Regression Models - Project

Lorena Zuniga

Executive Summary

For this project, Motor Trend (a magazine about the automobile industry) is interested in exploring the relationship between a set of variables and miles per gallon (MPG). They are interested in the following questions: 1- Is an automatic or manual transmission better for MPG ? 2- Quantify the MPG difference between automatic and manual transmissions The project use the dataset mtcars

Exploratory Analysis

Load the data:

```
data("mtcars")
str(mtcars)
```

In the am column: 0 is automatic transmission, 1 = manual transmission

We can observe that the MPG mean between each type of transmission is different:

```
 autom<- mtcars[mtcars$am==0,]
 manual <-mtcars[mtcars$am==1,]
 mean(autom$mpg)
```

```
## [1] 17.15
```

```
 mean(manual$mpg)
```

```
## [1] 24.39
```

Inference

It is necessary to determine if the MPG mean difference is significant, then we can perform a t-test. Alpha value =0.05

```
 t.test(autom$mpg,manual$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  autom$mpg and manual$mpg
## t = -3.767, df = 18.33, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.28  -3.21
## sample estimates:
## mean of x mean of y
##     17.15     24.39
```

The p-value is 0.001374, so in this case the difference between the MPG mean of the cars with manual transmission and the automatic cars is significant

Regression Analysis

We test a first simple linear regression model using am as predictor and mpg as outcome

```
summary (lm(mpg ~ am, data=mtcars))
```

The coefficient and intercept shows that manual transmission cars have 7.24 MPG and automatic transmission cars have 17.24 MPG. The R-squared = 0.338 , that shows the model only explains 33.8% of the variance.

Multivariable linear regression

In order to know which variables could be better to include in the multivariable linear regression model, we can get a correlation matrix, as follows:

```
cor(mtcars)[1,]
```

From the resultant matrix we can say that wt and hp could be tested as predictors

```
simpleModel <- lm(mpg ~ am,data=mtcars)
multiModel <- lm(mpg ~ am + hp + wt, data=mtcars)
```

Comparing both models with ANOVA:

```
anova(simpleModel, multiModel)
```

The p-value is 3.7e-09 and therefore the difference between both models is significant

```
summary(multiModel)
```

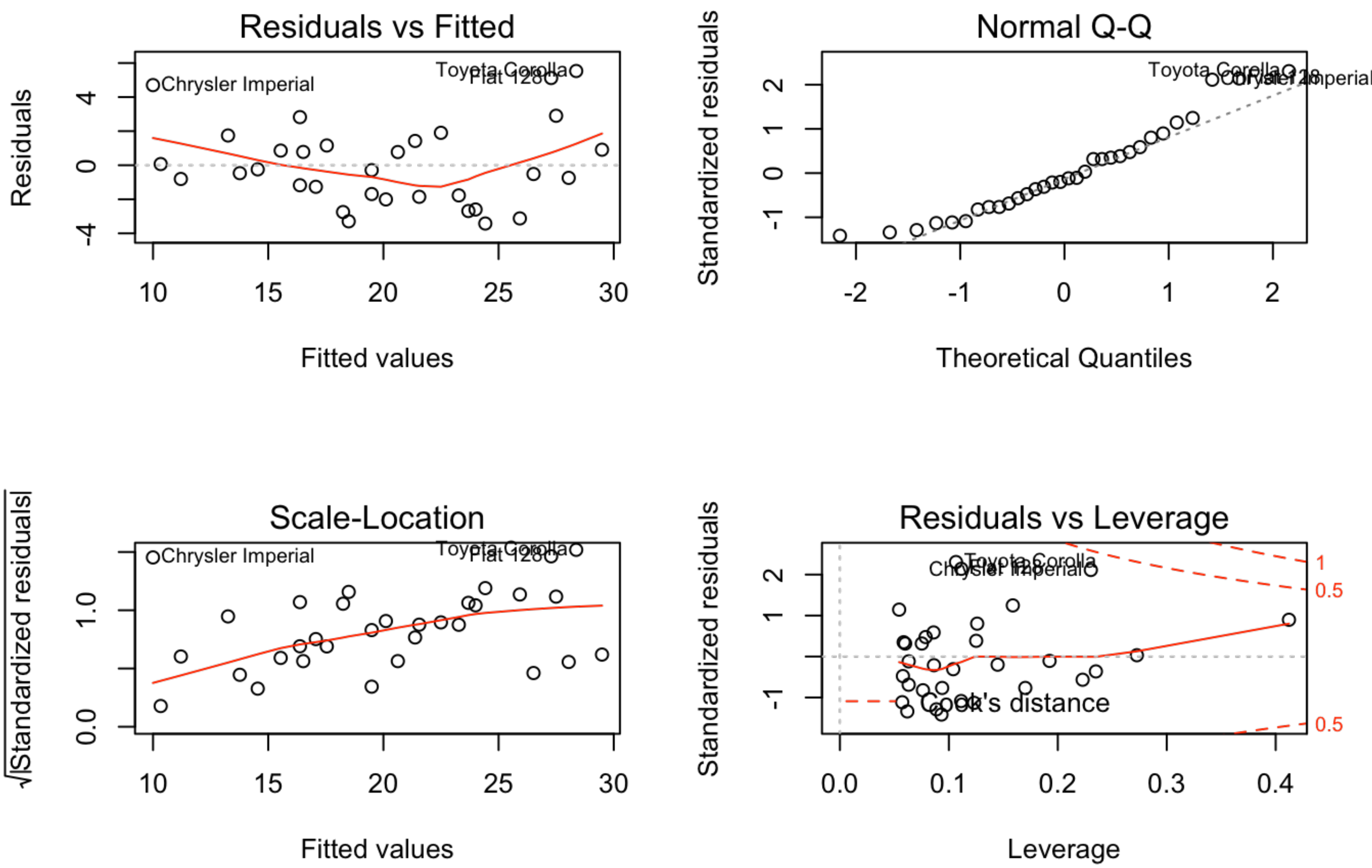
For the multivariable model the R-squared = 0.84, it is possible to say that the multivariable model explain the 84% of the variance.

Residuals

We can observe in the appendix that the residuals are homoskedastic

Apendix

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



The complete code for this project can be found here: (https://github.com/lorenazs/regressionmodels/blob/master/Project/project.Rmd (https://github.com/lorenazs/regressionmodels/blob/master/Project/project.Rmd))