Regression Models Course Project

Francisco Martín

October 17, 2018

Statement

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"

Executive Summary

We have to examine mtcars dataset (default in R) and explore how MPG is correlated with other car parameters such as the number of cylinders, horsepower, and automatic/manual transmission. In the end, we should be capable of giving an answer of how the type of transmission affects on miles per gallon using only linear regression analysis and give a set of figures which support our conclusions.

Exploratory Analysis

First of all, we should load mtcars dataset into our workspace

```
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

As we see, type of transmission is displayed in column "am", where 1 means automatic and 0 means manual transmission. I don't like how it looks, so I am going to change it to a factor column and see how data looks for each of one:

```
mtcars_cool <- mtcars
mtcars_cool$am <- factor(mtcars$am, labels = c("manual","automatic"))
summary(mtcars_cool[mtcars_cool$am == "automatic",]$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   15.00   21.00   22.80   24.39   30.40   33.90
```

```
summary(mtcars_cool[mtcars_cool$am == "manual",]$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   10.40   14.95   17.30   17.15   19.20   24.40
```

We can see mean value is about 7.24 more mpg in automatic than in manual (plot 1). We can adventure and do a big linear model looking for relations, but maybe it is better to look at correlations between different variables and mpg before doing any model:

```
##   mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb 
##  1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
```

There are four variables which have a stronger correlation. Those are “wt”, “cyl”, “disp” and “hp”. Let’s look how this variables and type of transmission affect on miles per galon doing different models:

```
lm_1 <- lm(formula = mpg ~ am, data = mtcars)
lm_2 <- lm(formula = mpg ~ am + wt, data = mtcars)
lm_3 <- lm(formula = mpg ~ am + wt + cyl, data = mtcars)
lm_4 <- lm(formula = mpg ~ am + wt + cyl + disp, data = mtcars)
lm_5 <- lm(formula = mpg ~ am + wt + cyl + disp + hp, data = mtcars)
lm_all <- lm(formula = mpg ~ ., data = mtcars)
```

We have trained six models. Let’s compare all of them using anova:

```
anova(lm_1, lm_2, lm_3, lm_4, lm_5, lm_all)
```

```
## Analysis of Variance Table
## 
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + cyl
## Model 4: mpg ~ am + wt + cyl + disp
## Model 5: mpg ~ am + wt + cyl + disp + hp
## Model 6: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1     30 720.90                                   
## 2     29 278.32  1    442.58 63.0133 9.325e-08 ***
## 3     28 191.05  1     87.27 12.4257   0.00201 ** 
## 4     27 188.43  1      2.62  0.3732   0.54782    
## 5     26 163.12  1     25.31  3.6030   0.07151 .  
## 6     21 147.49  5     15.63  0.4449   0.81206    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at summary we can see how well this model looking at how much variability it explains:

```
summary(lm_3)$coef[2,1]
```

```
## [1] 0.1764932
```

```
summary(lm_4)$coef[2,1]
```

```
## [1] 0.1290656
```

```
summary(lm_5)$coef[2,1]
```

```
## [1] 1.556492
```

Looks like hp is a key factor, because it changes the sign of the estimate am. Looking at this, and seing that wt and cyl has the same impact on model, let's train the final model:

```
lm_final <- lm(formula = mpg ~ am + wt + hp, data = mtcars)
summary(lm_final)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Looking at R-Squared term, it explains 83.99% of the variability. We can use this model to predict and explain general behaviour of mpg as we did in appendix (plots 2 and 3) doublechecking residuals and seing their are normally distributed.

Once we have doublecheck this model is valid, we can conclude automatic cars will run aproximatelly 2.08 more miles per galeon.

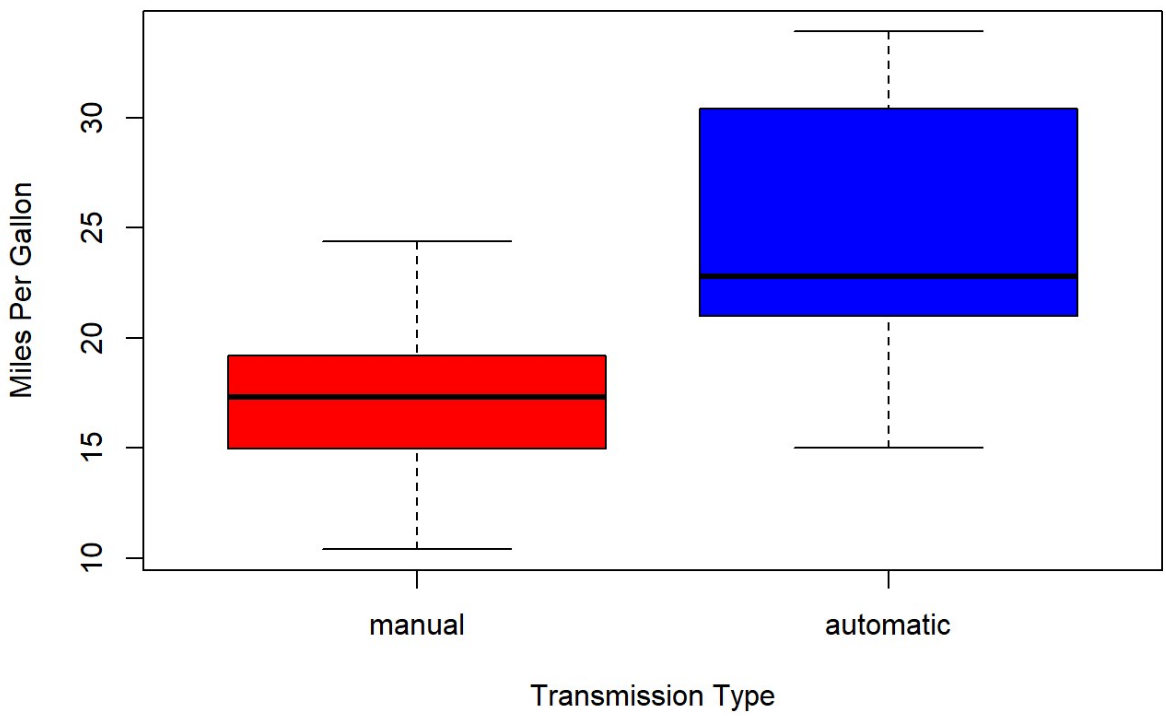
Conclusion

Looking at results obtained, we can conclude automatic car will travel 2.08 more miles per galeon. Based on this result, we can conclude that automatic cars are more efficient than manual cars.

Apendix (figures)

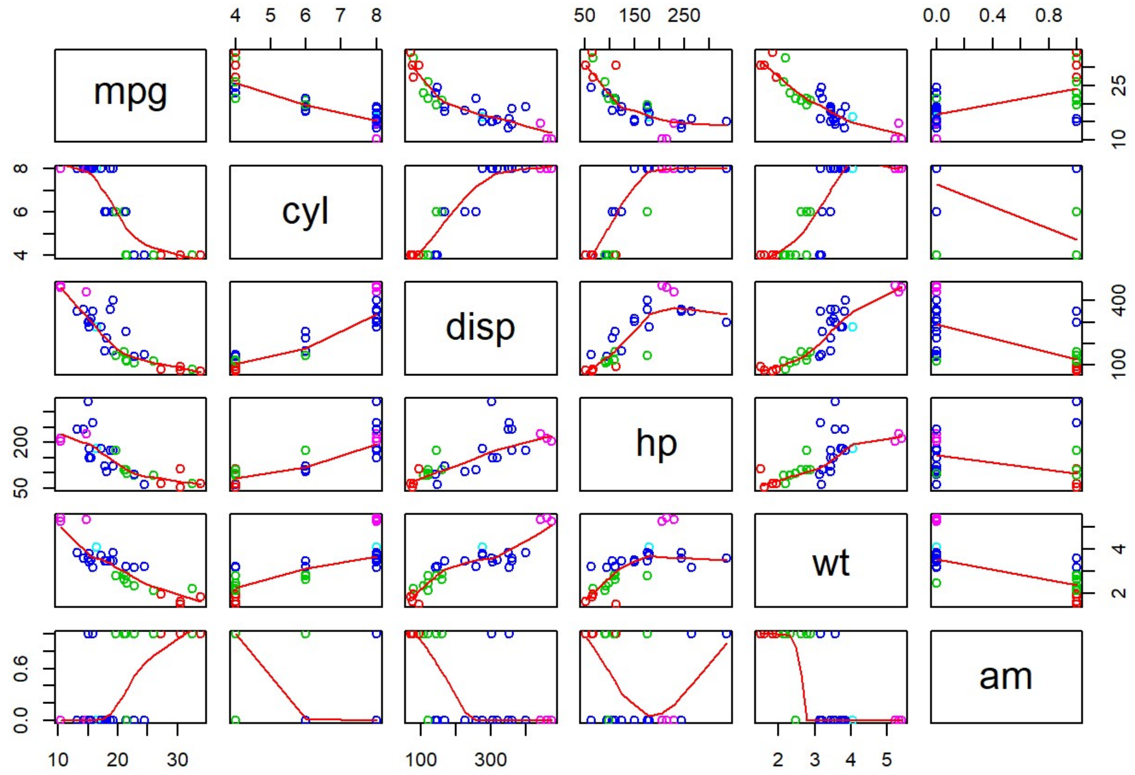
Plot 1 - BoxPlot of mpg per type of transmission

```
boxplot(mpg ~ am, data = mtcars_cool, col = (c("red","blue")), ylab = "Miles Per Ga
llon", xlab = "Transmission Type")
```



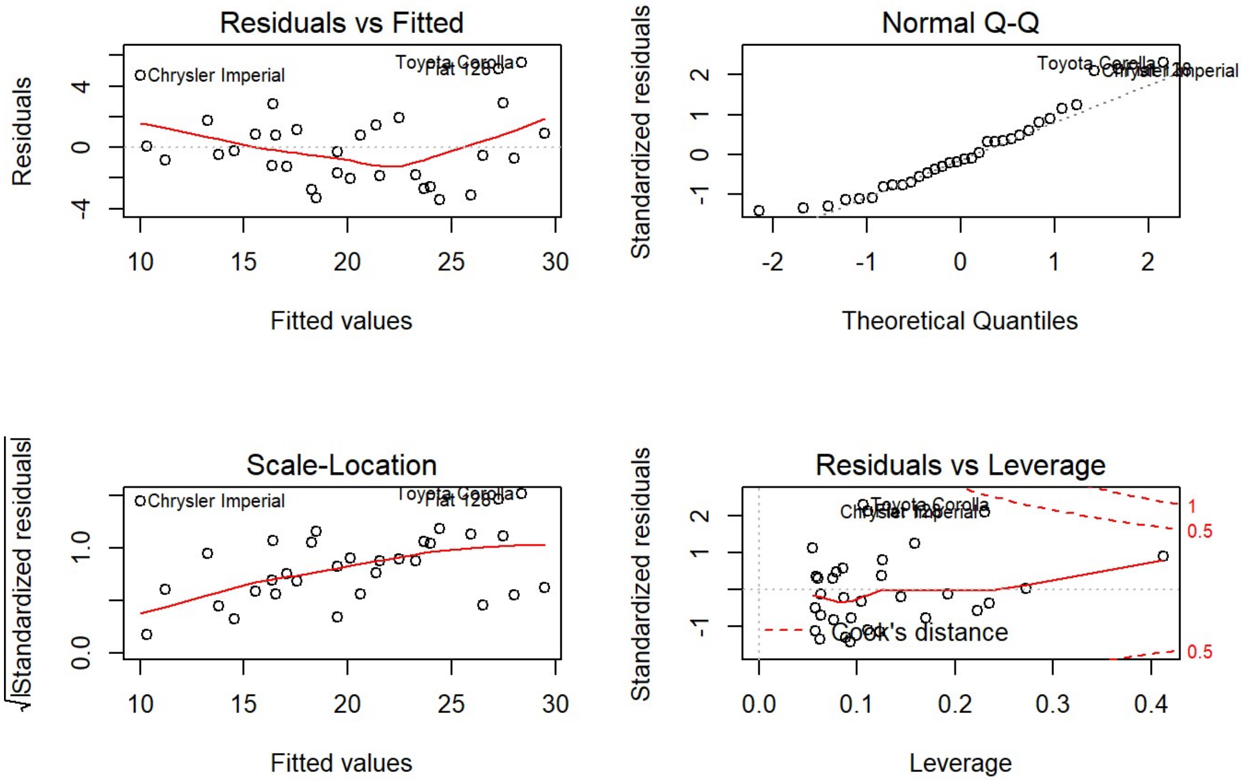
Plot 2 - Correlation of variables in final model

```
mtcarsplot <- mtcars[,c(1,2,3,4,6,9)]
pairs(mtcarsplot, panel = panel.smooth, col = 9 + mtcarsplot$wt)
```



Plot 3 - Plot of residuals in final model

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



Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Toyota Corolla
Fiat 128
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance