Introduction to Multiple Regression

Bachelor in Computer Science and Engineering

2020/21

1. Introduction

The file `Cardata.xlsx` contains data about a sample of vehicles. Among other variables we have the variable `price` of these vehicles. We want to know which variables affect the most the price of the vehicles. To this aim we construct a multiple regression model that explains the price of the cars. The quantitative variables of this file that can be interesting to explain the cars' prices are:

- `mpg`: miles per gallon of fuel.
- `cylinders`: number of cylinders of the motor.
- `weight`: weight of the vehicle (pounds).
- `displace`: cylinder capacity (cubic inches).
- `horsepower`: engine power.
- `accel`: time the vehicle takes to reach a speed of 60 mph

```
library(readxl)
CarData <- read_excel("CarData.xlsx")
head(CarData, 5)
```

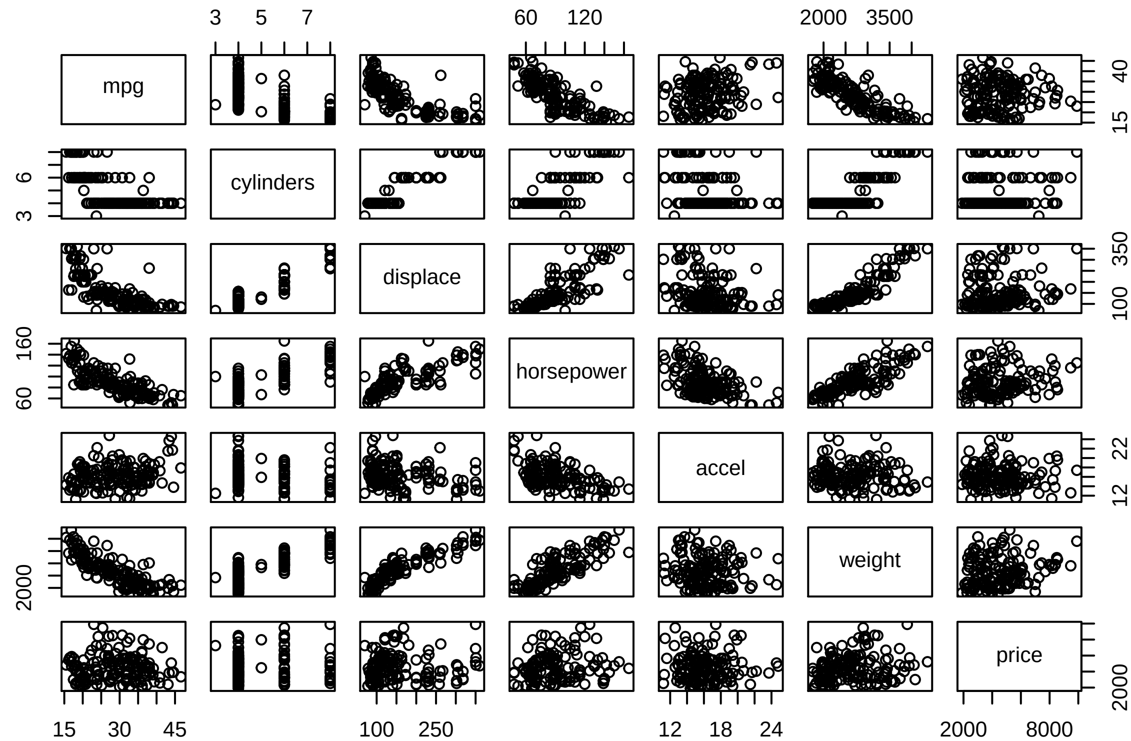
```
## # A tibble: 5 x 7
##     mpg cylinders displace horsepower accel weight price
##   <dbl>     <dbl>    <dbl>      <dbl> <dbl>  <dbl> <dbl>
## 1  43.1         4       90         48  21.5   1985  2400
## 2  36.1         4       98         66  14.4   1800  1900
## 3  32.8         4       78         52  19.4   1985  2200
## 4  39.4         4       85         70  18.6   2070  2725
## 5  36.1         4       91         60  16.4   1800  2250
```

2. Graphs XY

The multiple regression measures the relations that exists between a given variable, X_i, and another (dependent) variable, Y, after eliminating the influence of other additional variables. That means the information used by a multiple regression is not the same as the one we could plot by an XY graph. However it is useful to do a first graphical analysis of all data by means of XY graphs of each explicative variable X_i with Y. These graphs are useful to get a first impression about which variables have relation with the variable Y, if this relation is strong or weak, if it is linear or not, etc.

A simple way to visualize these graphs having many variables is

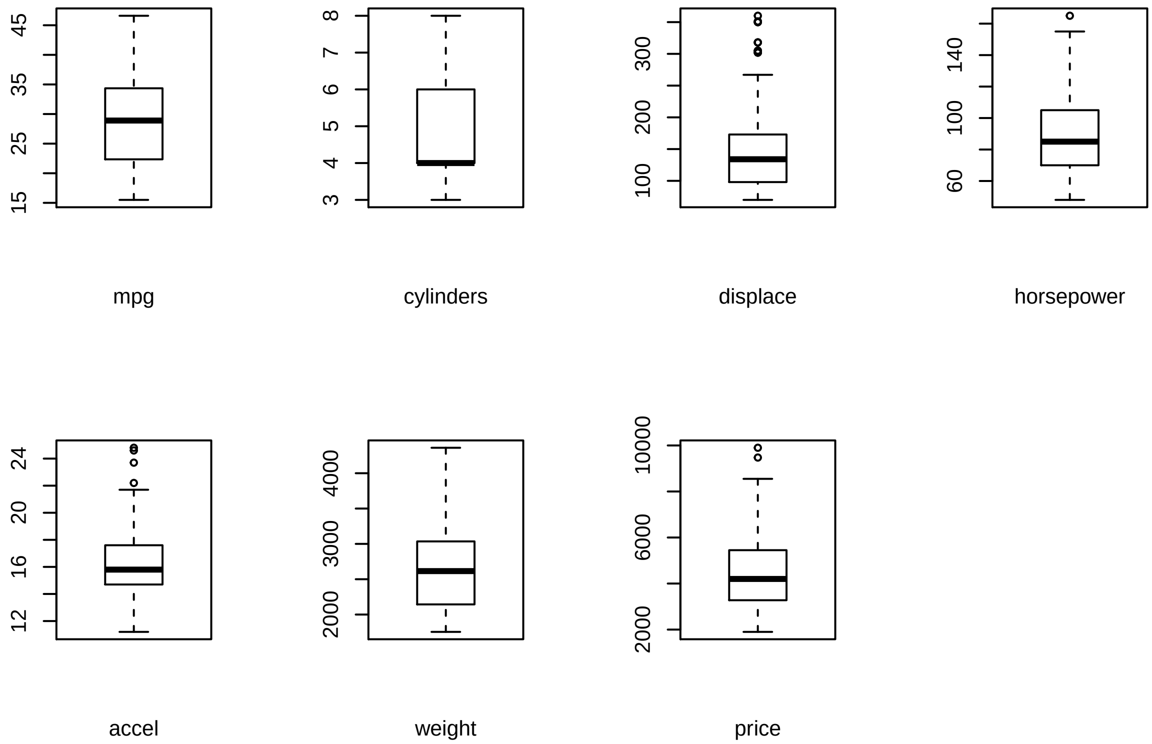
```
plot(CarData)
```



Of this matrix of XY's graphs, we are interested in the last line whose graphs show on the Y-axis the variable "price" and on the X-axes show the other variables. First important information we could get by looking at these graphs is the presence of atypical data. These outlier are confirmed at the boxplots below.

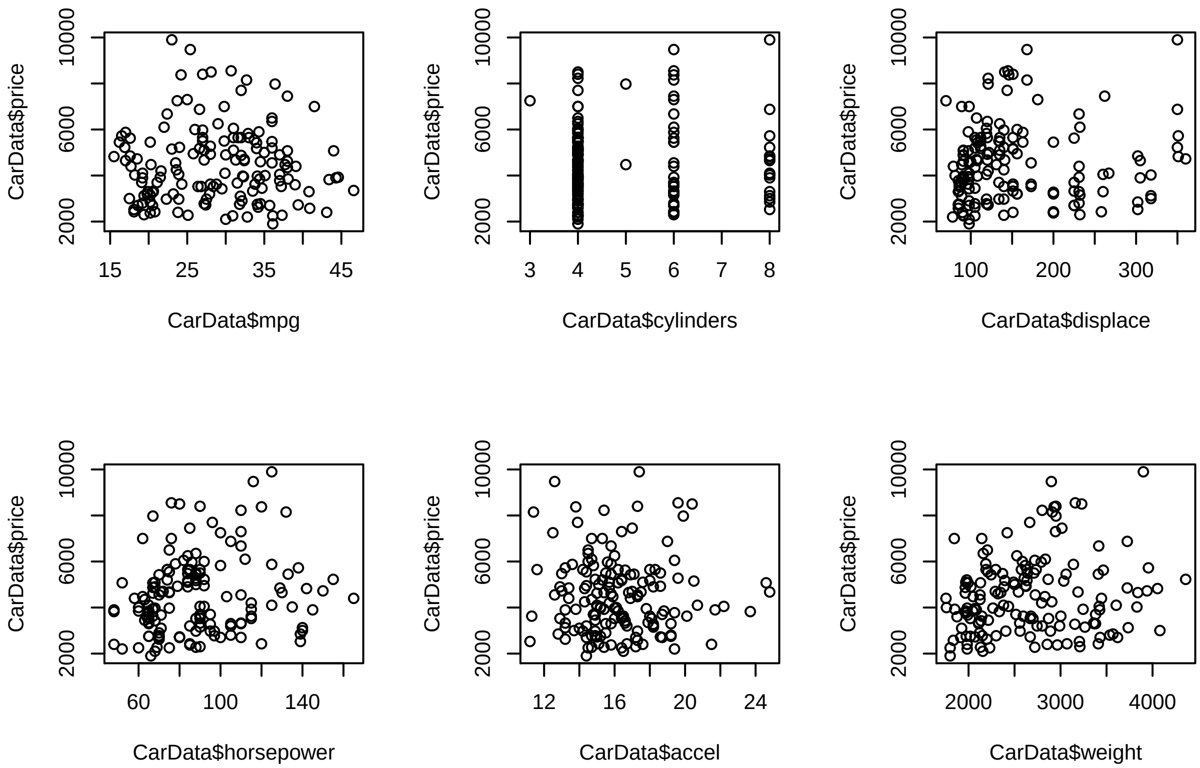
Also, it is interesting to see the univariate boxplots of each variable. We could use `boxplot (CarData)`, but note that this graph uses the same scale for all variables and does not return a good representation of all variables.

```
par(mfrow=c(2,4))
boxplot(CarData$mpg, xlab = "mpg")
boxplot(CarData$cylinders, xlab = "cylinders")
boxplot(CarData$displace, xlab = "displace")
boxplot(CarData$horsepower, xlab = "horsepower")
boxplot(CarData$accel, xlab = "accel")
boxplot(CarData$weight, xlab = "weight")
boxplot(CarData$price, xlab = "price")
```



To better view these relations, in the following, we plot the XY graph for each variable separately. The six XY graphs are:

```
par(mfrow=c(2,3))
plot(CarData$mpg, CarData$price)
plot(CarData$cylinders, CarData$price)
plot(CarData$displace, CarData$price)
plot(CarData$horsepower, CarData$price)
plot(CarData$accel, CarData$price)
plot(CarData$weight, CarData$price)
```



The most important aspects we could note are:

- The presence of anomalous points. They concern two luxury cars whose profiles are quite different from the ones of the other cars and that could affect negatively the results. The best thing to do for the analysis is to eliminate them.

```
CarData <- CarData[CarData$price<10000,]
```

- The influence of the all the variables is much dispersed. No variable alone has a very strong relation. We could expect that at the end the R^2 coefficient will be not very large.

3. Multiple Regression. Initial Model.

Since all the variable are related the results we would get with a simple regression and a multiple regression are very different. Therefore the best thing to do is to start with a model that contains all the variables (after having eliminated the atypical data). Next we will eliminate the variables that will result not significant. This elimination process has to be done one variable by one because the parameters of the model will depend on the full set of variables that are included in the analysis. Each time we eliminate one variable we have to recalculate all parameters and this can make that a variable that before seemed not significant can become significant.

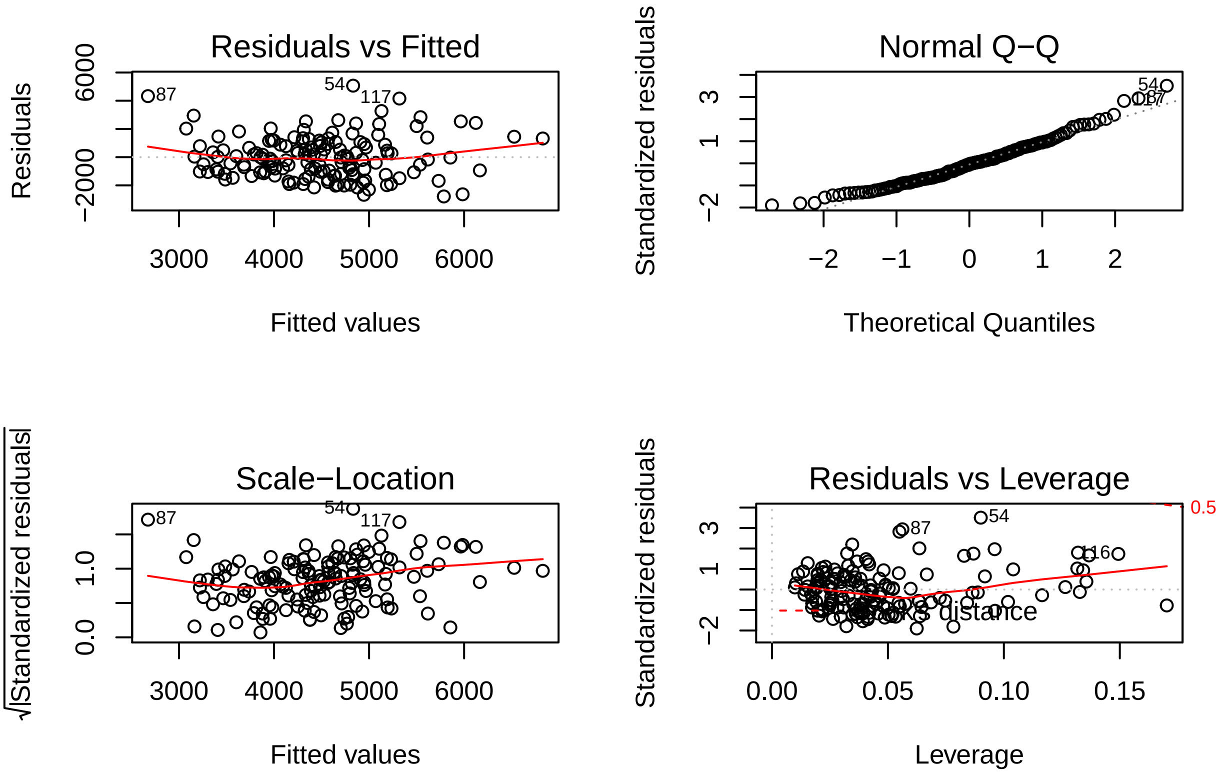
At the same time we have to look at the residual graphs in order to detect any no linearity that suggests to use a transformation. The multiple regression that uses all variables is shown in the following.

```
model <- lm(price ~ mpg + cylinders + displace + horsepower + accel + weight
             , data = CarData)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ mpg + cylinders + displace + horsepower +
##     accel + weight, data = CarData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2786.6 -1073.8   -68.6   928.0  5067.6
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5915.5438  2328.3168  -2.541   0.0121 *
## mpg           154.3406    32.9454   4.685 6.53e-06 ***
## cylinders      16.1369   254.8438   0.063   0.9496
## displace      -17.5516     6.8871  -2.548   0.0119 *
## horsepower     13.1809    14.5308   0.907   0.3659
## accel        -113.2531    82.2123  -1.378   0.1705
## weight          3.4705     0.8243   4.210 4.52e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1515 on 141 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.1955, Adjusted R-squared:  0.1613
## F-statistic: 5.712 on 6 and 141 DF,  p-value: 2.419e-05
```

The first thing to look at is the residuals to check if there are any patterns that could invalidate the regression. For this analysis, this does not appear to be the case. The regression results are valid and we can notice that there are variables that are not significant and that should be eliminated.

```r
par(mfrow=c(2,2))
plot(model)
```

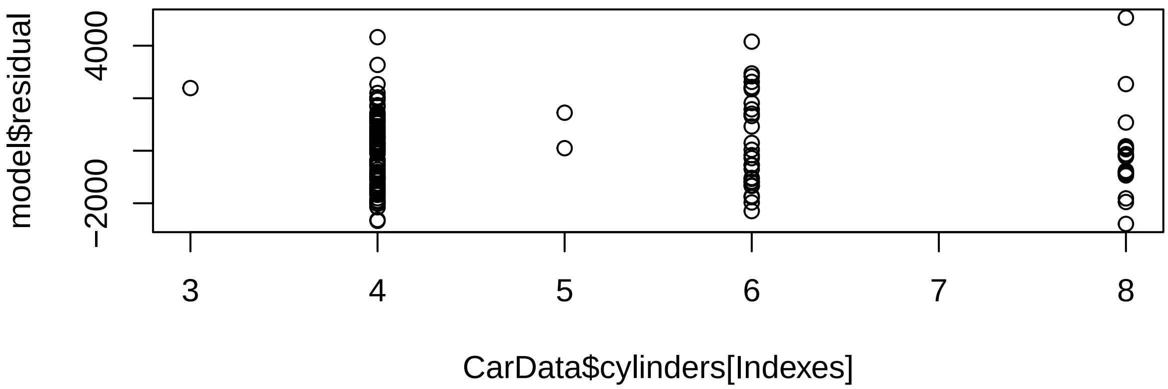


4. Elimination of no significant variables: Final Model.

The least significant variable is `cylinders`, i.e. the number of cylinders. Before eliminating it we notice that the reason of its low significance is due to the non linearity. To analyze the linearity of the relation we plot the residual Plot for this variable.[1]

```
Indexes = as.numeric(names(model$residuals))
plot(CarData$cylinders[Indexes], model$residual)
```

[1] The first line `Indexes = as.numeric(names(model$residuals))` finds the indexes of the observations used to estimate the regression model. The observations with `NA` in some variables are not used.



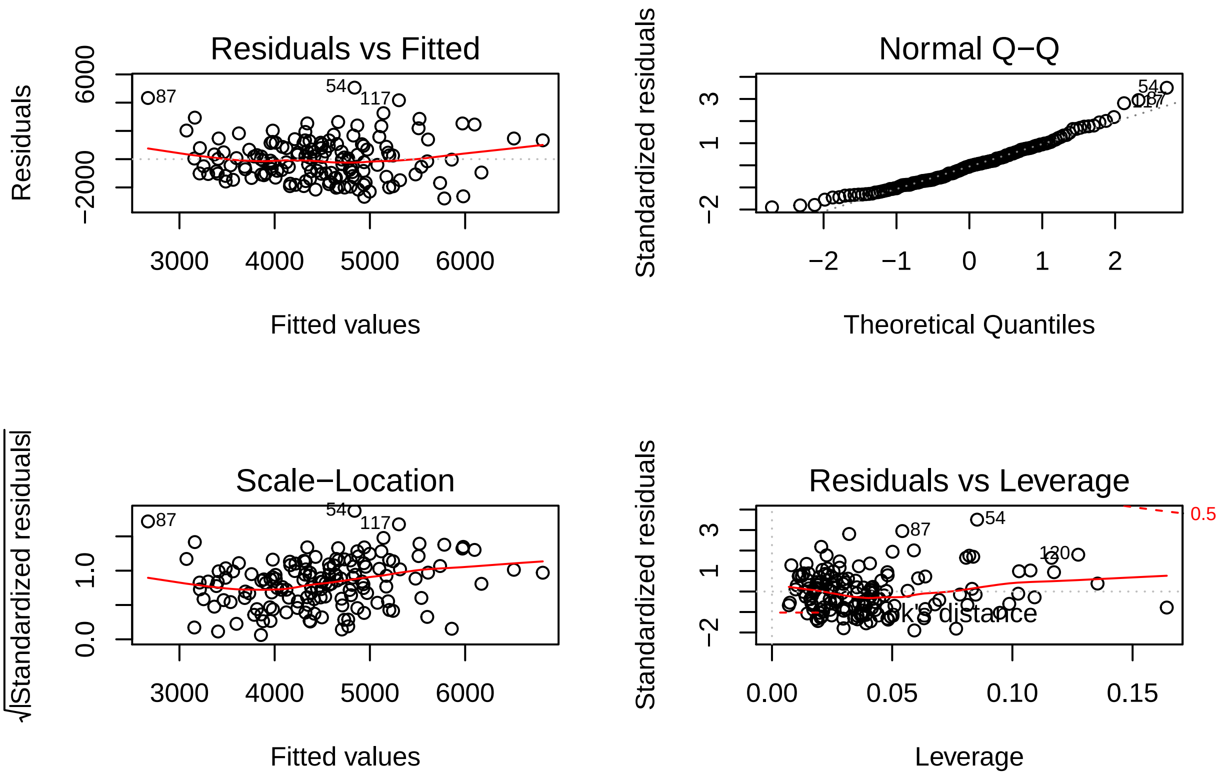
This graph shows that this variable has no marginal contribution to the model. Therefore the new model is:

```
model <- lm(price ~ mpg + displace + horsepower + accel + weight, data = CarData)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ mpg + displace + horsepower + accel + weight,
##     data = CarData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2781.2 -1080.8   -69.2   934.7  5061.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5888.9438  2282.0612  -2.581 0.010879 *
## mpg           154.4321    32.7981   4.709 5.87e-06 ***
## displace      -17.2371     4.7550  -3.625 0.000402 ***
## horsepower     13.3136    14.3286   0.929 0.354382
## accel        -112.6906    81.4438  -1.384 0.168634
## weight          3.4630     0.8127   4.261 3.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1509 on 142 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.1955, Adjusted R-squared:  0.1672
## F-statistic: 6.903 on 5 and 142 DF,  p-value: 8.5e-06
```

Now we remove the variable `horsepower` which in this new model seems less significant. The residual plots do not show any anomaly, so we can say that this variable does not make any informative contribution to the price that is not contained in the other explanatory variables.

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



The new model is now:

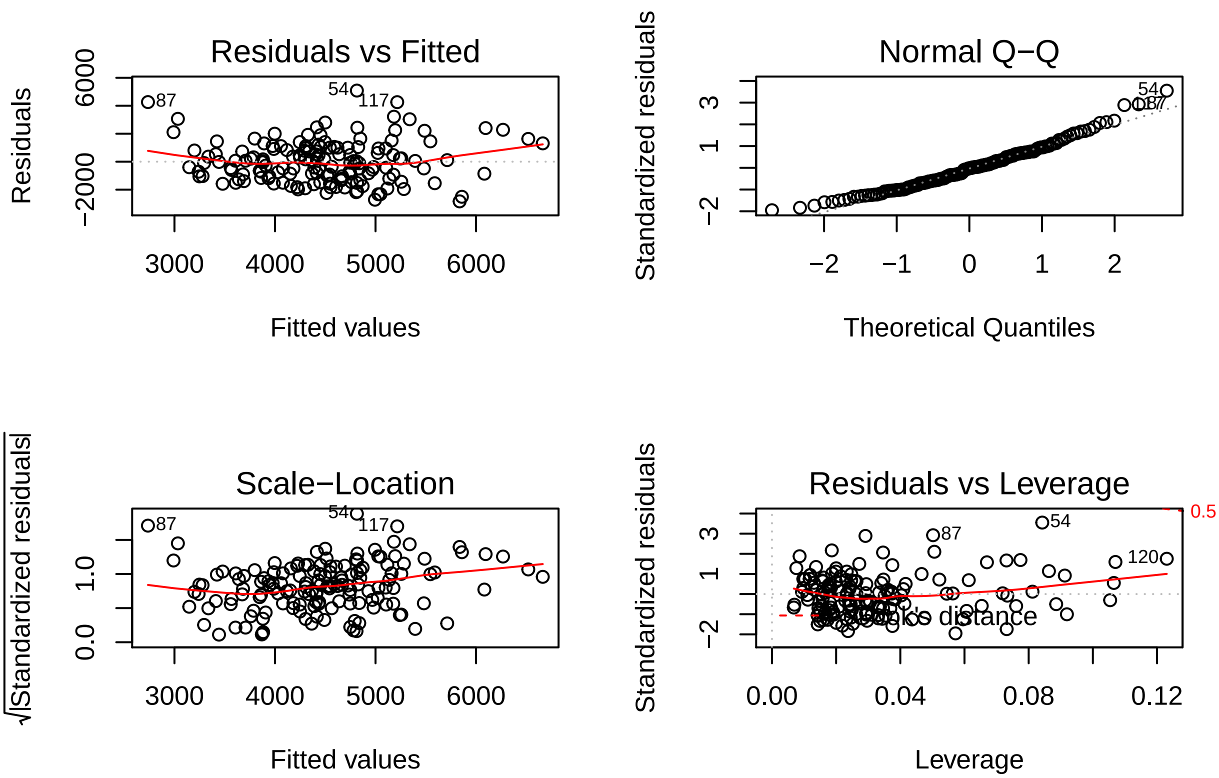
```
model <- lm(price ~ mpg + displace +accel + weight, data = CarData)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ mpg + displace + accel + weight, data = CarData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2836.6 -1044.9   -25.8   953.5  5085.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4754.045   1907.074  -2.493 0.013781 *
## mpg           147.013     32.027   4.590 9.42e-06 ***
## displace      -17.136      4.634  -3.698 0.000306 ***
## accel        -159.828     56.534  -2.827 0.005352 **
## weight          3.837      0.689   5.569 1.18e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1497 on 147 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.1892, Adjusted R-squared:  0.1672
## F-statistic: 8.577 on 4 and 147 DF,  p-value: 3.004e-06
```

```r
aov(model)
```

```
## Call:
##    aov(formula = model)
##
## Terms:
##                      mpg  displace     accel    weight Residuals
## Sum of Squares     45417   7026053    301103  69504169 329398241
## Deg. of Freedom        1         1         1         1       147
##
## Residual standard error: 1496.932
## Estimated effects may be unbalanced
## 1 observation deleted due to missingness
```

```r
par(mfrow=c(2,2))
plot(model)
```



This model contains all the informative variables (we can notice that `accel` that before was not significant now looks like significant) and only explains the 18.9% of the variability of the variable price. The residual vs. predicted graph does not show any visible pattern.

The resulting multiple regression model that explains the price of the cars is:

`price = -4754.045 + 147.013 mpg - 17.136 displace - 159.828 accel + 3.837 weight`

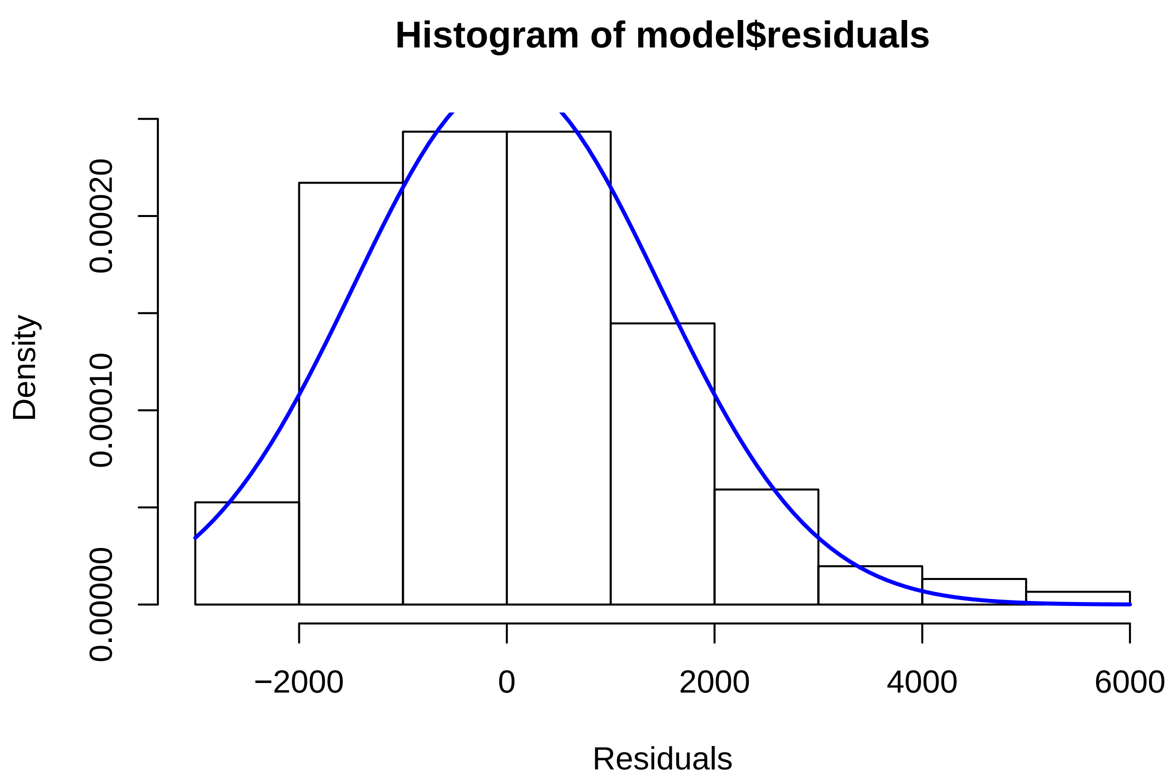
We should add to it only an error term that is a normal random variable with zero mean and estimated variance (residual variance) equal to `2240804` (`=329398241/147`).

From this model we can deduce that:

- Keeping the gasoline consumption, the weight, the displacement and the acceleration time constant, the number of cylinders and the power are not helpful to predict the car price.
- Keeping the weight, the displacement and the acceleration time constant, the cars that have less gasoline consumption (higher `mpg`) are more expensive.
- Keeping the gasoline consumption, the displacement and the acceleration time constant, the cars that weight more are more expensive.
- Keeping the gasoline consumption, the weight and the acceleration time constant, the cars that have higher displacement cost less. This result seems not very intuitive. Maybe it is due to the fact that keeping fixed the gasoline consumption the ones that have a smaller displacement are more efficient: more valves, turbocharged, electronic injectors, etc.

To complete the critical analysis of the model, we look at the normality of the residuals.

```
hist(model$residuals,
     probability = TRUE, # histogram has a total area = 1
     xlab = "Residuals")
curve(dnorm(x, mean(model$residuals), sd(model$residuals)),
      col="blue", lwd=2, add=TRUE, yaxt="n")
```



This histogram shows that the residuals have a light asymmetry but that in first approximation they can be considered normal. The `p-value` of the goodness-of-fit (chi-squared) test is 0.1841, and therefore we can conclude that the model is adequate.

```
library(nortest)
pearson.test(model$residuals)
```

```
##
##  Pearson chi-square normality test
##
## data:  model$residuals
## P = 16.158, p-value = 0.1841
```

5. Regression with binary variables

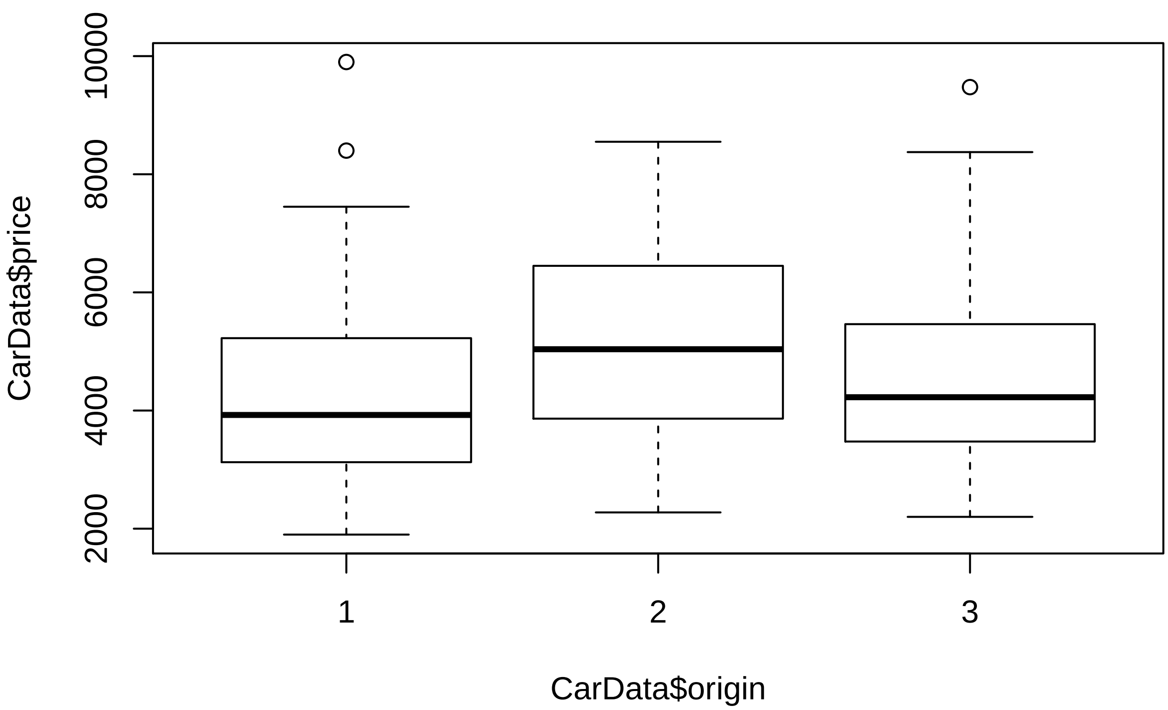
The file `Cardata2.xlsx` contains a variable `origin` that gives information about where the car has been produced. The meaning of its values is the following:

- America = 1,
- Europe = 2,
- Japan = 3.

We could use this variable to analyze if the mean price of the cars is influenced by the production region in a way that is not explicable by the information contained in the others variables, i.e. that it not captured by the above regression model. That is we are going to check if the price of car can be influenced by the fact that the car is American, European or Japanese.

First we make a descriptive analysis of the data by using boxplots:

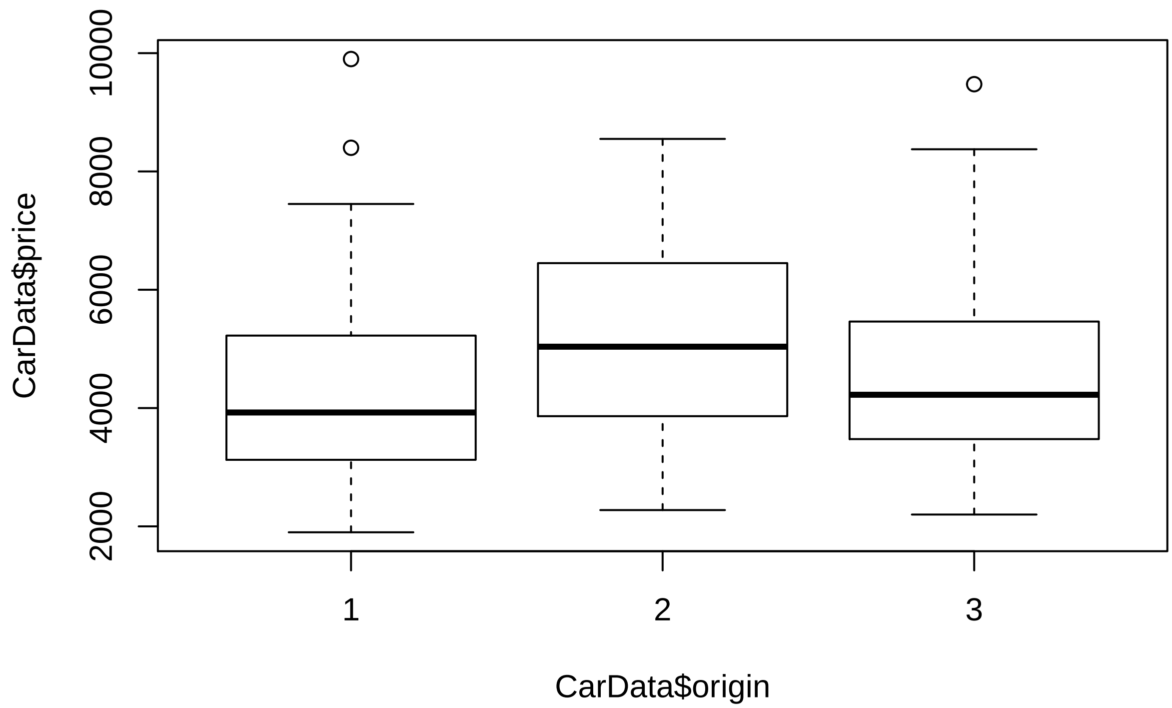
```
boxplot(CarData$price ~ CarData$origin)
```



Data show that the American cars are less expensive, then the Japanese ones and eventually the European cars. We want to know if these differences are significant and if they can be explained by the other regression variables.

We will use as reference point the mean price of the American cars, and so we will introduce two binary variables, one for the European cars (origin=2) and one for the Japanese cars (origin=3), and we will check if their coefficient are or are not significant.

```r
boxplot(CarData$price ~ CarData$origin)
```




```r
CarData$origin2 = (CarData$origin==2)
CarData$origin3 = (CarData$origin==3)
model <- lm(price ~ mpg + displace +accel + weight + origin2 + origin3, data = CarData)
summary(model)
```

```
##
## Call:
## lm(formula = price ~ mpg + displace + accel + weight + origin2 +
##     origin3, data = CarData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2935.0 -1035.3  -237.9   984.9  4950.6
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4596.3337  1928.8587  -2.383 0.018471 *
## mpg           127.4700    32.4612   3.927 0.000133 ***
## displace      -14.0419     4.7473  -2.958 0.003618 **
## accel        -154.3321    55.8231  -2.765 0.006439 **
## weight          3.6446     0.6955   5.240 5.56e-07 ***
```

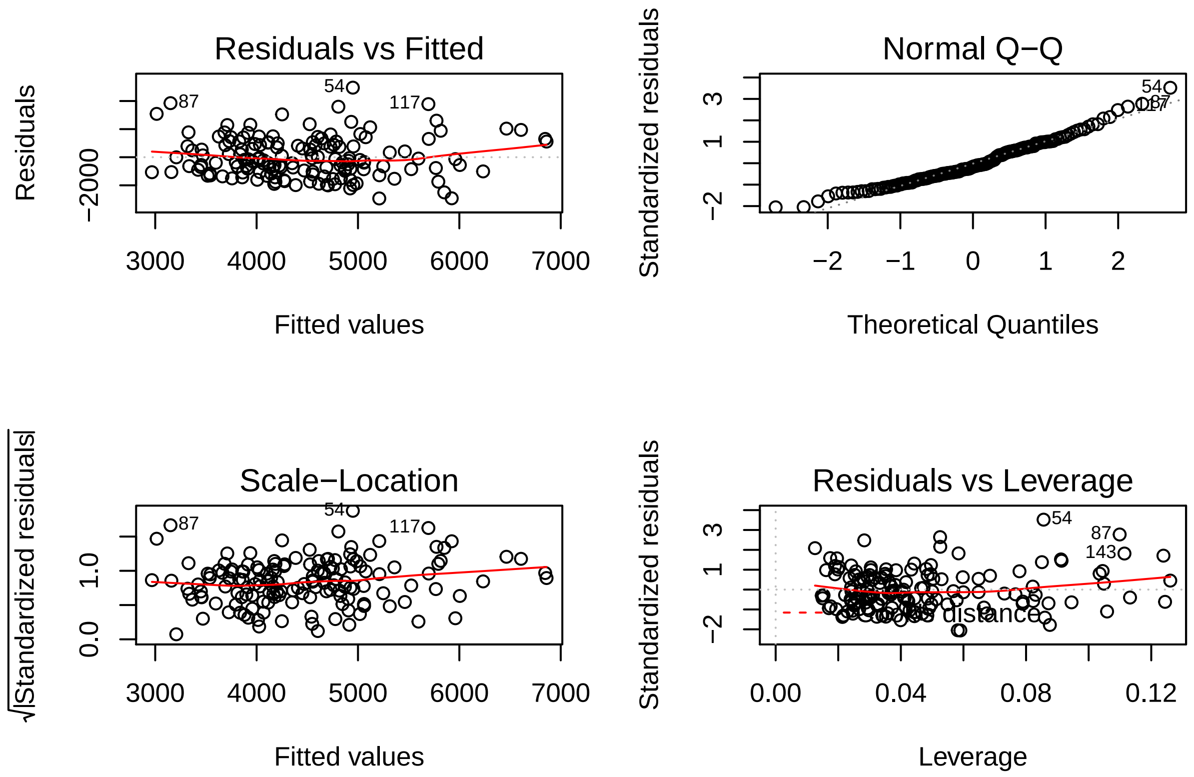
```
## origin2TRUE   835.1986   396.2150   2.108 0.036756 *
## origin3TRUE   787.5549   328.7206   2.396 0.017859 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1471 on 145 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2277, Adjusted R-squared:  0.1957
## F-statistic: 7.125 on 6 and 145 DF,  p-value: 1.14e-06
```

```r
aov(model)
```

```
## Call:
##    aov(formula = model)
##
## Terms:
##                       mpg  displace     accel    weight   origin2   origin3
## Sum of Squares      45417   7026053    301103  69504169   3209120  12420780
## Deg. of Freedom         1         1         1         1         1         1
##                 Residuals
## Sum of Squares  313768340
## Deg. of Freedom       145
##
## Residual standard error: 1471.027
## Estimated effects may be unbalanced
## 1 observation deleted due to missingness
```

We can notice that the binary variables are significant: the price of the cars depends on the production region and the difference cannot be explained by the different characteristics own by the vehicles. The American cars are the cheapest, in a set of car of given characteristics, then the Japanese cars (on average they cost 787.5 US dollars more than the American ones keeping equal the other factors) and eventually the European ones (on average they cost 835.199 US dollars more than the American ones keeping equal the other factors).

```r
par(mfrow=c(2,2))
plot(model)
```



The residual plots again do not show any pattern and therefore the regression model can be considered as valid.