Linear Regression Assignment

Diego Menin

Thursday, November 20, 2014

Disclaimer: This Documnent is part of the Course Project from the Linear regresison Course on the Coursera Data science Specialization. The information contained here is for general information purposes only. Any reliance you place on such information is therefore strictly at your own risk. I cannot accept any liability for the consequences of any actions taken on the basis of the information here provided.

Executive Summary

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

"Is an automatic or manual transmission better for MPG" "Quantify the MPG difference between automatic and manual transmissions"

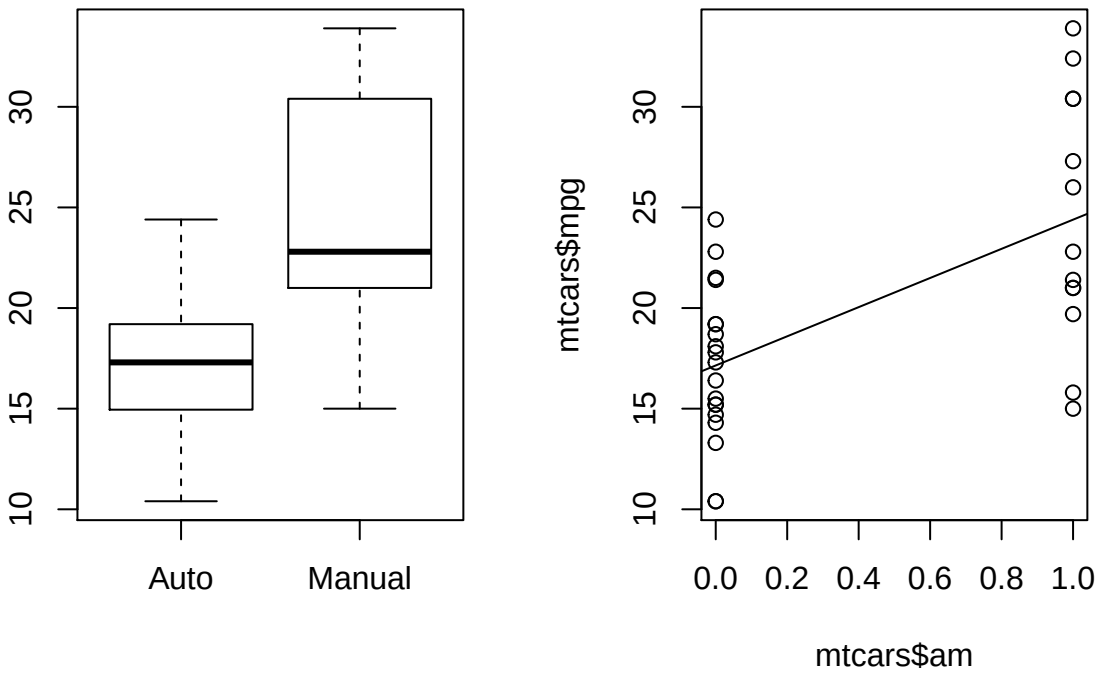
Analysis

In order to perform the Analyis, we'll load the mtcars data and perform a small data transformation by changing the "am" variable from numeric to factor (you can also run ?mtcars in R to see more details about the dataset)

```
cars <- mtcars
cars$am <- as.factor(cars$am)
levels(cars$am) <-c("Auto", "Manual")
```

A simple box plot shows that manual cars do more miles per gallon than Auto cars and a linear regresion predicting mpg using transmission shows that a change from Auto to Manual would result in a 7.2 increase in mpg

```
fit <- lm(mpg ~ am, data= cars)
par(mfrow = c(1,2))
plot(cars$am, cars$mpg)
#using the original mtcars dataset because on this dataset the am variable is
# a number so it automatic produces  a scaterplot
plot(mtcars$am, mtcars$mpg)
abline(fit)
```



Which is also the difference between the means

```
mean(cars[cars$am == "Manual",]$mpg) - mean(cars[cars$am == "Auto",]$mpg)
```

```
## [1] 7.245
```

We can also do a T-Test on the data

```
t.test(cars[cars$am == "Auto",]$mpg, cars[cars$am == "Manual",]$mpg)
```

```
##
##  Welch Two Sample t-test
##
## data:  cars[cars$am == "Auto", ]$mpg and cars[cars$am == "Manual", ]$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
```

Which results on a p-value of 0.001374 (bellow 5%) so quoting the output: “the alternative hypothesis is true: the difference in means is not equal to 0”

At this point we already know that the mean gallons per mile of "Manual"" cars is 7.244 MPGs higher than that of "Auto" cars and that the R^2 value is 0.3598, which means that this model only explains 35.98% of the variance. This is quite a poor fit so we'll try some Multi-Variable Analysis To find a better fitting model.

The question is: which other variables to include? Initially it would be intuitive (at least to me) to include hp (Gross horsepower), wt (weight) and qsec (1/4 mile time) because in theory the power, weight and speed would directly affect a car's fuel consuption (which is another way to look at miles per gallon; saying that your car does less miles per gallon of fuel it is the same as saying it consumes more fuel).

In order to check that, we are going to fit a model will all variables and them use the "step" function to check which variables best fit the model

```
fitall <- lm(mpg ~ ., data=cars)
goodfit<- step(fitall, trace=0)
summary(goodfit)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.481 -1.556 -0.726  1.411  4.661
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.618      6.960    1.38  0.17792
## wt            -3.917      0.711   -5.51    7e-06 ***
## qsec           1.226      0.289    4.25  0.00022 ***
## amManual       2.936      1.411    2.08  0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.46 on 28 degrees of freedom
## Multiple R-squared:  0.85,   Adjusted R-squared:  0.834
## F-statistic: 52.7 on 3 and 28 DF,  p-value: 1.21e-11
```

The function outputs 3 variable wt, qsec and "amManual", which confirms our "intuition". Actually if we check the p-values of our fitall model we can see that wt, am and qsec are the lowest p-values (and hp is the forth lowest), so not a bad prediction

```
summary(fitall)$coef[,4]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
##     0.51812     0.91609     0.46349     0.33496     0.63528     0.06325
##        qsec          vs    amManual        gear        carb
##     0.27394     0.88142     0.23399     0.66521     0.81218
```

Looking at the coeficinets we can see for example that the wt coeficient is -3.9165, which means that each point increased in mpg will decrease the weight in -3.9165. A better way to interpret that is: **Increasing the weight of a car by 3916.5 lb will cause a decrease of 1 mille per gallon on fuel consumption**, which makes sense because it is intuitive that a heavier car does less miles per 1 gallon of fuel. (Remember that the wt is sored in lb/1000)

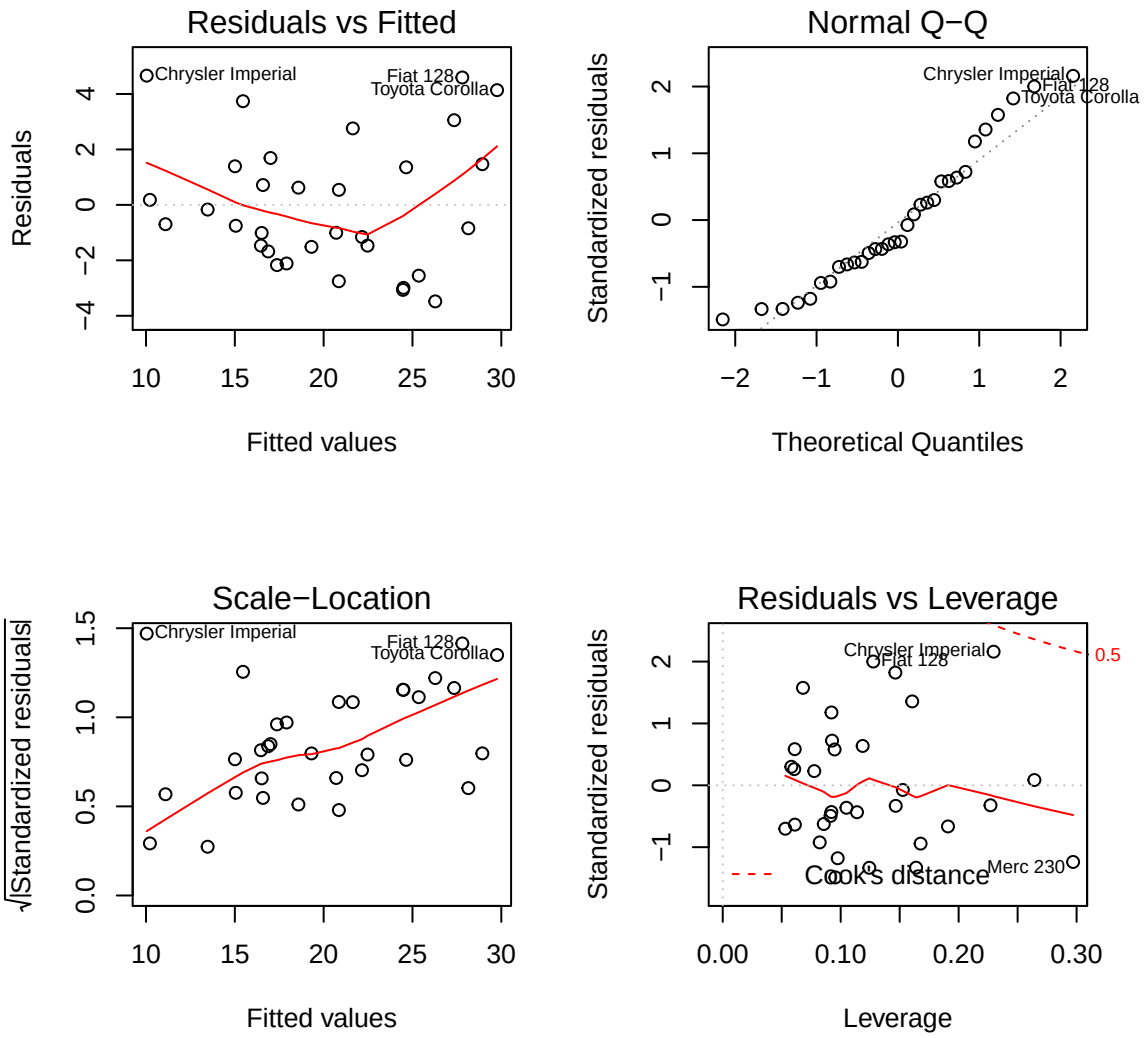
We can also see that R^2 is 0.8497 which means that this model explains 85% of the variance.

```
summary(fitall)$coef[,4]
```

```
## (Intercept)         cyl        disp          hp        drat          wt
##     0.51812     0.91609     0.46349     0.33496     0.63528     0.06325
##        qsec          vs    amManual        gear        carb
##     0.27394     0.88142     0.23399     0.66521     0.81218
```

Here we can graphically visualize the fited model

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



Conclusion

Cars with Manual Transmission are better than cars with automatic transmission by 2.9 mpg. However, transmission it is not the only factor that affects MPG; a car's weight and its acceleration (1/4 mile time)

also seem to affect the result.