Machine Learning\\auto-mpg.xlsx")
```

9. This question involves the use of multiple linear regression on the Auto data set.

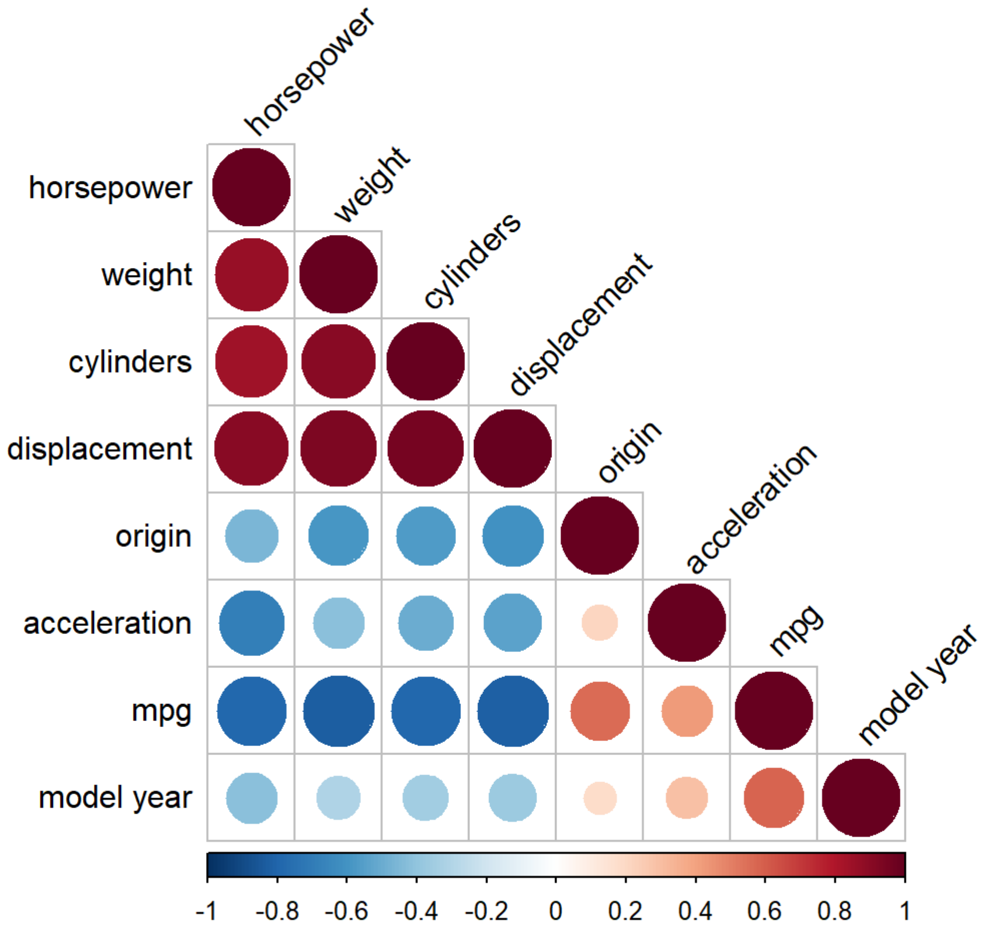
(a) Produce a scatterplot matrix which includes all of the variables in the data set.

```r
var_num <- select_if(data, is.numeric)

pairs(var_num)
```



(b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor() which is qualitative

```
rquery.cormat(var_num)
```


horsepower
weight
cylinders
displacement
origin
acceleration
mpg
model year
horsepower
weight
cylinders
displacement
origin
acceleration
mpg
model year
-1
-0.8
-0.6
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1

```
## $r
##              horsepower weight cylinders displacement origin acceleration  mpg
## horsepower            1
## weight             0.86      1
## cylinders          0.84    0.9         1
## displacement        0.9   0.93      0.95            1
## origin            -0.46  -0.59     -0.57        -0.61      1
## acceleration      -0.69  -0.42      -0.5        -0.54   0.21            1
## mpg               -0.78  -0.83     -0.78        -0.81   0.57         0.42    1
## model year        -0.42  -0.31     -0.35        -0.37   0.18         0.29 0.58
##              model year
## horsepower
## weight
## cylinders
## displacement
## origin
## acceleration
## mpg
## model year            1
##
## $p
##              horsepower   weight cylinders displacement  origin acceleration
## horsepower            0
## weight         1.4e-118        0
## cylinders      4.6e-107 1.1e-141         0
## displacement   1.5e-140 1.2e-177  1.7e-203            0
## origin          1.9e-21  2.6e-37   1.4e-34      7.9e-42       0
## acceleration    1.6e-56  3.2e-18   3.4e-27      5.4e-32 3.5e-05            0
## mpg               7e-81   3e-103   4.5e-81      1.7e-91   1e-34      1.8e-18
## model year      7.2e-18  4.2e-10     8e-13      2.3e-14 0.00029      4.8e-09
##                  mpg model year
## horsepower
## weight
## cylinders
## displacement
## origin
## acceleration
## mpg                0
## model year   4.8e-37          0
##
## $sym
##              horsepower weight cylinders displacement origin acceleration mpg
## horsepower   1
## weight       +          1
## cylinders    +          +      1
## displacement +          *      *         1
## origin       .          .      .         ,            1
## acceleration ,          .      .         .                   1
## mpg          ,          +      ,         +            .      .            1
## model year   .          .      .         .                                .
##              model year
## horsepower
## weight
## cylinders
## displacement
## origin
## acceleration
## mpg
## model year   1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

(c) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output.

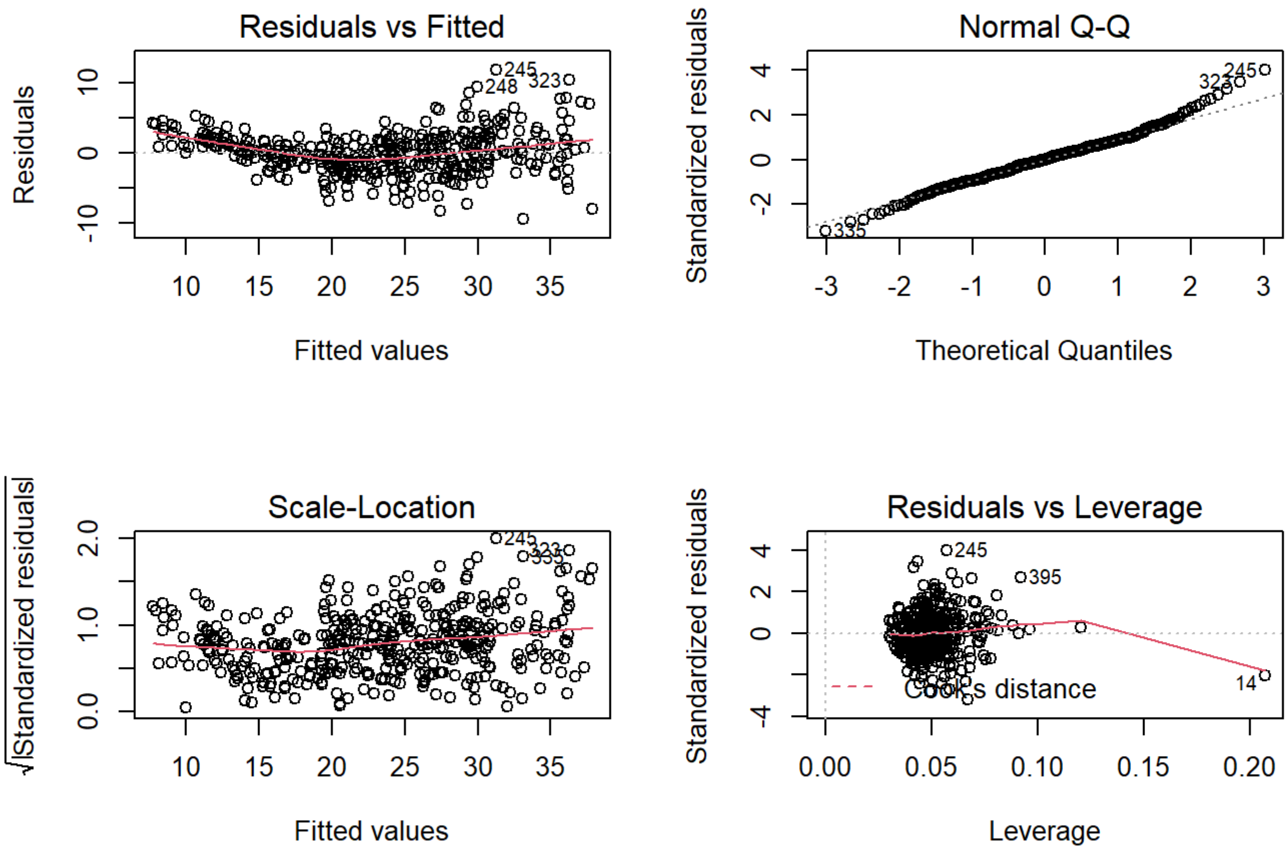
```
options(scipen=999)

lm1<-lm(mpg~cylinders+displacement+horsepower+weight+acceleration+factor(origin)+factor(`mode
l year`), data)

summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
##     acceleration + factor(origin) + factor(`model year`), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4288 -1.9194 -0.0287  1.7899 11.8399
##
## Coefficients:
##                          Estimate Std. Error t value            Pr(>|t|)
## (Intercept)            37.0199278  2.1417790  17.285 < 0.0000000000000002 ***
## cylinders              -0.1924289  0.3040482  -0.633             0.527195
## displacement            0.0172507  0.0072044   2.394             0.017139 *
## horsepower             -0.0240199  0.0136328  -1.762             0.078904 .
## weight                 -0.0061203  0.0006494  -9.424 < 0.0000000000000002 ***
## acceleration            0.0543880  0.0919344   0.592             0.554480
## factor(origin)2         2.5075851  0.5316558   4.717    0.000003398580565 ***
## factor(origin)3         2.5002584  0.5225230   4.785    0.000002470460191 ***
## factor(`model year`)71  1.0461869  0.8730131   1.198             0.231538
## factor(`model year`)72  0.0330325  0.8531036   0.039             0.969134
## factor(`model year`)73 -0.5322929  0.7718143  -0.690             0.490835
## factor(`model year`)74  1.6545531  0.9129699   1.812             0.070750 .
## factor(`model year`)75  0.9415172  0.8953739   1.052             0.293695
## factor(`model year`)76  1.7486166  0.8573480   2.040             0.042100 *
## factor(`model year`)77  3.2399161  0.8759807   3.699             0.000249 ***
## factor(`model year`)78  3.0821303  0.8333179   3.699             0.000249 ***
## factor(`model year`)79  5.3812526  0.8791655   6.121    0.000000002363133 ***
## factor(`model year`)80  9.5116004  0.9339482  10.184 < 0.0000000000000002 ***
## factor(`model year`)81  6.9070845  0.9223997   7.488    0.000000000000512 ***
## factor(`model year`)82  8.6173419  0.9031369   9.542 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.05 on 372 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.8547, Adjusted R-squared:  0.8473
## F-statistic: 115.2 on 19 and 372 DF,  p-value: < 0.00000000000000022
```

Among the characteristics of the car, only the variable `weight` has significance to the 1% level. The variable `displacement` shows statistical significance to the 5% level. The Adjusted R-Squared is 0.847. So, 84.7% of the variability in `mpg` is explaneid by the predictors. The cathegorical variable `year` suggests that cars perform more `mpg` , due mostly to the technological innovations.

(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

```
par(mfrow=c(2,2))
plot(lm1)
```



The `Normal -QQ` plot shows that residuals follow a normal distribution, because the points fits well the dashed straight line. The `Residuals vs Fitted` plot shows the horizontal red straight line that's a good indication that model follows a linear pattern.